Preface

We are very pleased to hold the Fourth Global WordNet Conference in Szeged, Hungary, following our tradition of alternating the meeting locations between different parts of the world.

The program includes 45 paper presentations and demos, two invited talks (Hitoshi Isahara, Adam Kilgarriff) and two topical panels. We received fewer submissions than in previous years; rather than reflecting a decrease in WordNet-related research this probably indicates increased "competition" with the many other conferences and workshops on lexical resources, computational linguistics and Natural Language Processing where work on WordNets is increasingly featured.

We are excited about several new WordNets whose creation is reported here: Croatian, Polish and South African languages. The language of the host country is highlighted with several papers on Hungarian WordNet.

We counted participants from 26 countries in Europe, Asia, Africa, the Near East and the US. Among the authors are many old WordNetters as well as new colleagues, some from countries as far away as Oman and South Africa.

The presentations cover a wide range of topics, including manual and automatic WordNet construction for general and specific domains, lexicography, software tools, ontology, linguistics, applications and evaluation. As in previous meetings, we expect lively discussions and exchanges that plant the seeds for new ideas and future collaborations.

Our thanks go to the Programme Committee who provided thoughtful and fair reviews in a timely fashion.

Christiane Fellbaum, Piek Vossen (for the Global WordNet Organization)

János Csirik, Dóra Csendes (for the Local Organizers)

November, 2007
Organisation

The Fourth Global WordNet Conference is organised by the University of Szeged, Department of Informatics in co-operation with the Global WordNet Association.

The conference home page can be found at http://www.inf.u-szeged.hu/gwc2008.

Programme Committee

Eneko Agirre (San Sebastian, Spain), Zoltan Alexin (Szeged, Hungary), Antonietta Alonge (Perugia, Italy), Pushpak Bhattacharyya (Mumbai, India), Bill Black (Manchester, UK), Jordan Boyd-Graber (Princeton, US), Nicoletta Calzolari (Pisa, Italy), Key-Sun Choi (Seoul, Korea), Salvador Climent (Barcelona, Spain), Dan Cristea (Iasi, Romania), Janos Csirik (Szeged, Hungary), Andras Csomai (Szeged, Hungary), Tomas Erjavec (Ljubljana, Slovenia), Christiane Fellbaum (Princeton, US), Julio Gonzalo (Madrid, Spain), Ales Horak (Brno, Czech Republic), Chu-Ren Huang (Taipei, Republic of China), Hitoshi Isahara (Kyoto, Japan), Neemi Kahuusk (Tartu, Estonia), Kyoko Kanzaki (Kyoto, Japan), Adam Kilgarriff (Brighton, UK), Claudia Kunze (Tuebingen, Germany), Birte Loenneker (Berkeley, US/Hamburg, Germany), Bernado Magnini (Trento, Italy), Palmira Marrafa (Lisbon, Portugal), Rada Mihalceva (Texas, US), Adam Pease (San Francisco, US), Karel Pala (Brno, Czech Republic), Ted Pedersen (Minneapolis, US), Bolette Pedersen (Copenhagen, Denmark), Emanuele Pianta (Trento, Italy), Eli Pociello (San Sebastian, Spain), German Rigau (San Sebastian, Spain), Horacio Rodriguez (Barcelona, Spain), Virach Sornlertlamvanich (Pathumthani, Thailand), Sofia Stamou (Patras, Greece), Dan Tufis (Bucarest, Romania), Tony Veale (Dublin, Ireland), Kadri Vider (Tartu, Estonia), Piek Vossen (Amsterdam, Netherlands)

Organisation Committee

János Csirik (Chair)
Dóra Csendes (Secretary)
Attila Tanács, Dóra Csendes, Veronika Vincze (Proceedings)
Veronika Vincze, Attila Almási, Róbert Ormándi (Helpers)
Christiane Fellbaum, Piek Vossen (Co-organisers)
# Table of Contents

## Papers

Consistent Annotation of EuroWordNet with the Top Concept Ontology ............... 3  
*Javier Álvez, Jordi Atserias, Jordi Carrera, Salvador Climent, Antoni Oliver, German Rigau*

SemanticNet: a WordNet-based Tool for the Navigation of Semantic Information ... 21  
*Manuela Angioni, Roberto Deonitis, Massimo Deriu, Franco Tuveri*

Verification of Valency Frame Structures by Means of Automatic Context  
Clustering in RussNet................................................................................................. 35  
*Irina V.Azarova, Anna S. Marina, Anna A. Sinopalnikova*

Some Issues in the Construction of a Russian WordNet Grid ......................... 44  
*Valentina Balkova, Andrey Sukhonogov, Sergey Yablonsky*

A Comparison of Feature Norms and WordNet .............................................. 56  
*Eduard Barbu, Massimo Poesio*

Enhancing WordNets with Morphological Relations: A Case Study from Czech,  
English and Zulu......................................................................................................... 74  
*Sonja Bosch, Christiane Fellbaum, Karel Pala*

On the Categorization of Cause and Effect in WordNet.................................. 91  
*Cristina Butnariu, Tony Veale*

Evaluation of Synset Assignment to Bi-lingual Dictionary .............................. 101  
*Thatasnee Charoenporn, Virach Sornerlamlavanich, Chumpol Mokarat, Hitoshi  
Isahara, Hammam Riza, Purev Jaimai*

Using and Extending WordNet to Support Question-Answering .................... 111  
*Peter Clark, Christiane Fellbaum, Jerry Hobbs*

Using GermaNet as a Semantic Resource for the Extraction of Thematic  
Structures: Methods and Issues............................................................................... 120  
*Irene Cramer, Marc Finthammer*

On the Utility of Automatically Generated WordNets .................................. 147  
*Gerard de Melo, Gerhard Weikum*
Methods and Results of the Hungarian WordNet Project ........................................... 311
Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik,
Gábor Prószéky, Tamás Váradi

Synset Based Multilingual Dictionary: Insights, Applications and Challenges ....... 321
Rajat Kumar Mohanty, Pushpak Bhattacharyya, Shraddha Kalele, Prabhakar Pandey, Aditya Sharma, Mitesh Kopra

Estonian WordNet: Nowadays.................................................................................. 334
Heili Orav, Kadri Vider, Neeme Kahusk, Sirli Parm

Event Hierarchies in DanNet .................................................................................... 339
Bolette Sandford Pedersen, Sanni Nimb

Building Croatian WordNet ...................................................................................... 349
Ida Raffaelli, Marko Tadić, Božo Bekavac, Željko Agić

Towards Automatic Evaluation of WordNet Synsets ............................................. 360
J. Ramanand, Pushpak Bhattacharyya

Lexical Enrichment of a Human Anatomy Ontology using WordNet.................. 375
Nils Reiter, Paul Buttelaar

Arabic WordNet: Current State and Future Extensions ......................................... 387
Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, Musa Alkhalfi, M. Antonia Martí, William Black, Sabri Elkateb, James Kirk, Adam Pease, Piek Vossen, Christiane Fellbaum

Building a WordNet for Persian Verbs................................................................. 406
Masoud Rouhizadeh, Mehrnoush Shamsfard, Mahsa A. Yarmohammadi

Developing FarsNet: A Lexical Ontology for Persian ............................................ 413
Mehrnoush Shamsfard

KUI: Self-organizing Multi-lingual WordNet Construction Tool .......................... 419
Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergrit Robkop, Hitoshi Isahara

Extraction of Selectional Preferences for French using a Mapping from
EuroWordNet to the Suggested Upper Merged Ontology ................................. 428
Dennis Spohr

Romanian WordNet: Current State, New Applications and Prospects .............. 441
Dan Tușiț, Radu Ion, Luigi Bozianu, Alexandru Ceaușu, Dan Ștefănescu

Enriching WordNet with Folk Knowledge and Stereotypes .............................. 453
Tony Veale, Yanfen Hao
Papers
Consistent Annotation of EuroWordNet with the Top Concept Ontology

Javier Álvez¹, Jordi Atserias², Jordi Carrera³, Salvador Climent³, Antoni Oliver³, and German Rigau¹

¹ Basque Country University.
² Web Research Group - Universitat Pompeu Fabra
³ Open University of Catalonia.

jibalgij@si.ehu.es, jordi.atserias@upf.edu, jcarrerav@uoc.edu, scliment@uoc.edu, aoliverg@uoc.edu, german.rigau@ehu.es

Abstract. This paper presents the complete and consistent annotation of the nominal part of the EuroWordNet (EWN). The annotation has been carried out using the semantic features defined in the EWN Top Concept Ontology. Up to now only an initial core set of 1024 synsets, the so-called Base Concepts, were ontologized in such a way.

1 Introduction

Componental semantics has a long tradition in Linguistics since the work of post-structuralists as Hjelmslev in the thirties [cf. [1]] or [2] among generativists. There is common agreement that this kind of lexical-semantic information can be extremely valuable for making complex linguistic decisions. Nevertheless, according to [1], componental analysis cannot be actually achieved due to three main reasons (being the first the most important): (1) the vocabulary of a language is too large, (2) each word needs several features for its semantics to be adequately represented and (3) semantic features should be organized in several levels.

Our work provides a good solution to these problems, since 65,989 noun concepts from WordNet 1.6 (WN16) [3] corresponding to 116,364 noun lexemes (variants) have been consistently annotated with an average of 6.47 features per synset, being those features organized in a multilevel hierarchy. Therefore, it might allow componental semantics to be tested and applied in real world situations probably for the first time, thus contributing to a wide number of NLP tasks involving semantic processing: Word Sense Disambiguation, Syntactic Parsing using selectional restrictions, Semantic Parsing or Reasoning.

Despite its wide scope, the work presented here is envisaged to be the first stage of an incremental and iterative process, as we do not assume that the current version of the EWN Top Concept Ontology (TCO) covers the optimal set of features for the aforementioned tasks. Currently, a second phase has started within the framework of
the KNOW Project\(^1\) in which the first version of the enriched lexicon is being used to label a corpus. We plan to use later this annotation for abstracting the semantic properties of verbs occurring in the corpus. This will lead, presumably, to a reformulation of the TCO, through addition, deletion or reorganisation of features.

In this paper, is organized as follows. After a brief summary of the state of the art (section §2), we present our methodology for annotating the nominal part of EWN (section §3). Then, we provide a qualitative analysis by providing some relevant examples (section §4). Section §5 summarizes a quantitative analysis and finally, section §6 provides some concluding remarks.

2 Previous Work and State of the Art

2.1 The EuroWordNet Top Ontology

The EWN TCO was not primarily designed to be used as a repository of lexical semantic information, but for clustering, comparing and exchanging concepts across languages in the EWN Project. Nevertheless, most of its semantic features (e.g. Human, Object, Instrument, etc.) have a long tradition in theoretical lexical semantics and have been postulated as semantic components of meanings. We will only describe here some of its major characteristics (see [4] for further details).

The EWN TCO (Fig. 1) consists of 63 features and it is primarily organized following [5]. Correspondingly, its root level is structured in three disjoint types of entities:

- 1stOrderEntity (physical things, e.g.: vehicle, animal, substance, object)
- 2ndOrderEntity (situations, e.g.: happen, be, begin, cause, continue, occur)
- 3rdOrderEntity (unobservable entities e.g.: idea, information, theory, plan)

1stOrderEntities are further distinguished in terms of four main ways of conceptualizing or classifying concrete entities:

- Form: as an amorphous substance or as an object with a fixed shape (Substance or Object)
- Composition: as a group of self-contained wholes or as a necessary part of a whole, hence the subdivisions Group and Part.
- Origin: the way in which an entity has come about (Artifact or Natural).
- Function: the typical activity or action that is associated with an entity (Comestible, Furniture, Instrument, etc.)

These main features are then further subdivided. These classes are comparable to the Qualia roles as described in [6] and are based on empirical findings raised during the development of the EWN project, when the classification of the Base Concepts (BCs) was undertaken. Concepts can be classified in terms of any combination of these four roles. As such, these top concepts function more as features than as ontological classes.

\(^1\) KNOW. Developing large-scale multilingual technologies for language understanding. Ministerio de Educación y Ciencia. TIN2006-15049-C03-02.
Although the main-classes are intended for cross-classification, most of the subdivisions are disjoint classes: a concept cannot be both an Object and a Substance, or Natural and Artifact. As explained below, feature disjunction will play an important role in our methodology.

2ndOrderEntities can lexicalize both nouns and verbs (as well as adjectives and adverbs) denoting static or dynamic situations, such as birth, live, life, love, die and death. All 2ndOrderEntities are classified using two different classification schemes:

- **SituationType**: the event-structure in terms of which a situation can be characterized as a conceptual unit over time
- **SituationComponent**: the most salient semantic component(s) that characterize(s) a situation

**SituationType** represents a basic classification in terms of the event-structure (in the formal tradition) or the predicate-inherent Aktionsart properties of nouns and verbs, as described for instance in [7]. **SituationTypes** can be Static or Dynamic, further subdivided in Property and Relation on the one side and UnboundedEvent and BoundedEvent on the other.

**SituationComponents** (e.g. Location, Existence, Cause, Mental, Purpose) emerged empirically when selecting verbal and deverbal Base Concepts in EWN. They resemble the cognitive components that play a role in the conceptual structure of events, as described in [8] and others. In fact, much in the same way as **Function** did for 1stOrderEntities, they are good candidates for encoding important semantic properties of words denoting situations.

Typically, **SituationType** represents disjoint features that can not be combined, whereas it is possible to assign any range or combination of **SituationComponents** to a word meaning. Each 2ndOrderEntity meaning can thus be classified in terms of an obligatory but unique **SituationType** and any number of **SituationComponents**.

Finally, 3rdOrderEntities was not further subdivided, since there appeared to be a limited number of BCs of this kind in EWN.

The TCO has been redesigned twice, first by the EAGLES expert group [9] and then by [10]. EAGLES expanded the original ontology by adding 74 concepts while the latter made it more flexible, allowing, for instance, to cross-classify features between the three orders of entities.

Moreover, the Global Wordnet Association [11] recently distributed a taxonomy consisting of 71 so-called Base Types which can be seen as semantic primitives or taxonomic tops playing a key role in large-scale semantic networks. The Base Types have been derived by refining the original set of BCs. They are connected to both EWN synsets and TCO features, and represent an important synthesis effort in order to achieve a more elegant and economic modelling of the TCO.
1stOrderEntity
   Form
      Object
      Substance
         Gas
         Liquid
         Solid
   Composition
      Group
      Part
   Origin
      Artifact
      Natural
         Living
         Animal
         Creature
         Human
         Plant
   Function
      Building
      Comestible
      Container
      Covering
      Furniture
      Garment
      Instrument
      Occupation
      Place
      Representation
         ImageRepresentation
         LanguageRepresentation
         MoneyRepresentation
      Software
      Vehicle
2ndOrderEntity
   SituationType
      Dynamic
         BoundedEvent
         UnboundedEvent
      Static
         Property
         Relation
   SituationComponent
      Cause
         Agentive
         Phenomenal
         Stimulating
   Communication
      Condition
      Existence
      Experience
      Location
      Manner
      Mental
      Modal
      Physical
      Possession
      Purpose
      Quantity
      Social
      Time
      Usage
3rdOrderEntity

Fig. 1. The EWN Top Concept Ontology
2.2 Ontological information in the Multilingual Central Repository

In the framework of the UE-funded MEANING project [12] a Multilingual Central Repository (MCR) was designed and implemented in order to act as a multilingual interface for integrating and distributing lexical-semantic knowledge [13]. The MCR follows the model proposed by the EuroWordNet project (EWN) [14], i.e. a multilingual lexical database with WordNets for several languages. It includes WordNets for English, Spanish, Italian, Catalan and Basque.

The EWN architecture includes the Inter-Lingual-Index (ILI), which is a list of records that interconnect synsets across WordNets. Using the ILI, it is possible to go from word meanings in one language or particular WordNet to their equivalents in other languages or WordNets. The current version of the MCR uses the set of Princeton WordNet 1.6 synsets as ILI.

In the MCR, the ILI is connected to three separate ontologies: the EWN TCO (described above), the Domain Ontology (DO) [15] and the Suggested Upper Merged Ontology (SUMO) [16]. The DO is a hierarchy of 165 domain labels, which are knowledge structures grouping meanings in terms of topics or scripts, e.g. Transport, Sports, Medicine, Gastronomy. SUMO incorporates previous ontologies and insights by Sowa, Pierce, Russell and Norvig and others and, compared to EWN TCO, is much larger and deeper. The WN-SUMO mapping [17] assigns only one SUMO category to every WN16 synset (being SUMO a large formal ontology), while the EWN TCO, as explained above, assigns a combination of a more reduced number categories. This makes the TCO much more suitable than that of SUMO for implementing componential semantics. While all the ILI is connected to the DO and to SUMO, only 1024 ILI-Records were connected to the TCO, i.e. those were selected as BCs in the EWN project.

2.3 Lexical Semantics for Robust NLP

Some NLP systems, such as knowledge-based Machine Translation systems usually include some kind of decision making (e.g. transfer module, PP-attachment) using lexical semantic features such as Human, Animate, Event, Path, Manner etc. [18]. Its use, however, is restricted to demo systems, e.g. [19] or, in real-world systems, to a limited number of lexical entries or/and to a very reduced number of semantic features, due to the difficulty of annotating a comprehensive lexicon with an exhaustive set of features.

However, WordNets are large lexical resources freely-available and widely used by the NLP community. Currently, they serve a wide number of tasks involving some degree of semantic processing. In most of these tasks, WordNets are used to generalize or abstract a set of synsets to a subsuming one by following the WordNet hierarchy up. The main problem is finding the right level of generalization; that is, finding the concept which optimally subsumes a given set of concepts; but it could be the case that the class which would optimally capture the generalization is not lexical, but abstract –thus having to be represented through features. It can also be the case

\[^2\] http://adimen.si.ehu.es/cgi-bin/wei5/public/wei.consult.perl
that WordNet simply is not the kind of taxonomy required, fact which can be due to several reasons: incompleteness, incorrect structuring, or perhaps that its structuring should be arranged differently for a particular NLP task.

Bearing these drawbacks in mind, some authors have turned to use the ontologies mapped onto WordNet to determine new sets of classes. For instance, [20] and [21] have already used the MCR including SUMO, DO, WN16 Semantic (Lexicographer’s) Files and a preliminary rough expansion of the TCO for Word Sense Disambiguation.

For many tasks, it seems that using a feature-annotated lexicon seems more appropriate than using the WordNet tree-structure, since (i) the WordNet hierarchy is not consistently structured [22] and (ii) a feature-annotated lexicon allows to make predictions based on measures of similarity even for words that, being sparsely distributed in WordNet, can only be generalized by reaching common hypernyms in levels too high in the hierarchy. Besides, a multiple-feature design allows to naturally depict semantically complex concepts, such as so-called dot-objects [6], e.g., intrinsically polysemic words such as “letter”, since a letter is something that can both be destroyed and carry information (as in “I burnt your love letter”). These aspects of meaning can be easily coded through using the EWN TCO, as shown in (1)

(1) “letter”: FUNCTION: LanguageRepresentation
FORM: Object

In this direction, [23] uses a lexicon augmented with EWN TCO features both to implement selectional restrictions to limit the search space when parsing and to perform type-coercion in a dialogue system.

3 Methodology

Our methodology for annotating the ILI with the TCO\(^3\) is based on the common assumption that hyponymy corresponds to feature set inclusion [24, p.8] and in the observation that, since WordNets are taken to be crucially structured by hyponymy “(...) by augmenting important hierarchy nodes with basic semantic features, it is possible to create a rich semantic lexicon in a consistent and cost-effective way after inheriting these features through the hyponymy relations” [9, pp. 204-205].

Nevertheless, performing such operation is not straightforward, as WordNets are not consistently structured by hyponymy [22]. Moreover, WordNets allow multiple inheritance. These are both drawbacks to overcome and situations to take advantage of by our methodology. As told above, within the EWN project, a limited set of lexical base concepts\(^4\) (the BCs) was annotated with TCO features. Despite being

\(^3\) We use WN16 since the ILI is drawn up on this version of WordNet

\(^4\) Base Concepts (BCs) should not be confused with Basic Level Concepts (BLCs) as defined by [26] but in a future work BCs can be taken as a starting set to define that of BLCs. Since BLCs are supposed to be richer in distinctive features and the most psychologically salient lexical categories, they can also be relevant for advanced NLP tasks.
largely general in meaning, this set did not cover all of the upper level nodes in the
WordNets. This was clearly a drawback for expanding features down all of WN1.6,
thus the first step of our work consisted of annotating the gaps up the hierarchy, from
the BCs to the Unique beginners. This was made semiautomatically: given that every
synset in WN1.6 originally belongs to a so-called Semantic File (a flat list of 45
lexicographer files), those synsets were assigned a TCO feature via a table of
expected equivalence between TCO nodes and Semantic Files.
This made the WN1.6 ready to be fully populated with at least one feature per
synset. Nevertheless, in many cases, synsets got more than one feature, for one or
more of the following reasons:

- They are BCs, so they were manually annotated with more than one
  feature
- In addition to their own manual annotation, they inherit features from
  one or more hypernyms
- They inherit features from different hypernyms, either located at
different levels in a single line of hierarchy or by the effect of multiple
  inheritance

An initial rough expansion was the first ground for revision and inspection,
following the strategy defined in [25]. The task has lasted for about three years and
has involved several re-expansion cycles.
The manual work has been based on TCO feature incompatibilities. It consisted in
automatically detecting co-occurrences in a synset of pairs of incompatible features.
The axiomatic incompatibilities are the following:

- 1stOrderEntity - 2ndOrderEntity
- 1stOrderEntity - 3rdOrderEntity
- 3rdOrderEntity - 2ndOrderEntity [except for SituationComponent]
- 3rdOrderEntity - Mental5
- Object - Substance
- Gas - Liquid - Solid
- Artifact - Natural
- Animal - Creature - Human - Plant
- Dynamic - Static
- BoundedEvent - UnboundedEvent
- Property - Relation
- Physical - Mental
- Agentive - Phenomenal – Stimulating

The first rough expansion described above caused the following number of feature
conflicts:

5 The incompatibility between 3rdOrderEntity and Mental and the compatibility between
3rdOrderEntity and Situation Components is explained below.
- 214 feature conflicts in 49 synsets caused by incompatible hand annotation
- 2247 feature conflicts in 743 synsets caused by hand annotation incompatible with inherited features
- 225,447 feature conflicts in 26,166 synsets caused by incompatibility between inherited features

The first type of conflicts usually indicates synsets causing ontological doubts to annotators within the EWN project (e.g. is “skin” an object or a substance?). The third type usually reveals errors in WordNet structure (i.e. ISA overloading [22]). The second type might be caused by either or both reasons.

The task consisted on manual checking feature incompatibilities in order to (i) adding or deleting ontological features, and (ii) setting inheritance blockage points. A blockage point is an annotation in WN1.6 which breaks the ISA relation between two synsets, thus no information can be passed by inheritance through it.

When a case of feature incompatibility occurred, the synset involved, together with its structural surroundings (hypernyms, hyponyms), was analyzed. If the problem was due to a WN1.6 subsumption error, the corresponding link was blocked and synsets below the blockage point are annotated with new TCO features.

Changes in the annotation were made and blockage points were set until all conflicts were resolved. Then a second re-expansion of TCO features was launched which resulted in a new (smaller) number of conflicts. Following this iterative and incremental approach, inheritance was being re-calculated and the resulting data was re-examined several times. Although such hand-checking is extremely complex and laborious, and despite the large number of conflicts to solve, the task ended up being feasible because working on the topmost origin of one feature conflict results in fixing many levels of hyponyms. For instance, leaf_1, “the main organ of photosynthesis and transpiration in higher plants”, is a synset that subcategorizes 66 kinds of leaves. It was originally categorized as Substance, but, being in that sense a bounded entity, it seemed clear that it cannot be assigned such TCO label. Therefore, fixing this case resulted in fixing as many as 66 conflicts downward with a single action.

The task has been carried out using application interfaces, which allowed access to the synsets and their glosses in three languages at the same time: English, Spanish and Catalan. The information that was relied on in order to make decisions was of the following kinds:

- Relational information regarding every synset and neighboring ones; i.e. the WN1.6 structure
- The nature of the feature conflict (any of the three types of incompatibility aforementioned)
- Synsets' glosses as provided by EWN
- Glosses, descriptions and examples of the TCO features as provided in [4]
- Usual word-substitution tests that acknowledge hyponymy, as in [27, pp. 88-92]

The task finished when finally a re-expansion of properties did not result in new conflicts. Then, two final steps were applied. First, as the TCO is itself a hierarchy, for every synset, its resulting annotation was expanded up-feature; e.g. if a synset
beared the feature Animal it was also labelled Living, Natural, Origin and 1stOrderEntity. Second, the whole noun hierarchy was been checked for consistency using several formal Theorem Provers like Vampire [28] and E-prover [29]. This step resulted in a number of new conflicts which were finally fixed.

This methodology has led to detect many more inconsistencies in WordNet and much deeper into the hierarchy than previous approaches (e.g. [30]).

This procedure can be seen as a shallow ontologization of WN1.6. That is, blocked links are reassigned to the TCO. This constitutes a pragmatic solution to the problem of the difficulty of complete WordNets ontologization. In this sense, our work will probably be the second one to ontologize the whole WordNet, after that with SUMO [17]. However, our coding (i) is multiple (SUMO links every synset to only one label of the ontology) and (ii) it is more workable since it uses a more intuitive and simple TCO.

Regarding the completion of the work, the possibility that some areas in the WordNet hierarchy have remained unexamined cannot be completely excluded, although a very large number of changes have been introduced: (i.e. more than 13.000 manual interventions). Moreover, it should be noticed that, when removing links or features to fix errors, all hyponimy lines involved by the action have been reexamined and reannotated in order not to loss information.

4 Examples and qualitative discussion

In this section some examples of our methodology are presented at work. Hereinafter, noun synsets are represented by one of their variants enclosed in curly brackets and TCO features by its name in italics, capitalized and enclosed in square brackets. Inherited features are marked ‘+’ while manually assigned features are marked ‘=’.

4.1 Bandung is not Java but a part of it

A simple but very typical case is the following, in which the conflict results from multiple inheritance and the incorrect use of hyponymy instead of meronomy in WN1.6:

\{
\text{Bandung}_1^{6} \text{[Artifact+ Natural+]}\}
\rightarrow \{\text{Java}_1 \text{[Natural+]}\}
\rightarrow \{\text{island}_1 \text{[Natural+]}\}
\rightarrow \{\text{city}_1 \text{[Artifact=}]\}

Clearly, Bandung is a city, but it is \textit{not} a Java (though it is \textit{part} of Java). This case is revealed thanks to incompatibility between \textit{Natural} and \textit{Artifact}. It is fixed by blocking the subsumption link between Bandung_1 and Java_1:

---

\(^6\) A city in the island of Java.
4.2 A drug is a substance

This case is less straightforward but as well quite representative of malfunctions in the WN1.6 hierarchy. In WN1.6, {artifact_1} is both glossed as "a man-made object" and an hyponym of {physical_object_1}. Thus, in EWN it was annotated with the TCO feature [Object], which stands for bounded physical things. Nevertheless, its hyponym {drug_1} subsumes substances, therefore, it was annotated in EWN as [Substance]. It seems clear that the WN1.6 builders wanted to capture the fact that drugs are artificial compounds (although there indeed exist natural drugs\(^7\)). But this fact, which is represented by the ISA relation between {drug_1} and {artifact_1} is not consistent with conceptualising {artifact_1} as a physical, bounded, object. In our work, feature expansion revealed the contradiction, since TCO features [Object] and [Substance] are incompatible:

\[
\{\text{artifact}_1 \text{ [Object=]}\} \\
\quad \rightarrow \{\text{article}_2 \text{ [Object+]}\} \\
\quad \rightarrow \{\text{antiquity}_3 \text{ [Object+]}\} \\
\quad \rightarrow \{\text{... [Object+]}\} \\
\quad \rightarrow \{\text{drug}_1 \text{ [Substance= Object+]}\} \\
\quad \rightarrow \{\text{aborticide}_1 \text{ [Substance= Object+]}\} \\
\quad \rightarrow \{\text{anesthetic}_1 \text{ [Substance= Object+]}\} \\
\quad \rightarrow \{\text{... [Substance=]} \text{ [Object+]}\}
\]

In this case, there were two possible solutions: either to underspecify {artifact_1} for Object and Substance, thus allowing it to subsume both kinds of entities, or blocking the subsumption relation between {drug_1} and {artifact_1}. We chose the latter solution because {artifact_1} mainly subsumes hundreds of physical objects in WN1.6. Moreover, this solution is consistent with the glosses and respects the statement of {artifact_1} as hyponym of {physical_object_1}. Therefore, it seems better to treat {drug_1} as an exception than to change the whole structure:

\[
\{\text{artifact}_1 \text{ [Object=]}\} \\
\quad \rightarrow \{\text{article}_2 \text{ [Object+]}\} \\
\quad \rightarrow \{\text{antiquity}_3 \text{ [Object+]}\} \\
\quad \rightarrow \{\text{... [Object+]}\} \\
\quad \rightarrow \{\text{drug}_1 \text{ [Substance=}\} \\
\quad \rightarrow \{\text{aborticide}_1 \text{ [Substance+]}\} \\
\quad \rightarrow \{\text{anesthetic}_1 \text{ [Substance+]}\} \\
\quad \rightarrow \{\text{... [Substance+]\}
\]

\(^7\) This fact prevents {drug_1} to be labelled [Artifact]. Only a number of its hyponyms can be done so.
If we conceptualize the annotation with the TCO not just as simple feature labelling but as connecting WN1.6 to an upper flat abstract ontology, this solution is equivalent to chopping off the \{drug_1\} subtree and link it to the \{Substance\} node of the TCO:

\[
\text{\begin{itemize}
\item \text{\{artifac_1\}}
\item \text{\{article_2\}}
\item \text{\{antiquity_3\}}
\item \text{\{...\}}
\item \text{\{Substance\}}
\item \text{\{drug_1\}}
\item \text{\{aborticide_1\}}
\item \text{\{anesthetic_1\}}
\item \text{\{...\}}
\end{itemize}}
\]

This vision was termed the shallow ontologization of WordNet in [25].

4.3 The Statue of Liberty

In this section, a complete case is described showing how one single feature conflicting in the bottom of the hierarchy reveals a chain of inconsistencies up to the upper levels of the taxonomy, thus resulting in hundreds of wrongly classified synsets. We also show how our methodology is applied to solve these problems.

One conflict between first order and second order features originally taking place in \{Statue_of_Liberty_1\} climbs up to \{creation_2\} reveals the big confusion existing in WN1.6 regarding art, artistic genres, works of art and art performances (the last being events). Fixing this involved blockages and feature underspecification throughout the hierarchy. Finally, one synset \{creation_2\} should be underspecified as it would need disjunction of properties to be properly represented, as it can be either an object or an event.

In order to facilitate the explanation, synsets are represented by a single intuitive word, the \textit{3rdOrderEntity} feature is more intuitively represented as \{Concept\} and only the more relevant synsets and features are shown.
Fig. 2. The case of the Statue. Initial situation.
As a starting point, there were four BCs manually annotated in EWN: {artifact} as [Object], {abstraction} as [Concept], {attribute} as [Property] and {sculpture} as [ImageRepresentation]. Figure 2 shows the clear-cut result of a direct expansion of properties by feature inheritance.

As a result of this process, several shocking annotations can be noticed at a first sight, for instance: (1) {musical composition}, {dance} and {impressionism} as [Object]; (2) {sculpture} as [Property], and (3) {Statue_of_Liberty} as [Concept].

Notice that we became aware of all this situation by inspecting the incompatibility of those TCO features inherited by {Statue of Liberty}. Due to multiple inheritance, the popular monument was taken to be an artifact, hence an object; but at the same time a kind of [art] —as e.g. {dance}, which is clearly an event, while {impressionism} is nothing but a concept. Moreover, {Statue of Liberty} appeared to be an abstraction, a [Concept], just as the geometric notion of a [plane]. Last, the statue also inherited [Property]. So, the result of applying full inheritance of ontological properties in WN1.6 resulted in multiple incompatible features eventually colliding at {Statue_of_Liberty}.

The analysis of the situation led to blockage of the following hierarchy paths, as it is shown in Figure 3:

- Between {artifact} and {creation}
- Between {art} and {dance} (but not between {art} and {genre})
- Between {plastic_art} and {sculpture}
- Between {three_dimensional_figure} and {sculpture}

Moreover, {creation} was underspecified by assigning the upmost neutral feature [Top] and [Property] was deleted in {attribute} since it is better represented by {attribute}’s hyponym {property} while the rest of hyponyms here considered (lines, planes, etc.) are, according to their glosses and relations, concepts.

The reasons behind these changes were the following:

(1) Although, intuitively, one might say that a creation is an artifact (for creations are made by men), according to the glosses and hyponyms one can realize that the synset {artifact} subsumes objects, while {creation} subsumes both objects and activities brought about by men (e.g. a “musical composition”). Therefore, {creation} can not inherit first order features, since they are incompatible with second order ones. Consequently, {creation} was here labeled as [Top] thus allowing its hyponyms to be further specified as entities or events since neither its gloss (“something that has been brought into existence by someone”) nor the lack of homogeneity of its hyponyms allowed to make a choice. In a more flexible version of the TCO, as that proposed by [10], [Origin] features could be also attributed to second and third order entities. This will allow to assign [Artifact] to synsets like {Creation}. We intend to evolve to a TCO like this in the future.

(2) Although, intuitively, one might say that dance is a kind of art, according to the glosses and other hyponyms it is realized that {art} refers to the concept (like e.g. {impressionism}) while {dance} refers to an activity. Therefore, while “art” and “impressionism” are considered ideas, “dance”, however, is an activity.
Although, intuitively, one might say that sculpture is a plastic art, according to the glosses and other hyponyms one can realize that, as regards the senses given, {sculpture} refers to physical objects, while {plastic_art} refers to the abstract concept — a type of {art}, such as {impressionism}.

Although, intuitively, one might say that a sculpture is a three dimensional figure, according to the glosses and other hyponyms it is realized that, {three_dimensional_figure} refers to the shape (the same as one-dimensional lines or two-dimensional planes, that is, abstract shapes). Therefore, in this sense, “sculptures” are objects while “figures” or “shapes” are geometrical abstractions.

The final result, as it can be seen in Figure 3, is a new quite reasonable labelling of the set of concepts, implicitly involving a reorganisation of the WN1.6 hierarchy. It is easy to realize how these limited set of decisions (four blockages, one feature deletion and few feature relabelling) subsequently affect hundreds of synsets. For instance, {creation} and {sculpture} relate to 713 and 28 hyponyms respectively.

4.4 Notes for further discussion

During the time devoted to carry on the work, a lot of interesting facts have been discovered about two objects of study: the structure of the noun hierarchy of WN1.6 and the nature of the EWN TCO features – as well as the mapping between both.

These facts are going to be further studied taken into consideration at least the following issues:

- To which extent those noun hierarchy problems correspond to those described in [22] or there are other kinds of facts distorting the WordNet structure
- Typical doubts or mistakes in the BCs annotation with the TCO carried on in the EWN Project
- Problems related to lack of clear definition of either synsets in WordNet or features in the TCO

For instance, a very common malpractice in EWN when annotating BCs with the TCO was that of the double coding of non-physical entities both as 3rdOrderEntity and Mental. Mental is a subfeature for 2ndOrderEntity and, as far as 2ndOrderEntity and 3rdOrderEntity are explicitly declared as incompatible, Mental and 3rdOrderEntity can not coexist. Therefore, Mental has to be deleted, since what the encoder was intuitively doing in these cases was telling twice that the synset stands for a mental or conceptual entity. In a future enhanced TCO, following [10], it would be better to allow Origin, Form, Composition and Function features to be applied to situations and concepts, instead of the current classification based on 3rdOrderEntity and Mental being disjunct. This will allow Concept to be cross-classified as for instance to classify “Neverland” as both Concept and Place in order to indicate that it is an imaginary location or to underspecify "creation" by classifying it simply as Artifact.
Fig. 3. The case of the Statue. Final result
5 Quantitative analysis

Summarizing, the whole process provided a complete and consistent annotation of the nominal part of WN1.6 which consist of 65,989 synsets nominals with 116,364 variants or senses. All 227,908 initial incompatibilities were solved by manually adding or removing 13,613 TCO features and establishing 359 blockage points. Now, the final resource has 207,911 synset-feature pairs without expansion and 427,460 synset feature pairs with consistent feature inheritance.

6 Conclusions and further work

We have presented the full annotation of the nouns on the EuroWordNet (EWN) Interlingual Index (ILI) with those semantic features constituting the EWN Top Concept Ontology (TCO). This goal has been achieved by following a methodology based on an iterative and incremental expansion of the initial labelling through the hierarchy while setting the inheritance blockage points. Since this labelling has been set on the ILI, it can be also used to populate any other WordNet linked to it through a simple porting process.

This resource is intended to be useful for a large number of semantic NLP tasks and for testing for the first time componential analysis on real environments.

Moreover, those mistakes encountered in WordNet noun hierarchy (i.e. false ISA relations), which are signalled by more than 350 blocking annotations, provide an interesting resource which deserves future attention.

Further work will focus on the annotation of a corpus oriented to the acquisition of selectional preferences. This, compared to state-of-the-art synset-generalisation semantic annotation, will result in a qualitative evaluation of the resource and in gaining knowledge for designing an enhanced version of the Top Concept Ontology more suitable for semantically-based NLP.

References


http://lpg.uoc.edu/files/wei-topontology.2.2.rar


Manuela Angioni, Roberto Demontis, Massimo Deriu, and Franco Tuveri

CRS4 - Center for Advanced Studies, Research and Development in Sardinia, Polaris - Edificio 1, 09010 Pula (CA), Italy
{angioni, demontis, deriu, tuveri}@crs4.it

Abstract. The main aim of the DART search engine is to index and retrieve information both in a generic and in a specific context where documents can be mapped or not on ontologies, vocabularies and thesauri. To achieve this goal, a semantic analysis process on structured and unstructured parts of documents is performed. While the unstructured parts need a linguistic analysis and a semantic interpretation performed by means of Natural Language Processing (NLP) techniques, the structured parts need a specific parser. In this paper we illustrate how semantic keys are extracted from documents starting from WordNet and used by an automatic tool in order to define a new semantic net called SemanticNet build enriching the WordNet semantic net with new nodes, links and attributes. Formulating the query through the search engine, the user can move through the SemanticNet and extracts the concepts which really interest him, limiting the search field and obtaining a more specific result by means of a dedicated tool called 3DUI4SemanticNet.

Keywords: Semantic net, Ontologies, NLP, 3D User Interface.

1 Introduction

The main aim of the DART ([1] and [2]) project is to realize a distributed architecture for a semantic search engine, facing the user with relevant resources in reply to a query about a specific domain of interest. In this paper we expose concepts and solutions related to the semantic aspects and to the geo-referencing features, designed to support the user in the information retrieval and to supply position based information strictly related to a specified area.

In order to reach this goal, a prototype able to enrich the WordNet [3] semantic net by means of new concepts often related to specific knowledge domains has been realized as part of the DART project[4]. This aspect of the problem bring us to distinguish between a specific and a generic context both in indexing and in retrieval of information, whether documents can be mapped or not on ontologies, vocabularies and thesauri.

The paper deals with two main aspects. The first is the definition of a semantic net, called SemanticNet, and the definition of a 3D user interface, called 3DUI4SemanticNet, that allows to navigate in the concepts through relations. The second puts in evidence the use of ontologies or structure descriptors, as in the case of
dynamic XML documents generated by Web services, RDF and OWL documents, and the possibility to build specialized Semantic Nets based on a specific context of use.

Section 2.1 describes how the system behaves in the general context while section 2.2 does it in specific contexts. Section 3 describes how the SemanticNet is obtained from WordNet and enriched by means of a certified source of documents. Section 4 proposes a use case related to a GIS (Geographical Information System) specific domain. Finally, section 5 describes the 3DUI4SemanticNet, a 3D navigation tool developed in order to give users a friendly interface to information contained in the SemanticNet.

2 Generic and Specific Contexts

Our main goal is to index and retrieve information both in a generic and in a specific context whether documents can be mapped or not on ontologies, vocabularies and thesauri. To achieve this goal, a semantic analysis process on structured and unstructured parts of documents is performed.

The first step of this process is to identify structured and unstructured parts of a document. The structured portions are identified by ontologies or structure descriptors, as in the case of dynamic XML documents generated by Web services, RDF and OWL documents, if they are known by the system.

In the second step the semantic analysis is performed. The unstructured parts need a linguistic analysis and a semantic interpretation performed by means of Natural Language Processing (NLP) techniques, while the structured parts need a specific parser. Finally, the system extracts the correct semantic keys that represent the document or a user query through a syntactic and a semantic analysis and identifies them by a descriptor which is stored in a Distributed Hash Tables, DHT [1]. These keys are defined starting from the semantic net of WordNet, a lexical dictionary for the English language. The system defines, then, a new semantic net, called SemanticNet, derived from WordNet [5] by adding new nodes, links and attributes.

Web resources managed by the system are naturally heterogeneous, from the contents point of view. The classification processing of web resources is necessary in order to identify the meaning of keywords by means of their context of use. The same keywords are the references the system uses both in indexing and searching phases.

2.1 The Generic Context

In a generic context the system doesn't know a formal or terminological ontology describing the specific semantic domain. So, the concepts defined in the structured part of the document by a formal or terminological ontology are evaluated using the WordNet point of view. The conceptual mapping could reduce the quality of information, but its importance is to allow a better navigation between concepts defined in two ontologies, using a shared semantics. In general, the more accurate the conceptual mapping will be, the better the response of the system. The module can
also extract the specialized semantics keys from the structured part of a document and return the generic semantic key mapped to these keys by means of the conceptual mapping.

Unstructured parts need a linguistic analysis and a semantic interpretation to be performed by means of NLP techniques. The main tools involved are:

- the Syntactic Analyzer and Disambiguator, a module for the syntactic analysis, integrated with the Link Grammar parser [6], a highly lexical, context-free formalism. This module identifies the syntactical structure of sentences, and resolves the terms' roles ambiguity in natural languages;
- the Semantic Analyzer and Disambiguator, a module that analyzes each sentence identifying roles, meanings of terms and semantic relations in order to extract “part of speech” information, the synonymy and hyponym relations from the WordNet semantic net. It also evaluates terms contained in the document by means of a density function based on the synonyms and hyponyms frequency [7];
- the Classifier, a module that classifies documents automatically. As proposed in WordNet Domains ([8] and [9]), a lexical resource representing domain associations between terms, the module applies a classification algorithm based on the Dewey Decimal Classification (DDC) and associates a set of categories and a weight to each document.

The analysis of structured parts, followed by the linguistic analysis and the semantic interpretation of unstructured parts, produces as results three types of semantic keys:

- a synset ID identifying a particular sense of a term of WordNet
- a category name given by a possible classification
- a key composed by a word and a document category, i.e. when the semantic key related to the word is not included in the WordNet vocabulary.

Finally, all semantic keys are used to index the document whereas in searching phase they are used in order to retrieve document descriptors using the SemanticNet through the concept of semantic vicinity.

2.2 The Specific Context

In a specific context the system adopts a formal or terminological ontology describing the specific semantic domain. The semantic keys are the identifiers of concepts in the specialized ontology and the classification is performed using the taxonomy defined in it.

Different analysis are performed in structured and unstructured parts of the document, as identified in the generic context. Moreover two types of indexing are performed: one with a set of generic semantic keys and the other with a set of specialized semantic keys.

The module extracts all of the specialized semantic keys each time a structured part related to a specific context is identified in a document. Otherwise the system
performs the same analysis as in the general context and adds the result to the generic semantic keys set. Then, the extracted keys have to be mapped into specialized semantic keys. In this way a conceptual mapping between the specific semantics domain ontology and the WordNet ontology is needed in the design phase of the module. However, the conceptual map is not always exhaustive. Each time a concept can not be mapped into a synset ID, it is mapped into a new unique code.

Structured parts of a document related to an explicit semantic (XML, RDF, OWL, etc) and defined by a known formal or terminological ontology are parsed by the system. The results are concepts of the ontology, so that the system has to know their mapping onto a concept of the WordNet ontology. The conceptual mapping of a specialized concept returns a WordNet concept or its textual representation as a code or a set of words. Through the conceptual mapping the relation 'SAME-AS' has been defined, it connects WordNet's synset IDs with concepts of the specialized ontology.

This new relation gives us the possibility to build a specialized Semantic Net (sSN) complying with the following properties:

1. Every sSN node is a concept in the specialized ontology.
2. Every sSN node has at least one sSN relation with another sSN node or it is related with a node in the SemanticNet by means of the 'SAME-AS' relation.
3. A sSN relation has to be consistent to the same relation in the SN. A similar example could be the 'broader' relation described in SKOS [10], which identifies a more general concept in meaning. The 'broader' relation can be mapped to the “IS-A” if the mapping is consistent to the “IS-A” relation derived from WordNet 'hyponymy' relation.

A generic semantic key extracted from the unstructured parts of the document can identify, in this way, a specialized semantic key if a mapping between the two concepts exists.

Once the semantic keys are extracted from the text, the system performs the classification related to the generic and the specific context. To index a document, the system needs to classify it and two set of semantic keys: one related to the generic context and the other one to the specialized context. In the search phase the system retrieves documents by means of a specialized semantic key, for example a URI [11], or a generic semantic key and the other one with a semantic vicinity in the sSN or SN.

An example of specific context module can be found in section 4 where the specific ontology is a terminological ontology for the GIS specific context.

3 Building the SemanticNet

In the context of DART, we think that the answer to an user query can be given providing the user with several kind of results, not always related in the standard way the search engines we use today do. To achieve this goal, we increase the semantic net of WordNet identifying valid and well-founded conceptual relations and links contained in documents in order to build a data structure, composed by concepts and correlation between concepts and information, that allows users to navigate in the

concepts through relations. Formulating the query, the user can move through the net and extract the concepts which really interest him, limiting the search field and obtaining a more specific result. The enriched semantic net can also be used directly by the system without the user being aware of it. In fact, the system receives and elaborates queries by means of the SemanticNet, extracts from the query the concepts and their related relations, then shows the user a result set related to the new concepts found as well as the found categories.

The automatic creation of a conceptual knowledge map using documents coming from the Web is a very relevant problem and it is very hard because of the difficulty to distinguish between valid and invalid contents documents. We therefore realized the importance of being able to access a multidisciplinary structure of documents, so a great amount of documents included in Wikipedia [12] was used to extract new knowledge and to define a new semantic net enriching WordNet with new terms, their classification and new associative relations.

In fact WordNet, as semantic net, is too limited with respect to the web language dictionary. WordNet contains about 150,000 words organized in over 115,000 synsets whereas Wikipedia contains about 1,900,000 encyclopedic information; the number of connections between words related by topics is limited; several word “senses” are not included in WordNet. These are only some of the reasons that convinced us to enrich the WordNet semantic net, as emphasized in [13] where authors identify this and 5 other weaknesses in the WordNet semantic net constitution.

We chose Wikipedia, the free content encyclopedia, excluding other solutions such as language specific thesaurus or on-line encyclopedia available only in a specific language. A conceptual map built using Wikipedia pages allows a user to associate a concept to other ones enriched with some relations that an author points out. The use of Wikipedia guarantees, with reasonable certainty, that such conceptual connection is valid because it is produced by someone who, at least theoretically, has the skills or the experience to justify it. Moreover, the rules and the reviewers controls set up guarantee reliability and objectivity and the correctness of the inserted topics. The reviewers also control the conformity of the added or modified voices.

What we are more interested in are terms and their classification in order to build an enriched semantic net, called “SemanticNet” to be used in the searching phase in the general context, while in the specific context we build a specialized Semantic Net (sSN). The reason for this is that varied mental association of places, events, persons and things depend on the cultural backgrounds of the users’ personal history. In fact, the ability to associate a concept to another is different from person to person. The SemanticNet is definitely not exhaustive, it is limited by the dictionary of WordNet, by the contents included in Wikipedia and by the accuracy of the information given by the system.

Starting from the information contained in Wikipedia about a term of WordNet, the system is capable of enriching the SemanticNet by adding new nodes, links and attributes, such as IS-A or PART-OF relations. Moreover, the system is able to classify the textual contents of web resources, indexed through the Classifier that uses WordNet Domains and applies a density function ([1], [2]), based on the computation of the number of synsets related to each term of the document. In this way, it is able to retrieve the most frequently used “senses” by extracting the synonyms relations given by the use of similar terms in the document. Through the categorization of the
document it can associate to the term the most correct meaning and can assign a weight to each category related to the content.

In fact, each term in WordNet has more than one meaning each corresponding to a Wikipedia page. We therefore need to extract the specific meaning described in the Wikipedia page content, in order to build a conceptual map where the nodes are the “senses” (synset of WordNet or term+category) and the links are given both by the WordNet semantical-linguistic relations and by the conceptual associations built by the Wikipedia authors. For example, the term tiger in WordNet corresponds to the Wikipedia page having http://en.wikipedia.org/wiki/Tiger as URL and the same title as the term, as showed in figure 1.

Such a conceptual map, allows the user to move on an information structure that connects several topics through the semantical-lexical relations extracted from WordNet but also through the new associations made by the conceptual connections inserted by Wikipedia users and extracted by the system.

From each synset a new node of the conceptual map is defined, containing information such as the synsetID, the categories and their weight. Through the information extracted from Wikipedia it is possible to build nodes, having a conceptual proximity with the beginning node, and to define, through these relations, a data structure linking all nodes of the SemanticNet. Each node is then uniquely identified by a semantic key, by the term referred to the Wikipedia page and by the extracted categories.

3.1 The Data Structure

The text of about 60,000 Wikipedia articles corresponding to a WordNet term and independently from their content has been analyzed, in order to extract new relations and concepts, and it has been classified, using the same convention used for the mapping of the terms of WordNet in WordNet Domains, assigning to each category a weight of relevance.

Only the categories having a greater importance are taken in consideration. The success in the assigning the categories has been evaluated to reach about 90%. Measures are made choosing a set of categories and analyzing manually the correctness of the resources classified under each category. Through the content classification of Wikipedia pages, the system assigns to their title a synsetID of WordNet, if it exists.

By analyzing the content of the Wikipedia pages, a new relation named “COMMONSENSE” is defined, that delineates the SemanticNet in union with the semantic relations “IS-A” and “PART-OF” given by WordNet.

The COMMONSENSE relation is a connection between a term and the links defined by the author of the Wikipedia page related to that term. These links are associations that the author identified and put in evidence. Sometimes this relation closes with a direct connection a circle of semantic relations between concepts of the WordNet semantic net. In such situation we consider valid these direct links because someone, in the Wikipedia resource, certified a relation between them. In fact, the vicinity between concepts is justified with the logic expressed from the author in the
SemanticNet: a WordNet-based Tool for the Navigation of Semantic Wikipedia article. The importance of these relations comes out each time the system give back results to a query, providing the user concepts that are related in WordNet or in the Wikipedia derived conceptual map.

In the SemanticNet, relations are also defined as oriented edges of a graph. The inverse relations are then a usable feature of the net. An example is the Hyponym/Hypernym relation, labeled in the SemanticNet as a “IS-A” relation with direction. The concept map defined is still under development and improvement. It is constituted by 25010 nodes each corresponding to a “sense” of WordNet and related concepts.
to a page of Wikipedia. Starting from these nodes, 371281 relations were extracted, part of which are also included in WordNet, but they are mainly new relations extracted from Wikipedia.

In particular 306911 COMMONSENSE relations are identified, where 48834 are tagged as IS-A_WN and 15536 as PART-OF_WN, that are relations identified by the system but also included in WordNet. Sometimes, it is also possible to extract IS-A and PART-OF relations from Wikipedia pages as new semantic relations, not contained in WordNet even if still now they are not added in our structure.

Terms not contained in WordNet and existing in Wikipedia are new nodes of the augmented semantic net. Terms which meanings are different in Wikipedia in respect with the meaning existing in WordNet will be added as new nodes in the SemanticNet but they will not have a synsetID given by WordNet.

In fig.1 a portion of the SemanticNet is described, starting from the node tiger. The text contained in the Wikipedia page corresponding to the term is analyzed and it is classified under the category Animals. In this way the system is able to exclude one of the two senses included in WordNet, the one having the meaning related with person (synsetID: 10012196), and to consider only the sense related with the category Animals (synset ID: 02045461). So, the net is enriched with other nodes having a conceptual proximity with the starting node tiger and the new relations extracted from the page itself as well as the relation included in WordNet can be associated to this specific node by the system. In particular, the green arrow means a COMMONSENSE relation between tiger and the concepts the Wikipedia authors have pointed out by the links. Some of them are terms not contained in WordNet at all (superpredator), others are included in WordNet but in the vocabulary they are not directly related with the term itself (lion). Other terms are included in WordNet and directly related with the term tiger yet (panthera, tigress) by PART-OF, HAS-MEMBER or IS-A relations (blue and red arrows).

Each term showed in the figure and related with the central node is a node itself, showed as a word for simplicity in the representation. It is really a specific sense of the term, related with the node tiger as animal with the specific meaning of the showed term and characterized by a synsetID, or by the couple <term, category> if it is not included in WordNet, by a specific category and by a description that unequivocally identify it.

For example, if the user is interested in information about the place where the greatest percentage of tigers lives but he doesn't know or he can not remember its name, by navigating the SemanticNet he can point out a COMMONSENSE relation between the term tiger and the term Indian subcontinent. Then he can refine the query limiting the search field to resources related with these two concepts.

Some concepts, such as lake, river or swimmer, do not always seem related with the concept tiger. But if we consider that tigers are strong swimmers and that they are often found bathing in ponds, lakes, and rivers, the relations pointed out turn out relevant for the concept itself and useful for users interested in this specific aspect of the animal.
3.2 Evaluation of the Information

Concerning the information added in the SemanticNet we have to distinguish between information gathered from WordNet and information extracted from Wikipedia documents. Properties of the information coming from WordNet are maintained in the structure used in the SemanticNet. Then we add new terms by means of a phase of classification of Wikipedia documents, in order to identify correctly information added in the map of concept.

As we said, COMMONSENSE relations added in the SN are the links contained in the Wikipedia pages related to a particular sense of a term. The main question is to identify which categories are associated with that specific sense. We conducted our tests on sets of documents, extracted from a total set of 47639 documents and evaluating only 5 categories: Plants, Medicine, Animals, Geography and Chemistry. In the evaluation we only consider if the classified document belongs or not to the specified category. The Classifier assigns a number of possible categories to each document with a weight associated. We selected the best result for each document by means of a minimum level of weight. In this way all terms added to the SemanticNet are always related in the correct way with others.

![Fig. 2: Classified documents for each category](image)

In Fig.3 the measures of the Classifier about 5 categories are showed. Results are validated by hand verifying all documents identified by the Classifier for each category.

![Fig. 3: Measures of the classification](image)
4 An Example of Specific Context

An example of a specific context is referred to a GIS context. In order to execute a coherent analysis of documents, two relevant problems have to be resolved. First is the choice of the taxonomy and thesaurus describing the specific semantic domain. The multilingual 2004 GEMET (GEneral Multilingual Environmental Thesaurus, version 1.0) [14] was chosen. GEMET consists in 5,298 terms hierarchically ordered with 109 top terms, 40 themes, 35 groups and 3 super groups.

Then the problem of mapping GEMET concepts into the WordNet concepts. For example, the GEMET term "topography" has the right meaning in the WordNet semantic net, while the term "photogrammetry" does not appear at all. In such cases, and generally in all specialized contexts or specific domains, we need to define a new semantic key for this kind of term that is the identifier used in the GEMET thesauri. In order to generate the conceptual mapping, a semi-automatic procedure was implemented. This procedure is very similar to the evaluation of the Wikipedia page for enriching the SemanticNet and it is based on two similar properties of GEMET and WordNet: GEMET and WordNet have a hierarchical structure. For example, the GEMET relations "narrower" and "broader" are similar to the "hypernym" and "hyponym" relations of WordNet. This similarity is also used to build the relation "IS-A" in the sSN. Another important property is the presence of textual description of concepts in both GEMET and WordNet.

Starting from the top term in the GEMET thesauri, the semantic keys from the textual description of a concept are extracted and its semantic vicinity is calculated with semantic keys extracted from the textual descriptions of concepts found in WordNet using the term related to the GEMET concept. The presence of some concepts such as the "narrower/broader" of the GEMET concept mapped to concepts "hyperonym/hyponym" of the concepts found in WordNet are also evaluated. The final evaluation of results was performed by a supervisor. A similar approach can be found at [15]. If the GEMET concept is not found in WordNet, the term related to the GEMET concept is used as the mapped semantic key.

Using such mapping makes it possible to classify the documents analysed in the generic context with the GEMET taxonomy and to generate the sSN which contains the relations "IS-A", derived from the "narrower" and "broader" GEMET relations, and the 'SAME-AS' relation, described in the section 2, which is derived from the conceptual mapping and connects the sSN node with the SN nodes. An example is showed in figure 4.

5 3D User Interface for SemanticNet

As described before, the SemanticNet allows users to navigate in the concepts through relations. This feature include the need to overcome the limits of traditional user interface, specially in terms of effectiveness and usability. For this reason, in parallel with the development of the SemanticNet we investigate new paradigm and model of UI, focusing on 3D visualization to allows the user to change the point of view, improving the perception and the understanding of contents [16]. The main idea is to
improve the user experience during the phases of searching and extracting of the concepts. An essential requirement for such goal is to guarantee a better usability in information navigation by concrete representations and simplicity [17]. We develop an alternative tool (3DU4SemanticNet) for the browsing of the Web resources by means of the map of concept. Formulating the query through the search engine, the user can move through the SemanticNet and extract the concepts which really interest him, limiting the search field and obtaining a more specific result.

This tool works according to the user preferences and current user context, providing a 3D interactive scene that represent the selected portion of SemanticNet. Moreover, it provides a three-dimensional view that guarantee a better usability in

![SemanticNet Diagram](image)

**Fig. 4:** A segment of the specialized SemanticNet relative to the concept “Tiger”.

Term contained in WordNet

Term contained in GEMET
terms of information navigation, and it could also provide different layouts for different cases [18] and [19]. By example, a user could navigate through the net, moving from a term to another looking for relations and additional information. In order to keep the user into a Web context, the scene is described by an X3D document [20][21]. This choice has been driven by the features of this language, specially because it is a standard based on XML, and it is the ISO open standard for 3D content delivery on the web supported by a large community of users. The X3D runtime environment is the scene graph which is a directed, acyclic graph containing the objects represented as nodes and object relationships in the 3D world. The basic structure of X3D documents is very similar to any other XML document. In our application this structure is built starting from a description of the relations between terms. This description is provided by a GraphML [18] file, that represent the structural properties of SemanticNet.

As X3D, GraphML is based on XML and unlike many other file formats for graphs, it does not use a custom syntax. Its main idea is to describe the structural properties of a graph and a flexible extension mechanism to add application-specific data; from our point of view, this fits the need to describe a portion of the SemanticNet starting from the singular nodes. The following figure shows the X3D model built from the GraphML provided for a part of the SemanticNet defined for the term “tiger”, disambiguated and assigned to the “Animals” WordNet Domains category.

Fig. 5: X3D representation for the term “tiger” described by GraphML
6 Conclusions and Future Work

In this paper we have described the SemanticNet as part of the project DART, Distributed Agent-Based Retrieval Toolkit, currently under development. We have focused the efforts into the idea to provide a user friendly tool, able to reach and filter relevant information by means of a conceptual map based on WordNet. One of the main aspect of this work has been the right classification of resources. It has been very important in order to enrich the WordNet semantic net with new contents extracted from Wikipedia pages and with concepts coming from the GEMET thesaurus.

The SemanticNet is structured as a highly connected directed graph. Each vertex is a node of the net and the edges are the relations between them. Each element, vertex and edge, is labeled in order to give the user a better usability in information navigation even through a dedicated 3D tool.

Future works will include the improvement of the SemanticNet, by extracting new nodes and relations, and the measuring of the user preferences in the navigation of the net in order to give a weight to the more used paths between nodes.

Moreover the structure of nodes, as defined in the net, allows to access the glosses given by WordNet and Wikipedia contents. The geographic context gives the user further filtering elements in the search of Web contents. In order to make the implementation of modules of a specialized context easier, its conceptual mapping and the definition of the specialized semantic net, a future work will describe the SemanticNet with a simple formalism like SKOS. Therefore the system could be more flexible in indexing and searching of web resources.

References


10. SKOS, Simple Knowledge Organisation Systems, http://www.w3.org/2004/02/skos/


Verification of Valency Frame Structures by Means of Automatic Context Clustering in RussNet

Irina V. Azarova¹, Anna S. Marina², and Anna A. Sinopalnikova³

¹ Department of Applied Linguistics, St-Petersburg State University, Universitetskaya nab. 11, 199034 St-Petersburg, Russia.
² Department of Lexicography, Institute of Linguistic Studies, Tuchkov pereulok 9, 199053 Saint-Petersburg, Russia.
³ Brno University of Technology, Bozetechova 2, 61266 Brno, Czech
ivazarova@gmail.com, a_s_marina@rambler.ru, sino@fit.vutbr.cz

Abstract. The major point of the RussNet technique is a specification of valency frames for synsets. Parameters of valency frames are employed for word meaning and synsets differentiating in the procedure of thesaurus construction and automatic text analysis for word disambiguation. Valency description is calculated on the basis of statistically stable context features in the text corpus: morphologic, syntactic, and semantic. The automatic classification of verb contexts with unambiguous morphology annotation is discussed in the paper. The goal of this technique is differentiation of semantic types for verbs. The procedure is fulfilled with a help of morphology tag distributions in some context window for verbs from different semantic trees of RussNet. The optimal width of a distribution window, an appropriate tag set, and clustering results are discussed. This procedure may be helpful at various stages of analysis, especially for valency frame verification in some semantic tree.

1 Introduction

The computer thesaurus RussNet¹ developed at the Department of Applied Mathematic Linguistics of Saint-Petersburg State University inherited the main principles of WordNet construction method [1]. RussNet is based on the corpus of modern texts (dated from 1985 up to nowadays) including 21 million of words, the major part of which (60%) are articles on various topics from newspapers and magazines, covering thematic diversity of the common Russian language [2].

The RussNet was not translated from the WordNet prototype, its construction involves some additional components in its structure, that were oriented to its usage in automatic text analysis [3].

The basic node in RussNet – the synset – may include several members (words or multiword expressions), which are ordered by their frequency of appearance in the corpus contexts in the particular sense described by the synset. This frequency is

¹ http://www.phil.pu.ru/depts/12/RN
measured in ipm-s. In order to fix frequency distribution of a polysemous word we use manual marking up of meanings (WM) in the contexts of a random corpus sample. We investigated the necessary size of a sample, and learnt that a random sample of 100–150 contexts represents the same distribution of WMs characteristic for the whole set of contexts in the corpus. The possible error of portion frequencies hardly exceeds 1% (which may be, however, crucial for scarce WMs). Our results coincide with previous investigations mentioned by [4], in which the marking up of WMs in the samples were compared.

The next issue for investigation was whether there is some particular type of word meaning frequencies distribution. As WM frequencies may be so similar that the least distortion may be undesirable. We found out that the most common distribution of WMs in the corpus (for polysemious words) is rather specific: in approximately 80% cases such word has a distinct first meaning, which is the most frequent and usually occurs in 50–70% contexts of the corpus. There are so called low-frequency WMs (occurring in 1–3% of contexts), which are unrealistic to order according their frequencies. Other WMs fill the array between the first and low-frequency in a manner of decreasing numbers.

The procedure of marking up WMs in a corpus sample requires calculation of frequencies for grammatical context markers and listing of semantic classes of words on the suggested valency positions in order to substantiate making decision concerning word senses differentiation. Stable context markers are included into valency frames expanding standard WordNet structure. Below valency frame specification is described in detail.

## 2 Valency Frame Specification

The idea of extending a computer lexicon by syntagmatic information is rather common in various WordNet-dictionaries as well as traditional ones [5], [6], [7].

In our project we compile valency frame description on the basis of stable features in the marked up contexts from random corpus samples [3]. The general parameters of the valency frame are

- its active or passive attribute, which reflects a syntactic position of the described synset as a head (dominant) element or a daughter (subordinate) one;
- the attribute of a syntactic construction – predicative or attributive – in which the valency frame primarily occurs.

The valency frame may include several valencies with following parameters

- its obligatory or optional attribute, which correlates with the frequency of valency occurrence in the corpus: prevailing (100% \( \leq f \leq 66\% \)), rather stable (65% \( \leq f \leq 35\% \)); less stable (<35%) valencies are considered unreliable and are not listed;
- the semantic function of a valency position (so called semantic role), which in RussNet is interpreted as an argument position in the whole set of arguments involved in some semantic tree (e.g. verbs of movement).

The valency may have several ways of manifestation, each of them is characterised by

- its semantic scope in terms of semantic trees, subtrees, groups of trees in RussNet;
Verification of Valency Frame Structures by Means of Automatic…

- morphological and syntactic features, i.e. frequent surface expression (e.g. the particular preposition for nouns, the aspect form for infinitive, an adverb from the particular group, etc.).

The described structure is used for automatic disambiguation [3]: the parsing of the phrase or sentence is mapped against valency frames of words in the construction. In case of matching between a parse structure and obligatory valency frames, it is considered verified, otherwise, optional valency frames may show preferred analysis.

The open question is whether inheritance of valency frame parameters in troponymic verbal trees exists, in what form and for what parameters. We had a preliminary research basing on three semantic verb trees in RussNet [8] and, though didn’t receive unequivocal confirmation of the inheritance scheme, it may be stated that context features of the major part of verbs from a particular semantic tree share a lot of statistically stable parameters.

In order to prove this stability we investigated the automatic clustering of verb contexts as representatives of WMs in RussNet. This research is presented below.

3 Motivation: How to Use a Morphological Marking up of the Text for Sense Disambiguation

In our research we chose for a starting point the approach of [4]. In this work the disambiguation procedure for a verb serve was described, in which 3 distributions were used: main POS tags, additional markers (e.g. punctuation marks, prepositions, etc.), and lexical items. This investigation demonstrated that POS tags and some other features afford to reliably (80–83%) differentiate meanings of this polysemious verb.

The results depend on the width of the analysis window, and the substantial amount of contexts in the training set.

The authors drew conclusions comparing their approach with similar ones
1. initial processing of the text (e.g. syntactically connected fragments) doesn’t affect the results crucially;
2. it was unachievable to differentiate low frequency WMs, because it was hardly possible to compile quality training set;
3. it was easier to differentiate homonyms (or contrast WMs) than similar WMs;
4. the huge amount of a training set improves the results but not to the same extent as processing time increased for preparing these sets.

3.1 Preliminary Results: the POS Tag Distribution for Different Semantic Verb Groups

The mentioned research might be valid only for languages with a fixed word order, and hardly applicable for languages with freer word order. We decided for the beginning to fulfil “pilot” check: compare distributions of 9 verbs from two groups: verbs of movement (идти ‘to go’, пойти ‘to start walking’, выйти ‘to go out’, вернуться ‘to return’, ходить ‘to walk’) and communication verbs (сказать ‘to say’, говорить ‘to speak’, спросить ‘to ask’, ответить ‘to answer’, просить ‘to
We selected 200 contexts from the corpus per verb in its first meaning. The width of the analysis window was chosen [-6…+6], its zero position being a key verb form and other positions being occupied by neighbouring words and punctuation marks.

**Fig. 1.** A correlated distribution of frequencies for the preposition tag in the contexts of travel verbs.

We compared 3 distributions for: (1) punctuation marks; (2) main POS: nouns, adjectives, verbs, adverbs; (3) close classes and syntactic POS: pronouns & numerals, conjunctions, prepositions, particles. After distribution calculated, some tags appeared to be too scarce (< 5% occurrences), they were eliminated from further investigation.

**Fig. 2.** An uncorrelated distribution of frequencies for the adjective tag in the contexts of communication verbs.

The distribution for a tag was presented as a graph defined over the analysis window \(i=[-6...+6]\), \(f_i\) showing the overall occurrence of the described tag in the \(i\)-th position of all training contexts for a particular verb. The average for the group was calculated. The graphs show correlation of tag distributions in groups or its absence. It appeared that some tags have specific distribution throughout the whole window, e.g. nouns, verbs, pronouns, commas, quotation marks, colons, dashes – for communication verbs, and nouns, adverbs, pronouns, conjunctions, prepositions, commas – for travel verbs. Some tags have a particular distribution only in the right
part of the context (e.g. conjunctions for communication verbs) or left context (e.g. prepositions for communication verbs). On Fig. 1 the correlated distribution of prepositions in contexts of verbs of movement is shown, on Fig. 2 – the uncorrelated distribution of adjectives for verbs of communication.

Distribution data showed that it is feasible to use them for automatic differentiating of some groups of verbs, however, the width of the distribution analysis window and the appropriate tag set should be examined in detail.

4 The Optimal Width of the Distribution Window

The research reported in this paper involves 51 frequent Russian verbs from 21 semantic groups taken from [9]. Each verb was represented by 200 contexts chosen at random from our corpus, which were unambiguously marked up with morphological tags. At first we chose the maximal window of [-10…+10] positions and the tag set including POS marker plus case specification for substantives and the aspect value for verbs (e.g. Nnom, Aloc, Vperf, etc). Punctuation marks are represented by one tag PM. Tag distributions are calculated in all positions of the maximum window for the contexts of each verb, thus each position was represented by the vector of tag frequencies. If some tag doesn’t occur in i-th position, its frequency is zero.

Distributions were compared according to the vector model [10], [11] with the cosine similarity. For example, in i-th position in the window and distributions $a$ and $b$ the similarity is equal to:

$$\text{sim}(a, b) = \frac{\sum_{N} a_{ij} \times b_{ij}}{\sqrt{\sum_{N} a_{ij}^2 \times \sum_{N} b_{ij}^2}}$$

(1)

In Tab. 1 a fragment of the positional similarity matrix for verbs is shown, similarity is measured in per cents.

<table>
<thead>
<tr>
<th>Verb1</th>
<th>Verb2</th>
<th>all</th>
<th>-10</th>
<th>-5</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>брать 'to take'</td>
<td>иметь 'to have'</td>
<td>81</td>
<td>95</td>
<td>93</td>
<td>88</td>
<td>57</td>
<td>14</td>
<td>85</td>
<td>94</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>казаться 'to appear'</td>
<td>84</td>
<td>97</td>
<td>94</td>
<td>90</td>
<td>63</td>
<td>39</td>
<td>88</td>
<td>94</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>идти 'to go'</td>
<td>88</td>
<td>96</td>
<td>89</td>
<td>93</td>
<td>90</td>
<td>58</td>
<td>91</td>
<td>92</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>хотеть 'to wish'</td>
<td>91</td>
<td>93</td>
<td>95</td>
<td>94</td>
<td>87</td>
<td>83</td>
<td>85</td>
<td>91</td>
<td>94</td>
</tr>
<tr>
<td>быть 'to be able'</td>
<td>моющий 'to be able'</td>
<td>82</td>
<td>93</td>
<td>84</td>
<td>89</td>
<td>86</td>
<td>45</td>
<td>65</td>
<td>92</td>
<td>89</td>
</tr>
</tbody>
</table>
It is easily seen that the “distant” positions of the distributions look very similar, thus they are non-specific for any verb group.

Fig. 3. The stemma of verb clustering in the [-10]th position of the distribution window.

In order to visualise results of similarity measurement we used automatic cluster analysis [12]. We represent the results of clustering on stemmas (see Fig. 3–5), which reflect the order of grouping in reverse order: leaves are the “closest” verbs, and the whole verb group is shown at the root node.

Fig. 4. The stemma of verb clustering in the [+1]st position of the distribution window.
It is possible to assess the clustering quality so that to choose the optimal window width (and other parameters). We consider clustering to be interpretable if verbs from the same semantic tree are put together in one cluster, and inexplicable – if there are a lot of representatives from different semantic groups mixed in one cluster. If there were not morphological correlations among contexts of different verbs with similar or relative meanings, we’d never received any explicable groupings.

Results of [4] show that a range of positions in the window has a cumulative outcome. So we compared verb distributions per each position and per position range and received expected result: there were no separate position in the window [-10...+10], which was sufficient by itself to show reliable clustering. On Fig. 3, 4, 5 there are stemmas for clustering in the [-10] position, [+1] position and the better range [-3,+5]. Grey background marks verbs from the same semantic group, numbers at the top nodes show the step of clustering and an average similarity

5 An Optimal Tag Set for Distribution Capture

The tag set (TS) is another key point of distribution description. The first variant of the tag set showed interpretable results, however, it was significant to what extent it may deviate clustering. We tried 3 tag sets: the 1st TS was described above, the 2nd TS was simple POS tagging without specification of grammar category values (e.g. N, A, Adv, V, Pron, etc.); the 3rd TS was a kind of POS tag generalisation: all substantives were united under one tag, however, case specification for substantives were added (Nom, Gen, Dat, Acc, Abl, Loc, V, Adv, etc.). The comparison of clustering parameters shows that clustering with the 2nd TS produces a flatter structure without elaboration of the inner structure in groups, the clustering with the 3rd TS creates more detailed structure than the 1st TS.

Fig. 5. The stemma of verb clustering in the [-3,+5] position range of the distribution window with the 1st TS
6 The Structure of Clusters

The clustering of verbs with the help of morphological tagging of contexts, distribution window [-3,+5] and 1st or 3rd TSs afford us to differentiate 13 groups, in which 70% of verbs are united. Some groups are very close to WordNet classification into semantic trees: verbs of communication, cognition, stative, motion, emotion, possession, contact, modal, creation, perception. Moreover, we tried to use united morphological distributions as centres of clusters, and receive interesting structure for further cluster structuring. For example, communication and cognition verbs are very close and united at the very early stage of clustering, the same for possession and contact verbs. It is interesting that verbs – aspect pairs are united at one step, though we reported in our paper [8] that they very often have different meaning.

1. сказать ‘to say’, ответить ‘to answer’, спросить ‘to ask’;
2. понимать ‘to understand, imperf.’, знать ‘to know’, понять ‘to understand, perf.’, помнить ‘to remember’, думать ‘to think’;
3. сидеть ‘to sit’, лежать ‘to lie’, стоять ‘to stand’;
4. взять ‘to take, perf.’, брать ‘to take, imperf.’, получить ‘to receive’, иметь ‘to have’;
5. идти ‘to go’, ехать ‘to ride/drive’, пойти ‘to start walking’;
6. ненавидеть ‘to hate’, любить ‘to love’, чувствовать ‘to feel’;
7. бросить ‘to throw’, послать ‘to send’;
8. мочь ‘can/to be able’, успеть ‘to manage/succeed’, хотите ‘want/to wish’;
9. делать ‘to do, imperf.’, сделать ‘to do, perf.’;
10. видеть ‘to see, imperf.’, увидеть ‘to see, perf.’;
11. жить ‘to live’, работать ‘to work’;
12. дать ‘to give, perf.’, давать ‘to give, imperf.’;
13. остаться ‘to stay’, оказаться ‘to be found’.

7 Conclusion and Future Work

The reported research shows that it is highly probable to differentiate verbs from different semantic groups with the help of morphology tagging. This procedure may be helpful at the preliminary stages of corpus contexts processing, at the stage of valency frame verification for some semantic tree, and during automatic text processing. For different purposes it is essential to formulate the similarity measure of a particular context with cluster patterns.

In order to understand why some verbs were not clustered according to their semantic type, we examined these cases, and discovered that primarily the failure was connected with random character of the training set. It was supposed that due to the mentioned above prevailing of the first WMs in the corpus contexts, we receive representation of the semantic class with low noise. But in cases of contrast meaning when verb meanings appear rather as homonyms, than similar WMs, we received a mixed distribution. After refining some of them, we saw appropriate clustering results.
Another perspective of this approach is a hypothesis that other main POS may be processed in the same manner. Now we investigate similar distributions for nouns.

References

Some Issues in the Construction of a Russian WordNet Grid

Valentina Balkova\textsuperscript{2}, Andrey Sukhonogov\textsuperscript{1}, and Sergey Yablonsky\textsuperscript{1,2}

\textsuperscript{1} Petersburg Transport University, Information Systems Department, Moscow av., 9, St.-Petersburg, 190031, Russia
\textsuperscript{2} Russicon Company, Kazanskaya str., 56, ap.2, 190000, Russia
v_balk@front.ru, asukhonogov@rambler.ru, serge_yablonsky@hotmail.com

Abstract. This paper deals with the development of the Russian WordNet Grid. It describes usage of Russian and English-Russian lexical language resources and software to process Russian WordNet Grid for Russian language and design of a XML-markup of the grid resources. Relevant aspects of the DTD/XML format and related technologies are surveyed.

1 Introduction

The Semantic Web aims to add a machine tractable layer to compliment the existing web of natural language hypertext. In order to realise this vision, the creation of semantic annotation, the linking of web pages to ontologies, and the creation, evolution and interrelation of ontologies must become automatic or semi-automatic processes.

Computational lexicons (CL) provide machine understandable word knowledge. That is important for turning the WWW into a machine understandable knowledge base Semantic Web. CL supply explicit representation of word meaning with word content accessible to computational agents. Word meaning in CL is linked to word syntax and morphology and has multilingual lexical links.

Computational lexicons are key components of HLT and usually have such typology:

- monolingual vs. multilingual;
- general purpose vs. domain (application) specific;
- content type (morpho-syntactic, semantic, mixed, terminological).

Today such types of CL are designed

- network based (hierarchy/taxonomy — WordNet \cite{1, 2, 10}, heterarchy — EuroWordNet \cite{3});
- frame based (Mikrokosmos, FrameNet);
- hybrid (SIMPLE).

Application of WordNet for different tasks on the Semantic Web requires a representation of WordNet in RDF and/or OWL \cite{4-6}. There are several conversions available (from WordNet’s Prolog format to RDF/OWL) which differ in design choices and scope. It is expected that the demand for WordNet in RDF/OWL will
grow in the coming years, along with the growing number of Semantic Web applications.

The WordNet Task Force of the W3C’s Semantic Web Best Practices Working Group aims at providing a standard conversion of WordNet. There are two main motivations that support the development of a standard conversion:

- the development through the W3C’s Working Group process results in a peer-reviewed conversion that is based on consensus of the participating experts. The resulting standard provides application developers with a resource that has the desired level of quality for most common purposes;
- a standard improves interoperability between applications and multilingual lexical data [10].

Semi-automatic integration and enrichment of large-scale multilingual lexicons like WordNet is used in many computer applications. Linking concepts across many lexicons belonging to the WordNet-family started by using the Interlingual Index (ILI). Unfortunately, no version of the ILI can be considered a standard and often the various lexicons exploit different version of WordNet as ILI.

At the 3rd GWA Conference in Korea there was launched the idea to start building a WordNet grid around a Common Base Concepts expressed in terms of WordNet synsets and SUMO definitions (http://www.globalwordnet.org/gwa/gwa_grid.htm). This first version of the Grid was planned to be build around the set of 4689 Common Base Concepts. Since then only three languages with essentially various number of synsets and different WordNet versions were placed in the Grid mappings (English – 4689 synsets with WN 2.0 mapping, Spanish – 15556 synsets with WN1.6 mapping and Catalan - 12942 synsets with WN1.6 mapping). But there is yet no official format for the Global WordNet Grid. So far there are just only 3 files in the specified format. As alternative another possible solution can use the DTD from the Arabic WordNet: http://www.globalwordnet.org/AWN/DataSpec.html.

This paper deals with the development of the Russian WordNet Grid. It describes usage of Russian and English-Russian lexical language resources and software to process WordNet Grid for Russian language (4600 synsets with WN 2.0 mapping) and design of a XML/RDF/OWL-markup of the grid resources. Relevant aspects of the DTD/XML/RDF/OWL formats and related technologies are surveyed.

2 Conceptual model

The three core concepts in WordNet are the synset, the word sense and the word. Words are the basic lexical units, while a sense is a specific sense in which a specific word is used. Synsets group word senses with a synonymous meaning, such as {car, auto, automobile, machine, motorcar} or {car, railcar, railway car, railroad car}. There are four disjoint types of synset, containing exclusively nouns, verbs, adjectives or adverbs. There is one specific type of adjective, namely an adjective satellite.

Furthermore, WordNet defines seventeen relations, of which
- ten between synsets (hyponymy, entailment, similarity, member meronymy, substance meronymy, part meronymy, classification, cause, verb grouping, attribute);
Valentina Balkova, Andrey Sukhonogov, and Sergey Yablonsky

- five between word senses (derivational relatedness, antonymy, see also, participle, pertains to);
- “gloss” (between a synset and a sentence);
- “frame” (between a synset and a verb construction pattern).

In the Table 1 the set of relations in different WordNet realization are summarized, where S – any synset, N – noun synset, V – verb synset, A – adjective synset, R – adverb synset, WS – any word sense, NS – noun sense, VS – verb sense, AS – adjective sense, RS – adverb sense.

3 XML structure of the Russian WordNet Grid

Several Russian lexical resources were used for the Russian WordNet Grid development and the test version of the English-Russian WordNet [7]. We’ve done porting of the original English and Russian WordNet into XML using the DTD for the XML structure from http://www.globalwordnet.org/gwa/gwa_grid.htm and the DTD from the Arabic Wordnet: http://www.globalwordnet.org/AWN/DataSpec.html;

The standard DTD for the Russian grid XML structure and the English/Russian XML format for the Grid is shown on Fig. 1, 2. The grid of English and Russian local WordNets is realized as a virtual repository of XML databases accessible through web services. Basic services devoted to the management of the actual versions of Princeton and Russian WordNets.

Unfortunately, no version of the grid can be considered a standard because the various grids exploit different versions of WordNet, have different numbers of entries and there is no mappings of the multilingual grids on new versions of WordNet.

4 RDF/OWL structure of the Russian WordNet Grid


We’ve done porting of the original English and Russian WordNet Grid into RDF (Resource Description Framework) and OWL (Ontology Web Language). All specific Russian WordNet classes/properties (Table 2) are defined in another name space – rwn (in Princeton WordNet we have wn name space).

Still there are open issues how to support different versions of WordNet in XML/RDF/OWL and how to define the relationship between them and how to integrate WordNet with sources in other languages. Although again the TF did not focus on solving this problem as it is out of scope, we have tried to take this into account in our design, e.g. by making Words separate entities with their own URI. This allows them to be referenced directly and related to structures representing words in other RDF/OWL sources.
### Table 1. Relations between synsets in the different WordNet realizations (Russian WordNet, Princeton WordNet and EuroWordNet)

<table>
<thead>
<tr>
<th>N</th>
<th>Relation</th>
<th>Part of speech</th>
<th>Russian WordNet</th>
<th>Princeton WordNet</th>
<th>EuroWordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>N-&gt;N hyponymOf</td>
<td>@</td>
<td>HAS_HYPERONYM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N-&gt;N hasHyponym</td>
<td>~</td>
<td>HAS_HYPERSONY</td>
</tr>
<tr>
<td>1.</td>
<td>Hyponym</td>
<td></td>
<td>V-&gt;V troponymOf</td>
<td>@</td>
<td>HAS_HYPERONYM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>V-&gt;V hasTroponym</td>
<td>~</td>
<td>HAS_HYOONYM</td>
</tr>
<tr>
<td>2.</td>
<td>Troponym</td>
<td></td>
<td>N-&gt;N hasMeronym</td>
<td></td>
<td>HAS_MERONYM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N-&gt;N hasMemberMeronym</td>
<td>#m</td>
<td>HAS_MERO_MEMBER</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N-&gt;N hasSubstanceMeronym</td>
<td>#s</td>
<td>HAS_MERO_PORTION</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N-&gt;N hasPartMeronym</td>
<td>#p</td>
<td>HAS_MERO_PART</td>
</tr>
<tr>
<td></td>
<td></td>
<td>meronymOf</td>
<td></td>
<td></td>
<td>HAS_HOLONYM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>memberMeronymOf</td>
<td></td>
<td></td>
<td>HAS_HOLO_MEMBER</td>
</tr>
<tr>
<td></td>
<td></td>
<td>substanceMeronymOf</td>
<td>%s</td>
<td></td>
<td>HAS_HOLO_PORTION</td>
</tr>
<tr>
<td></td>
<td></td>
<td>partMeronymOf</td>
<td></td>
<td></td>
<td>HAS_HOLO_PART</td>
</tr>
<tr>
<td>3.</td>
<td>Meronymy</td>
<td></td>
<td>N-&gt;N attribute</td>
<td>=</td>
<td>xpos_hyponym</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A-&gt;N valueOf</td>
<td>=</td>
<td>xpos_hyponym</td>
</tr>
<tr>
<td>4.</td>
<td>Attribute</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Derivation</td>
<td>S&lt;-&gt;S</td>
<td>relatedForm</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>-------------</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-&gt;S</td>
<td>domainCategory</td>
<td>;c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-&gt;S</td>
<td>domainCategoryMember</td>
<td>-c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-&gt;S</td>
<td>domainRegion</td>
<td>;r</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-&gt;S</td>
<td>domainRegionMember</td>
<td>-r</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-&gt;S</td>
<td>domainUsage</td>
<td>;u</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-&gt;S</td>
<td>domainUsageMember</td>
<td>-u</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6. DomainLabel</th>
<th>S&lt;-&gt;S</th>
<th>nearAntonym</th>
<th>NEAR_ANTONYM</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS&lt;-&gt;WS</td>
<td>Antonym</td>
<td>! ANTONYM</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7. Antonymy</th>
<th>V&lt;-&gt;V</th>
<th>sameGroupAs</th>
<th>$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>8. VerbGroup</th>
<th>V&lt;-&gt;V</th>
<th>isSubeventOf</th>
<th>IS_SUBEVENT_OF</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-&gt;V</td>
<td>hasSubevent</td>
<td>HAS_SUBEVENT</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>9. Entailment</th>
<th>V-&gt;V</th>
<th>causes</th>
<th>CAUSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-&gt;V</td>
<td>isCausedBy</td>
<td>IS_CAUSED_BY</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10. Causation</th>
<th>WS&lt;-&gt;WS</th>
<th>seeAlso</th>
<th></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>11. AlsoSee</th>
<th>WS&lt;-&gt;WS</th>
<th>isDerivedFrom</th>
<th>IS_DERIVED_FROM</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS&lt;-&gt;WS</td>
<td>hasDerived</td>
<td>HAS_DERIVED</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>12. Derived</th>
<th>A&lt;-&gt;A</th>
<th>similarTo</th>
<th>&amp;</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>13. SimilarTo</th>
<th>WS&lt;-&gt;WS</th>
<th>participleOf</th>
<th>&lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS&lt;-&gt;WS</td>
<td>hasParticiple</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. 1 DTD for the Russian grid
### Table 3. Specific Russian WordNet classes/properties

<table>
<thead>
<tr>
<th>№</th>
<th>Class</th>
<th>Property</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Word</td>
<td>&amp;rdfs;Literal &amp;wnr;vowelPosition</td>
<td>Position of the stress for every lemma in Russian WordNet.</td>
</tr>
<tr>
<td>2.</td>
<td>Word</td>
<td>&amp;xsd;nonNegativeInteger &amp;wnr;paradigmID</td>
<td>Lemma’s paradigm number. One lemma in general has many paradigms.</td>
</tr>
<tr>
<td>3.</td>
<td>WordSense</td>
<td>&amp;rdfs;Literal &amp;wnr;glossaryWord</td>
<td>Russian WordNet has glossaries for every word.</td>
</tr>
<tr>
<td>4.</td>
<td>WordSense</td>
<td>&amp;xsd;nonNegativeInteger &amp;wnr;senseNumber</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>WordSense</td>
<td>&amp;xsd;nonNegativeInteger &amp;wnr;synsetPosition</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>WordSense</td>
<td>&amp;rdfs;Literal &amp;wnr;styleMark</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>WordSense</td>
<td>&amp;rdfs;Literal &amp;wnr;isDominant</td>
<td>Dominant property.</td>
</tr>
<tr>
<td>8.</td>
<td>WordSense</td>
<td>#WordSense/#Idiom &amp;wnr;hasIdiom</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Idiom</td>
<td>&amp;rdfs;Literal &amp;wnr;idiom</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>Idiom</td>
<td>&amp;rdfs;Literal &amp;wnr;idiomDefinition</td>
<td></td>
</tr>
</tbody>
</table>
Table 4. Equivalent Classes

<table>
<thead>
<tr>
<th>№</th>
<th>W3C RDFS</th>
<th>Russian WordNet OWL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>SynSet</td>
<td>Synset</td>
</tr>
<tr>
<td>2.</td>
<td>NounSynSet</td>
<td>Noun</td>
</tr>
<tr>
<td>3.</td>
<td>VerbSynSet</td>
<td>Verb</td>
</tr>
<tr>
<td>4.</td>
<td>AdjectiveSynSet</td>
<td>Adjective</td>
</tr>
<tr>
<td>5.</td>
<td>AdverbSynSet</td>
<td>Adverb</td>
</tr>
</tbody>
</table>

Table 5. Russian WordNet OWL

<table>
<thead>
<tr>
<th>№</th>
<th>Russian WordNet (OWL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Synset</td>
</tr>
<tr>
<td>2.</td>
<td>owl:ObjectProperty</td>
</tr>
<tr>
<td>3.</td>
<td>glossaryEntry</td>
</tr>
<tr>
<td>4.</td>
<td>exampleSentences</td>
</tr>
<tr>
<td>5.</td>
<td>owl:TransitiveProperty</td>
</tr>
<tr>
<td>6.</td>
<td>hasHyponym</td>
</tr>
<tr>
<td>7.</td>
<td>nearAntonym</td>
</tr>
<tr>
<td>8.</td>
<td>seeAlso</td>
</tr>
<tr>
<td>9.</td>
<td>relatedForm</td>
</tr>
<tr>
<td>10.</td>
<td>Noun</td>
</tr>
<tr>
<td>11.</td>
<td>Verb</td>
</tr>
<tr>
<td>12.</td>
<td>Adjective</td>
</tr>
<tr>
<td>13.</td>
<td>Adverb</td>
</tr>
<tr>
<td>14.</td>
<td>AdjectiveSatellite</td>
</tr>
<tr>
<td>15.</td>
<td>owl:ObjectProperty</td>
</tr>
<tr>
<td>16.</td>
<td>hasMeronym</td>
</tr>
<tr>
<td>17.</td>
<td>memberMeronymOf</td>
</tr>
<tr>
<td>18.</td>
<td>hasMemberMeronym</td>
</tr>
<tr>
<td>19.</td>
<td>substanceMeronymOf</td>
</tr>
<tr>
<td>20.</td>
<td>hasSubstanceMeronym</td>
</tr>
<tr>
<td>21.</td>
<td>partMeronymOf</td>
</tr>
<tr>
<td></td>
<td>Description</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>22.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>hasPartMeronym</code></td>
</tr>
<tr>
<td></td>
<td>#Noun/#Noun</td>
</tr>
<tr>
<td>23.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>isCausedBy</code></td>
</tr>
<tr>
<td></td>
<td>#Verb/#Verb</td>
</tr>
<tr>
<td>24.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>causes</code></td>
</tr>
<tr>
<td></td>
<td>#Verb/#Verb</td>
</tr>
<tr>
<td>25.</td>
<td><code>owl:SymmetricProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>sameGroupAs</code></td>
</tr>
<tr>
<td></td>
<td>#Verb/#Verb</td>
</tr>
<tr>
<td>26.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>isDerivedFrom</code></td>
</tr>
<tr>
<td></td>
<td>#WordSense/#WordSense</td>
</tr>
<tr>
<td>27.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>hasDerived</code></td>
</tr>
<tr>
<td></td>
<td>#WordSense/#WordSense</td>
</tr>
<tr>
<td>28.</td>
<td><code>owl:TransitiveProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>isSubeventOf</code></td>
</tr>
<tr>
<td></td>
<td>#Verb/#Verb</td>
</tr>
<tr>
<td>29.</td>
<td><code>owl:TransitiveProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>hasSubevent</code></td>
</tr>
<tr>
<td></td>
<td>#Verb/#Verb</td>
</tr>
<tr>
<td>30.</td>
<td><code>owl:SymmetricProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>similarTo</code></td>
</tr>
<tr>
<td></td>
<td>#Adjective/#Adjective</td>
</tr>
<tr>
<td>31.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>attribute</code></td>
</tr>
<tr>
<td></td>
<td>#Noun/#Adjective</td>
</tr>
<tr>
<td>32.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>valueOf</code></td>
</tr>
<tr>
<td></td>
<td>#Adjective/#Noun</td>
</tr>
<tr>
<td>33.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>domainUsage</code></td>
</tr>
<tr>
<td></td>
<td>#Synset/#Synset</td>
</tr>
<tr>
<td>34.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>domainUsageMember</code></td>
</tr>
<tr>
<td></td>
<td>#Synset/#Synset</td>
</tr>
<tr>
<td>35.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>domainCategory</code></td>
</tr>
<tr>
<td></td>
<td>#Synset/#Synset</td>
</tr>
<tr>
<td>36.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>domainCategoryMember</code></td>
</tr>
<tr>
<td></td>
<td>#Synset/#Synset</td>
</tr>
<tr>
<td>37.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>domainRegion</code></td>
</tr>
<tr>
<td></td>
<td>#Synset/#Synset</td>
</tr>
<tr>
<td>38.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>domainRegionMember</code></td>
</tr>
<tr>
<td></td>
<td>#Synset/#Synset</td>
</tr>
<tr>
<td>39.</td>
<td><code>WordSense</code></td>
</tr>
<tr>
<td></td>
<td><code>owl:Class</code></td>
</tr>
<tr>
<td>40.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>inSynSet</code></td>
</tr>
<tr>
<td></td>
<td>#WordSense/#Synset</td>
</tr>
<tr>
<td>41.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>containsWordSense</code></td>
</tr>
<tr>
<td></td>
<td>#Synset/#WordSense</td>
</tr>
<tr>
<td>42.</td>
<td><code>Word</code></td>
</tr>
<tr>
<td></td>
<td><code>owl:Class</code></td>
</tr>
<tr>
<td>43.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>senseOf</code></td>
</tr>
<tr>
<td></td>
<td>#WordSense/#WordSense</td>
</tr>
<tr>
<td>44.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>hasSense</code></td>
</tr>
<tr>
<td></td>
<td>#WordSense/#WordSense</td>
</tr>
<tr>
<td>45.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>frequency</code></td>
</tr>
<tr>
<td></td>
<td>#WordSense/xsd;double</td>
</tr>
<tr>
<td>46.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>lemma</code></td>
</tr>
<tr>
<td></td>
<td>#Word/ &amp;rdfs;Literal</td>
</tr>
<tr>
<td>47.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>senseKey</code></td>
</tr>
<tr>
<td></td>
<td>#WordSense/&amp;rdfs;Literal</td>
</tr>
<tr>
<td>48.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>participleOf</code></td>
</tr>
<tr>
<td></td>
<td>#WordSense/#WordSense</td>
</tr>
<tr>
<td>49.</td>
<td><code>owl:ObjectProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>hasParticiple</code></td>
</tr>
<tr>
<td></td>
<td>#WordSense/#WordSense</td>
</tr>
<tr>
<td>50.</td>
<td><code>owl:SymmetricProperty</code></td>
</tr>
<tr>
<td></td>
<td><code>antonym</code></td>
</tr>
<tr>
<td></td>
<td>#WordSense/#WordSense</td>
</tr>
</tbody>
</table>
51. TopOntology owl:Class

52. owl:ObjectProperty hasItem #TopOntology/#Synset

53. owl:ObjectProperty index #TopOntology/rdfs:Literal

54. owl:ObjectProperty name #TopOntology/rdfs:Literal

55. owl:ObjectProperty broaderItem #TopOntology/#TopOntology

56. owl:ObjectProperty narrowerItem #TopOntology/#TopOntology

### 5 Developing and Managing the WordNet Semantic Web Models

For managing WordNet Semantic Web models we use the Multilingual WordNet Editor [8] together with XMLSpy 2007 and Oracle 10g/11g that provides important XML/RDF/OWL support for data modeling and editing of XML/RDF/OWL WordNet models.

Oracle Database 11g incorporates native RDF/RDFS/OWL support, enabling WordNet application to benefit from a scalable, secure, integrated, efficient platform for semantic data management. Ontological datasets, containing 100s of millions of data items and relationships, can be stored in groups of three, or "triples" using the RDF data model. Oracle Database 11g enables such repositories to scale into the billions of triples, thereby meeting the needs of the most demanding applications of...
WordNet. Managing semantic data models within Oracle Database 11g introduces significant benefits over file-based or specialty database approaches:

- **Low Cost of Ownership:** Semantic applications can be combined with other applications and deployed on a corporate level with data stored centrally, lowering ownership costs. Beyond the advantage of central data storage and query, service oriented architectures (SOA) eliminate the need to install and maintain client-side software on the desktop and store and manage data separately, outside of the corporate database.

- **Low Risk:** RDF and OWL models can be integrated directly into the corporate DBMS, along with existing organizational data, XML and spatial information, and text documents. This results in integrated, scalable, secure high-performance WordNet applications that could be deployed on any server platform (UNIX, Linux, or Windows).

- **Performance and Security:** For mission-critical semantic data models Oracle provides the security, scalability, and performance of the industry’s leading database, to manage multi-terabyte RDF datasets and server communities ranging from tens to tens of thousands of users.

- **Open Architecture:** The leading semantic software tool vendors have announced support for the Oracle Database 11g RDF/OWL data model. In addition, plug-in support is now available from the leading open source tools.

- **Native inference** using OWL and RDFS semantics and also user-defined rules.

- **Querying of RDF/OWL data** and ontologies using SPARQL-like graph patterns embedded in SQL.

- **Ontology-assisted querying** of enterprise (relational) data storage,

- **Loading, and DML** access to semantic data.

  Based on a graph data model, RDF triples are persistent, indexed, and queried, similar to other object-relational data types. Oracle database capabilities to manage semantics expressed in RDF and OWL ensure that WordNet developers benefit from the scalability of the Oracle database to deploy high performance enterprise applications.

**References**

A Comparison of Feature Norms and WordNet

Eduard Barbu and Massimo Poesio
Center for Mind/Brain Sciences, Rovereto, Trento, Italy
eduard.barbu@email.unitn.it
poesio@dit.unitn.it

Abstract. Concepts are the most important objects of study in cognitive science, the building blocks of any theory of mind. Most theories of conceptual organization of semantic memory including the classical theory of concepts assume a featural representation of concepts. The importance of features in contemporary theories of semantic memory posed the researchers the hard problem of finding psychologically relevant concept description. It is believed that the most reliable method for achieving this goal is asking the human subjects in controlled experiments to produce features for a set of concepts. The purpose of this paper is to compare the featural description of concepts in two psychological feature norms with the featural descriptions of concepts in WordNet. To perform this comparison we mapped the concepts in the two feature norms on Princeton WordNet 2.1 and then automatically extracted the potential features from a suitable semantic neighborhood of the respective concepts.

Keywords: concepts, featural representation, feature norms, semantic memory, WordNet, comparison

1 Introduction

Concepts are the most important objects of study in cognitive science, the building blocks of a theory of mind. The debate over the nature and the structure of concepts is as old as the philosophical reflection. In the contemporary cognitive science there are three main tenets about the nature of concepts: concepts are mental representations, cognitive abilities or Fregean senses.

Because this paper will focus on the structure and not on the nature of concepts it is essential for the purpose of subsequent discussion to assume that concepts are mental representations and not cognitive abilities of some sort. We also assume that the humans can access and report the content of these representations. Moreover, we will concentrate not on the structure of concepts in general but on the structure of those concepts lexicalized in (English) language.

Any discussion of the structure of concepts should begin with the older and for some still appealing theory of concepts called the classical theory. It has its roots in the work of philosophers like Plato and Aristotle and till the second part of the 20th century this theory has been practically unchallenged. According to the modern formulation of the classical theory there are two types of lexical concepts: primitive and complex. The complex concepts have a definitional structure composed of other
A Comparison of Feature Norms and WordNet

concepts that specify their necessary and sufficient conditions. If we redefine the constituents entering in the description of the complex concepts we get to define all the concepts using the finite stock of primitive concepts. From now on we will call the description of complex concepts a featural description and the components of the description, features. In this paper we will use interchangeably, when there is no possibility of confusion, the terms feature, property and relation. If we take as example the well-known concept “bachelor”, its featural description contains the features “is a man” and “is unmarried”.

Of course, the stock of concepts that have a featural description in terms of necessary and sufficient conditions is much bigger and not controversial as it is “bachelor”. We can consider countless examples from mathematics: prime number, even number, vectorial space, equilateral triangle, etc.

According to the classical theory, when we classify an object in the world we check to see what the features of the object to be classified are and then we assign the object the category that uniquely fits its description. For example, when we classify a particular dog we verify that the perceptual features we extract from the interaction with that particular exemplar (presumably these are like “has four legs”, “has a head”, etc.), match our mental description of the concept dog.

During the second part of the 20th century the classical theory of concepts came under heavy attack from many quarters: philosophy, psychology and the newly born cognitive science. The main reproach from the psychological perspective has been that the classical model does not predict either the typicality effects or the category fuzziness. The typicality effects refer to the fact that people tend to rank the members of natural categories according to how good examples they are of the respective category. For instance, a sparrow is considered a more typical example of a bird than a chicken. This phenomenon cannot be explained by the classical theory of concepts because, according to it, all members of a category have equal status. Category fuzziness, on the other hand, refers to the fact that some categories have indeterminate boundaries. For example, both answers to the question: “Is a carpet a furniture?” seems to be inadequate [1]. Again this is a problem for the classical theory of concepts because it does not allow for category indeterminacy.

In psychology the first theories that were proposed as alternatives to the classical theory were the twin theories “prototype theory” and “exemplar theories”. These theories succeed where the classical theory failed, namely in explaining both category fuzziness and typicality effects. But for achieving this they had to reject one major tenet of classical theory, namely that concepts have necessary and sufficient conditions. What is important for us is that they did not question the other main classical contention the fact that concepts in the case of prototype theory and the individuals that define a concept in the case of exemplar theories are featural representations.

The importance of features in contemporary theories of semantic memory posed the researchers the hard problem of specifying psychologically relevant concept descriptions. It is believed that the most reliable method for achieving this goal is asking the human subjects in controlled experiments to produce features for a set of concepts. The purpose of this paper is to compare the featural description of concepts in the psychological feature norms with the featural descriptions of concepts in WordNet. To make this comparison we mapped the concepts in the two feature norms
on Princeton WordNet 2.1 and then automatically extracted the potential features from a suitable semantic neighborhood of the respective concepts. For assessing the quality of the proposed automatic procedure we manually performed a comparison between 20 concept descriptions found in feature norms and their descriptions in WordNet.

The rest of the paper is organized as follows. In the first part we introduce the two feature norms and compare them quantitatively and qualitatively. The second part of the paper presents the algorithm for extracting the potential features from Princeton WordNet. The third part gives a quantitative and qualitative comparison between each of the feature norms and the feature extracted automatically. We conclude the paper presenting some related work and the conclusions.

2 Feature Norms

The empirical question a researcher working with featural concept descriptions for testing the hypothesis about semantic memory organization confronts is how to derive a set of features that approximates the mental representation of concepts.

In a paper about semantic memory impairment Farah and McClelland [2] showed how modality specific semantic memory system could account for category deficits after the brain damage. They implemented a neural network model of semantic memory based on the hypothesis that the functional and visual features have a different distribution for living and for non-living things. The proportion of visual versus functional features for living and non-living things was estimated based on a set of dictionary definitions. Their approach has been criticized on the ground that the features that are extracted from dictionary definitions do not provide a good model of the human mental representation.

A better alternative for feature generation is to ask people in controlled psychological experiments to make explicit the content of their semantic memory. In a celebrated series of experiments in the 70’s Rosch and Mervis [3] asked their subjects to produce features for twenty members of six basic level categories. Subsequently they demanded the subjects to rank the respective members according to how good examples they are for the respective categories. For example, the subjects were asked to rank the concepts chair, piano and clock in function of how representative examples they are for the category furniture. One major finding of their study was that typicality of a concept is highly correlated with the total cue validity for the same concept. That is, the most typical items are those that have many features in common with other members of the category and few features in common with members outside the category. Subsequent research replicated the results of Rosch and Mervis, but nowadays it is acknowledged that besides cue validity there are other factors that determine the typicality [4].

Following Rosch, other researchers [5, 6] built feature norms and used them for investigations of the semantic memory. The norms became the empirical material for constructing computational theories about information encoding, storage and retrieval from semantic memory. Following the line of research that started with Rosch and
Mervis, the norms are also used to examine the relation between semantic representations and prototypicality.

According to our knowledge only two feature norms are publicly available. The first one has been built by Garrard and his colleagues [7]. The norm was produced by asking 20 people to provide featural descriptions for a set of 64 concepts of living and nonliving things. McRae and his collaborators [8] have acquired the second feature norm, the largest norm to date, asking 725 subjects to list features for 541 living and not living basic level concepts. From now on, we will refer to these feature norms as Garrard database and McRae database respectively.

The methodology for building the norms differs in some details from one researcher to the other. For example, unlike Rosch and Mervis neither Garrard nor McRae impose their subjects time limits for feature listing task.

In Garrard’s experiment each stimulus (the concept for which the subjects should provide descriptions) was presented on a separate page and the task of the subjects was to fill in the fields present on the page. The fields classified the type of features that the subject should provide: classification features (under the head Category), descriptive features (under is field), parts (under has field) and abilities (under can field).

In McRae’s experiment the stimuli were shown on empty pages. In a task description session experimenters hinted the subjects about the nature of the description that they were expected to provide. As we will see later these methodological differences can account only partially for the dissimilarities in the concept description between Garrard and McRae databases.

To get a feeling of what kind of features are listed in the experiments we present the partial description of the concept “apple” as registered in Garrard database: Apple = {“is a fruit”, “has pips”, “has skin”, “is round”, “has stalk”, “has flesh”, “has core”, “is red”, “is green”, “is sweet”, “has leaves”, “is juicy”, “is coloured”, “is sour”, “has white flesh”, “is small”, “is edible”, “can be cooked”, “can fall”, “can be picked”, “can ripen”, “can rot”}

In addition to the featural description of concepts, the databases contain a wealth of interesting information. We will mention only three fields that are particularly important. Dominance is a field indicating the number of subjects that listed a certain feature. It reflects the “weight” of a certain feature in the mental representation of a concept: the higher the dominance for a specific feature, the greater the importance of the respective feature. Distinctiveness reflects the percent of members of a category for which a specific feature is listed. It is a measure of how good the individual features are in distinguishing the categories. For example, “has trunk” is a highly distinctive feature for the category elephant because it helps distinguishing the members of this class from the other animals not members, whereas “has tail” is a lowly distinctive feature because the elephants share this feature with other animals.

The third field, the most significant from our point of view, gives a classification of feature types in the databases. Unfortunately the two databases contain different feature classifications. Garrard database has a relative simple but nevertheless controversial feature classification. The features are classified as categorizing, sensory, functional or encyclopedic. The categorizing features taxonomically classify the stimulus (e.g. lion “is an animal”), the sensory features are those features that are grounded in a sensory modality (e.g. the bus “is coloured” or apple “is sour”), the
functional features describes an activity or the use someone makes of an item (monkeys “can run”, a brush “can apply paint”) and the encyclopedic features are the features that cannot be classified as superordinate, sensory or functional. Sometimes the way Garrard and colleagues are making use of this classification is puzzling. For example, they are classifying some features that denote parts of the sophisticated modern apparatus as being sensory (as for example the rotor or the control of an helicopter). Even if some can argue that we “see” these parts, their identification as parts is largely based on the knowledge of the structure and the functions of a modern vehicle.

More interesting is the classification employed by McRae and colleagues in classifying the features in their database. They are using a taxonomic classification, a slightly modified version of Wu and Barsalou taxonomy [9], a taxonomy derived from studies on human perception. Among the principles that Wu and Barsalou considered when constructing the taxonomy were: the introspective experience of the subjects when they generate feature norms, the modality specific regions of the brain, the frame theory of Fillmore and others.

Their taxonomy has two levels; at the coarsest level the features are classified as taxonomic properties, entity properties, situational properties or introspective properties. The taxonomic properties are those properties that taxonomically classify an entity, entity properties denote general properties of an entity, situational properties are properties characteristic to situations and introspective properties are properties of subject mental states.

At the next level each mentioned category of properties is again further divided. The modified Wu and Barsalou taxonomy used by McRae has at the second level 27 categories. To make things clear let’s consider a partial description of the concept accordion: {“a musical instrument” (Taxonomic::Superordinate), “has keys” (Entity::External Component), “produces music” (Entity::Entity Behaviour)}.

In the above description three features of the concept accordion are listed. The first feature states that the accordion is a musical instrument, and according to Wu and Barsalou category it is classified at the first level as being a taxonomic feature and at the second level as being a superordinate feature (Taxonomic::Superordinate). The other two features “has keys” and “produces music” are classified as being “Entity::External Component” and “Entity::External Behavior” feature type respectively. The Entity::Component properties denote those features that are external components of the object to be described whereas Entity::Behaviour features denote activities component of the behavior of the object under description.

Before making the comparison between WordNet and these databases we want to compare the databases with each other (Table 1).
A Comparison of Feature Norms and WordNet

Table 1. A quantitative evaluation of Garrard and McRae databases

<table>
<thead>
<tr>
<th></th>
<th>Garrard Database</th>
<th>McRae Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts</td>
<td>62(^1)</td>
<td>541</td>
</tr>
<tr>
<td>Feature-concept Pairs</td>
<td>1657</td>
<td>7275</td>
</tr>
<tr>
<td>Average Number of F/C</td>
<td>26.7</td>
<td>13.4</td>
</tr>
</tbody>
</table>

The first row of the Table 1 lists the number of concepts in each database; the second row gives the number of concept-feature pairs in the each of the two databases and the last row lists the average number of feature per concept for each database. Please observe that in Garrard database the average number of feature per concept is twice bigger than in McRae database. Perhaps the strategy that Garrard used for feature production task fields paid off.

For the qualitative comparison of the databases we semi-automatically mapped the databases. First we identified the common concepts in the two databases and then we semi-automatically mapped the concept-feature pairs. In most cases the mapping between concept-features pairs is one to one but in some cases the mapping is one to many or many to one.

The mapping between McRae and Garrard databases revealed the results presented in the tables 2 and 3. From table 2 it can be seen that we found a set of 50 concepts common to both databases (Mapped Concepts). For this common set of concepts we list the number of concept-feature pairs present in each database (“Garrard CF pairs” and “McRae CF pairs” respectively). Finally, “Common Mapped Pairs” field represents the number of concept-feature pairs the databases have in common.

Table 2. The mapping between Garrard and McRae databases

<table>
<thead>
<tr>
<th></th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapped Concepts</td>
<td></td>
</tr>
<tr>
<td>Garrard CF pairs</td>
<td>1326</td>
</tr>
<tr>
<td>McRae CF pairs</td>
<td>765</td>
</tr>
<tr>
<td>Common Mapped Pairs</td>
<td>430</td>
</tr>
</tbody>
</table>

Table 3. A per feature type comparison between the Garrard and McRae databases

<table>
<thead>
<tr>
<th>Relation classification</th>
<th>CFPM</th>
<th>CFPG</th>
<th>CFPG / CFPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Made Of</td>
<td>32</td>
<td>27</td>
<td>0.84</td>
</tr>
<tr>
<td>Superordinate</td>
<td>67</td>
<td>54</td>
<td>0.80</td>
</tr>
<tr>
<td>External Component</td>
<td>171</td>
<td>129</td>
<td>0.75</td>
</tr>
<tr>
<td>Entity Behaviour</td>
<td>91</td>
<td>60</td>
<td>0.65</td>
</tr>
<tr>
<td>External Surface Property</td>
<td>102</td>
<td>64</td>
<td>0.62</td>
</tr>
<tr>
<td>Internal Component</td>
<td>21</td>
<td>13</td>
<td>0.61</td>
</tr>
<tr>
<td>Internal Surface Property</td>
<td>18</td>
<td>11</td>
<td>0.61</td>
</tr>
</tbody>
</table>

\(^1\) Garrard published only 62 from the 64 concepts for which he collected featural descriptions
As one can see looking at the same table, 56% of concept-feature pairs listed in McRae database are present in Garrard database, but only 32% of the concept-feature pairs in Garrard database are also present in McRae database. The problem is how to make sense of these differences. The second finding namely that 68% of the concept feature pairs in Garrard database are not in McRae database can be explained by the methodological difference between the authors. One can argue that Garrard subjects had only to fill in the fields already on the page, whereas McRae subjects should have produced the features with no help. More problematic is how to interpret the first number, 43% of the concept-feature pairs in the McRae database are not in the Garrard database. This fact poses serious problems for the computational theories of the semantic memory based on feature norms but we will not address the problem in this paper.

Table 3 gives the database comparison using Wu and Barsalou taxonomy. The first column lists some salient relation types in Wu and Barsalou taxonomy omitting the first level of classification. For the set of common concepts in the two databases, the second column gives the number of concept-feature pairs classified with a certain relation type in McRae database (CFPM). Thus we find that 32 concept-feature pairs were classified as instances of “Made Of” relation type, 67 as exemplifying the Superordinate relation type and so on. The third column looks at the same statistic for the concept-feature pairs that are in the intersection between McRae and Garrard database (CFPG). For example from the 32 concept-feature types classified as instances of “Made Of” relation type in the McRae database, 27 have been mapped on Garrard database. The last column gives the report of the last two columns. We eliminated those relation types that classified less than 11 concept-feature pairs or had a score in the last column less than 0.51.

One can see that the feature types successfully mapped from McRae database to Garrard database are parts (“Made Of”, “External Component”, “Internal Component”), taxonomic features (Superordinate), the features classified under “Entity Behaviour” and the features that denote external and internal surface properties.

3 WordNet Feature Extraction

The procedure for building the feature norms is a time consuming one: for example McRae and his colleagues started the feature collection in the 90’s.

Hoping that we can find an automatic procedure for producing featural concept description we want to see how the features norms compare with WordNet. WordNet is a resource built starting from psycholinguistic principles and aiming to be a model of human semantic memory. The feature norms, as we showed before, are built having in mind the computational modeling of semantic memory. Therefore one would expect to find in WordNet many of the features that are produced by the subjects in the psychological experiments.

To automatically compare the concept descriptions in the two databases with the concept descriptions in WordNet we mapped the concepts in the databases to
A Comparison of Feature Norms and WordNet

WordNet concepts. The mapping procedure has two steps: the first one is fully automatic and the second one is manual.

In the automatic step we try to guess which is the most likely assignment between the words that were offered as stimuli in the databases and the corresponding WordNet synsets. First we generate all synsets that contain the stimuli words in the databases and their respective hyperonyms up to the root of the WordNet tree. Then from the Category field in Garrard database and from the Superordinate property types in the McRae database we generate the classification of the database concepts. Afterwards we perform the intersection between the words that classify the stimulus in the databases and the hyperonyms in WordNet. In case that the intersection is not empty and not two senses of a word have the same hyperonym in WordNet we can automatically find the synset corresponding to the stimulus. For example the word apple present in both databases is classified in both of them as a fruit. There are two senses of the word apple in WordNet, the first one (apple#1) refers to the apple as a fruit and the second one (apple#2) refers to the apple as a tree. One of the hyperonym synsets of apple#1 has as one of its members the word fruit. Therefore we find that the apple in both databases should be mapped on the first sense of apple in WordNet (apple#1).

In the second step we manually map the stimuli words that could not be mapped automatically and we also briefly recheck the accuracy of automatic mapping.

Before presenting the algorithm for WordNet feature extraction we give some useful term definitions: projection set, semantic neighborhood and WordNet feature.

**Definition 1 (Projection Set).** The set of synsets that represent the mappings of the concepts in the databases onto WordNet is called the projection set.

We have two projection sets, one for each database: Garrard project set and McRae project set respectively. When we use the term projection set without qualification we refer to the mapping of both projection sets.

**Definition 2 (Semantic Neighborhood).** The semantic neighborhood of a synset $s$ is a graph $<N, R>$ where $N$ is a finite set of nodes representing WordNet synsets and $R$ is a set of relations linking the nodes.

The algorithm for feature extraction considers only two semantic relations in $R$: hyperonymy and meronymy. We choose the hyperonymy relation because it is an inheritance and transitive relation therefore along this relation a concept inherits all the featural descriptions of its superordinates. We included the meronymy relation because the parts are among the most salient feature types produced by the subjects in the feature generation task.

**Definition 3 (WordNet Feature).** A WordNet feature of a concept is any word in the synsets of its semantic neighborhood and any noun, adjective or verb in the glosses of the synsets of its semantic neighborhood.

The feature extraction for the synsets in the projection set is performed from the semantic neighborhood of each synset.
Considering among potential features of a concept any noun adjective or verb seems to overestimate the number of real features present in WordNet. Remember that we want to see which features in the databases are also present in WordNet. Therefore the generation of a reasonable number of “false” features would not affect at all the comparison because the set of real features is a subset of generated WordNet features.

The algorithm for the extraction of WordNet features for the concepts represented by the synsets in the projection set has three steps. In the first step we generate the semantic neighborhoods of each synset in the projection set. In the second step we part of speech tag and lemmatize all the glosses of the synsets from the semantic neighborhood. The part of speech tagging and the lemmatization is performed with TreeTagger, a language independent part of speech tagger, developed by the Institute for Computational Linguistics of the University of Stuttgart. The tagger uses an English parameter file trained on Penn Treebank. In the third step we extract all the WordNet features and eliminate possible duplicate features. Figure 1 shows a part of the semantic neighborhood of the synset apple. A node of the graph is labeled with its corresponding synset; the synset is followed by its gloss. The edges of the graph are labeled with the semantic relations in the above-mentioned R set.

Running the algorithm for the toy example above we obtain the following potential features for the concept apple: fruit, red, yellow, green, skin, sweet, tart, crisp, whitish, flesh, edible, reproductive, body, seed, plant, vegetable, grow, market, peel.

The above algorithm allows us to make a global comparison between the database features and WordNet features. To perform a much finer comparison we will classify the synsets in the projection set and then compare the database features and WordNet features per category.

To find a suitable classification, first we generate the WordNet tree along the hyperonym relation starting from the synsets in the projection set. We treat the synsets in the projection set as the objects to be classified and any category subsuming the synsets in the projection set as a potential classifier. There are many potential classifications one can find but we would like to find a classification that forms a partition of the objects to be classified. We also want that the clear-cut categories formed to be basic level categories.
Synset: edible fruit
Gloss: edible reproductive body of a seed plant especially one having sweet flesh

Synset: produce, green goods, green groceries
Gloss: fresh fruits and vegetable grown for the market

Synset: apple
Gloss: fruit with red or yellow or green skin and sweet to tart crisp whitish flesh

Synset: peel, skin
Gloss: the rind of a fruit or vegetable

In figure 2 we see a part of the classification tree whose leaves are synsets from the projection set. The problem one confronts is where the tree should be cut to form a good partition. Should we cut the tree at the node “musical instrument” and classify with its label all the leaves that fall under the node or should we cut the tree at the nodes “free reed instrument” and “woodwind” and classify with these other two labels the leaves that fall under them?

Fig. 1. The semantic neighborhood of the synset apple

Fig. 2. A part of the classification tree
Because we want to produce basic level categories, cutting the tree at the node musical instrument seems the obvious solution. We explored the possibility of finding an automatic resolution of the problem. Ideally an algorithm should take as input the hyperonimic tree and produce as output a good partition of it.

The algorithm we tested cut the tree at the nodes that give the smaller generalization possible. A node from the hyperonimic tree gives the smaller possible generalization if it dominates at least two synsets from the projection set. After we collect all the categories satisfying the above condition we retain only those that form a partition of the objects to classify. For example, applying the algorithm for the toy example in figure 2 one cuts the tree at the nodes “free reed instrument” or “musical instrument”. Please observe that the tree cannot be cut at the node woodwind because this node dominates only one leaf and it does not give us any useful generalization. Subsequently we observe that the category “musical instrument” dominates the category “free reed instrument” and that only the category “musical instrument” gives us a partition of the objects to classify. Unfortunately this straightforward method does not produce satisfying results because it generates many artificial categories.

The other automatic approach we considered was to use the classifications already present in the two databases: the category field for Garrard database and the Superordinate relation type for the McRae database. One can argue that the categories thus obtained are basic-level because the subjects of the psychological experiments produced them. This method however leaves unclassified the word stimuli for which the subjects in McRae experiment did not produced categories. We chose to take a middle path. Starting from the categories produced by the subjects in each of the two experiments and inspecting the classification tree we came up with a much better category set. The partition we obtained came with the cost of not being able to cover the whole space. For the Garrard database the following categories form a partition of 50 concepts: {“implement”, “bird”, “mammal”, “fruit”, “container”, “vehicle”, “reptile”}. One can see that the class of animals is split into reptiles, mammals and birds an then we have the partition of tools (implement), fruits and vehicles. For the McRae database instead the partition has 16 categories and covers 345 concepts: {“clothing”, “implement”, “fruit”, “furniture”, “mammal”, “plant”, “appliance”, “weapon”, “container”, “musical instrument”, “building”, “vehicle”, “fish”, “reptile”, “insect”, “bird”}

4 Results and discussion

For each of the two databases we performed a global comparisons with WordNet, a comparison by feature type and then a per category comparison using the category partitions presented in the final part of the section 3. To make possible the automatic comparison between feature norms and WordNet we had to make two simplifying assumptions. In both Garrard and McRae databases “has legs” and “has four legs” for example are considered to be distinct features. We neglect the cardinality and collapse this features in one: “has legs”. We also considered that in the case when a feature expresses a two place relation and the relation is not explicitly defined in WordNet (e.g. meronym or hyperonym), the presence of the arguments of the relation in
WordNet are sufficient for deciding that the relation linking the arguments in WordNet is the same relation expressed by the database feature. For example if we want to decide if the feature “used for cooking” for the concept “pot” exists in WordNet and we find the word cooking in the semantic neighborhood of the concept pot then we assume that the relation that holds between pot and cooking is the functional relation “used for”. For most features in the databases this is true but there are some cases when our second assumption is false. Table 4 shows the proportion of the concept-feature pairs in the databases one can find in WordNet.

**Table 4.** A global comparison between feature norms and WordNet

<table>
<thead>
<tr>
<th>Database</th>
<th>CF pairs in database</th>
<th>CF pairs in WordNet</th>
<th>Percent in WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>McRae</td>
<td>6925</td>
<td>2108</td>
<td>30%</td>
</tr>
<tr>
<td>Garrard</td>
<td>1537</td>
<td>342</td>
<td>22%</td>
</tr>
</tbody>
</table>

The “CF pairs in database” column lists the number of concept-feature pairs in each database whereas the “CF pairs in WordNet” column gives the number of concept-feature pairs in the intersection between each database and WordNet. The last column shows the percent of the features in the databases estimated to be in WordNet. One can see that the percent of concept-feature pairs in the McRae database also found in WordNet is higher that the percent of features in the Garrard database that are in WordNet (30 % vs. 22%).

In the next two tables we see what features types are better represented in WordNet. Tables 5 and 6 list in the first column the feature types, in the second column the number of the typed concept feature pairs in the respective database, in the third column the number of concept-feature pairs in the intersection between WordNet and the databases for each feature type, and in the last column the percent of concept-feature pairs found in WordNet for each feature type.

**Table 5.** Per feature type comparison between Garrard database and WordNet

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>CF pairs in database</th>
<th>CF pairs in WordNet</th>
<th>Percent in WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorizing</td>
<td>115</td>
<td>83</td>
<td>72%</td>
</tr>
<tr>
<td>Sensory</td>
<td>737</td>
<td>190</td>
<td>25%</td>
</tr>
<tr>
<td>Encyclopedic</td>
<td>241</td>
<td>26</td>
<td>11%</td>
</tr>
<tr>
<td>Functional</td>
<td>444</td>
<td>43</td>
<td>10%</td>
</tr>
</tbody>
</table>
Table 6. Per feature type comparison between McRae database and WordNet

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>CF pairs in database</th>
<th>CF pairs in WordNet</th>
<th>Percent in WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superordinate</td>
<td>588</td>
<td>470</td>
<td>80%</td>
</tr>
<tr>
<td>External Component</td>
<td>926</td>
<td>442</td>
<td>48%</td>
</tr>
<tr>
<td>Internal Component</td>
<td>168</td>
<td>64</td>
<td>38%</td>
</tr>
<tr>
<td>Origin</td>
<td>59</td>
<td>16</td>
<td>27%</td>
</tr>
<tr>
<td>Contingency</td>
<td>91</td>
<td>24</td>
<td>26%</td>
</tr>
<tr>
<td>External Surface Property</td>
<td>1175</td>
<td>306</td>
<td>26%</td>
</tr>
<tr>
<td>Made Of</td>
<td>471</td>
<td>122</td>
<td>26%</td>
</tr>
<tr>
<td>Function</td>
<td>1098</td>
<td>281</td>
<td>25%</td>
</tr>
<tr>
<td>Participant</td>
<td>183</td>
<td>44</td>
<td>24%</td>
</tr>
<tr>
<td>Internal Surface Property</td>
<td>179</td>
<td>40</td>
<td>22%</td>
</tr>
<tr>
<td>Location</td>
<td>455</td>
<td>84</td>
<td>18%</td>
</tr>
<tr>
<td>Associated Entity</td>
<td>153</td>
<td>22</td>
<td>14%</td>
</tr>
<tr>
<td>Systemic Property</td>
<td>293</td>
<td>38</td>
<td>13%</td>
</tr>
<tr>
<td>Entity Behavior</td>
<td>495</td>
<td>63</td>
<td>13%</td>
</tr>
<tr>
<td>Action</td>
<td>184</td>
<td>20</td>
<td>11%</td>
</tr>
<tr>
<td>Evaluation</td>
<td>105</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>

Looking at the tables 5 and 6 we see which feature types are better represented in WordNet and which are lacking. Table 5 gives the comparison for Garrard database, the features being classified according to the categorization employed by Garrard and colleagues. As one expects the feature type better covered by WordNet is the classification type, 72% of the classification features produced by the subjects in Garrard experiment are found in WordNet. All other feature types are not so well represented, the second place being adjudicated by the sensory features with 25%. As we argued in section 2, Garrard classification is very crude and therefore not very informative.

Much more interesting is the comparison for McRae database (in the table 3 we do not list all the feature types in Wu and Barsalou taxonomy. We do not show the feature types that classify less than 50 concept-feature pairs).

Meeting our expectations the best feature type in terms of coverage is the superordinate type (78%), this is the only feature type that has coverage over 50% in WordNet. The relations that denote parts and that correspond to various types of meronymy are relatively well-represented occupying position 2, 3 and 7.

The features classified under “External Surface Property” label, occupies the fifth place. The high position in the table for the external surface properties can be explained by the fact that the definitions of many concepts denoting concrete objects list properties of their external surfaces (e.g. shape, color). For example, the definition of the concept apple contains the attributes red, green and yellow, all being external surface properties according to Wu and Barsalou taxonomy.

The last feature type, labeled “evaluation”, has no representation in WordNet. The features typed as evaluation reflect subjective assessment of objects or situations (for
example the evaluation that a bag is useful, that a blouse is pretty or the shark is dangerous).

A comparison between databases and WordNet using the category partitions discussed above is given in the table 7 (we show only the top seven categories in McRae partition):

<table>
<thead>
<tr>
<th>Garrard category</th>
<th>Percent Overlap</th>
<th>McRae category</th>
<th>Percent Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fruit</td>
<td>37%</td>
<td>Fish</td>
<td>52%</td>
</tr>
<tr>
<td>Bird</td>
<td>34%</td>
<td>Fruit</td>
<td>43%</td>
</tr>
<tr>
<td>Implement</td>
<td>25%</td>
<td>Vehicle</td>
<td>43%</td>
</tr>
<tr>
<td>Container</td>
<td>21%</td>
<td>Bird</td>
<td>42%</td>
</tr>
<tr>
<td>Mammal</td>
<td>20%</td>
<td>Plant</td>
<td>37%</td>
</tr>
<tr>
<td>Vehicle</td>
<td>20%</td>
<td>Musical instrument</td>
<td>36%</td>
</tr>
<tr>
<td>Reptile</td>
<td>17%</td>
<td>Weapon</td>
<td>33%</td>
</tr>
</tbody>
</table>

The columns 1 and 3 represent the set of category partitions in the databases. The columns 2 and 4 represent the percent of the concept feature pairs for each category that are present in WordNet. The best-represented categories in WordNet for Garrard database are fruits and birds whereas for McRae database the best represented categories in WordNet are fish, fruit, vehicle and birds.

For assessing the accuracy of our automatic procedure we performed a manual comparison between 20 WordNet concept descriptions and each of the two corresponding database descriptions. The 20’s concept set contains 10 concepts that our algorithm says have the highest overlap with the databases and 10 concepts that have the lowest overlap with the databases.

A manual mapping between the database concept description and WordNet concept description revealed that the number of concept-feature pairs common to databases and WordNet is bigger than the estimation given by our algorithm. There are three reasons for this fact. The first reason is that some features present in WordNet are expressed using different words from the words used to register the same features in the two databases. For example in McRae database one of the features of the concept “anchor” is “found on boats”. The definition of the concept “anchor” in WordNet contains a semantically close word: vessel (WordNet says that vessel is a hyperonym of boat). If the words in the WordNet glosses would have been semantically disambiguated our algorithm would have exploited this information and improved the automatic estimation. However even a WSD of the words in the glosses does not completely solves our problem because is a notorious fact the WordNet makes very fine sense discriminations and many features that are near synonyms of the words in the glosses would not be found.

The second reason for the inaccuracy of our automatic procedure is related with a general problem of feature norms. It is assumed for methodological simplicity that features listed in the feature production task are independent. However the above
statement is known to be false. One of the most important relations that link the features is the entailment relation. For example the concept trolley’s features: “used for carrying things” and “used for moving things” are related by entailment. If someone carries some things with a trolley he will always move the things. The relation of entailment holds also between some features in the feature norms and some features in WordNet. The functional feature of the concept “anchor” “used for holding the boats still” is logical equivalent with the feature “prevents a vessel for moving” found in the WordNet gloss of the concept anchor.

The third reason why the automatic comparison fails to reveal the true overlap between databases and WordNet is the incompleteness of WordNet. A very salient feature that the human subjects produce when they describe concrete objects are the parts of the respective objects but many concepts from the projection set lack the meronyms in PWN 2.1. We do think that a manual comparison with a complete WordNet would show an overlap of approximately 40% with McRae database and 30% with Garrard database.

The comparison between feature norms and WordNet reveals some potential improvements for the future WordNet versions. To find what feature types are lacking one needs to inspect Table 6 and evaluate any feature type except Superordinate type and the feature types that are related with parts. We will briefly discuss three-feature type present in feature norms but lacking or underrepresented in WordNet: the evaluation, the associated entity and the function feature types.

As we argued above even if the evaluation features are an important part of the semantic representation for some concepts they are totally missing from WordNet. We do not think that every possible subjective evaluation should find a place in WordNet but only the most salient ones. For example the evaluation that sharks are generally considered dangerous or that hyenas are seen as being ugly should be part of the WordNet entry for shark and hyena respectively.

Other interesting feature type under-represented in WordNet is the associated entity feature type. As many of the concepts presented as stimuli in feature generation tasks denote concrete objects, the mental representation for these concepts includes the knowledge of the entities that we normally associate with these objects in the situations we typically encounter them. For example we typically associate an anchor with the chains or ropes it is attached to or we associate an apple with the worms it may be infested or even we associate bagpipes with Scotland.

The function or role that an entity serves for an agent is an important part of the meaning of the respective entity. We use keys to lock or open doors, we empty and fill the basket we use trolley for transporting things and the garages are used for storing cars. Many of these important functional features are lacking from WordNet (only 281 of the 1098 function features in McRae database are present in WordNet). If WordNet wants to be a model of the human semantic memory it should rethink its structure to accommodate the feature types present in feature norms.
5 Related work

We are not aware of other work that compared the feature norms concept descriptions with WordNet concept descriptions. However there has been a lot of dedicated effort to the concept extraction from web or corpora and in some cases there were attempts to compare these concepts descriptions with feature norms. Some of this work has sought to extract information about attributes such as parts and qualities [10, 11]. Almuhareb and Poesio developed supervised and unsupervised methods for feature extraction from the Web based on ideas from Guarino [12] and Pustejovsky [13] among others, and showed that focusing on extracting ‘attributes’ and ‘values’ leads to concept descriptions that are more effective from a clustering perspective – e.g., to distinguish animals from tools or vehicles – than purely distributional descriptions. They extracted candidate attributes using constructions inspired by [14] such as “the X of the car is…” and then removed false positives using a statistical classifier.

Recently Poesio and collaborators [15] evaluated the concept descriptions automatically extracted from one of the biggest corpora in existence against three feature norms, the two presented in this paper plus a feature norm produced by Vinson and Vigliocco. They made an in depth comparison between the three feature norms including the computation of statistical correlations between the feature norms concept descriptions and corpora concept descriptions.

More generally our work is connected with the ontology learning effort in the natural language processing and semantic web community and with various work in psychology that tries to understand the human conceptual system using empirical methods.

6 Conclusions and further work

The comparison between the concept description in the databases Garrard and McRae and the comparison between these database concept description and WordNet revealed some interesting results. First we saw that 56 % of concept feature pairs listed in McRae database are present in Garrard database and 32 % of the concept-feature pairs in Garrard database are present in McRae database. We also found using an automatic procedure that 30 % percents of the concept-feature pairs in McRae database are found in WordNet and 22 % of concept-feature pairs in Garrard database are present in WordNet. We argued that an ideal comparison between the two databases and WordNet would reveal a bigger overlap, which will be comparable with the overlap between the two psychological databases.

Using Wu and Barsalou taxonomy and the manual comparison of 20 concept descriptions in the databases and WordNet we showed that WordNet description lack or under represent important features types present in the feature norms such as: evaluation, associated entity and functional feature types. We firmly believe that any future improvement of WordNet should take into consideration the feature norms.

We also stressed the fact that the features in the feature norms are not independent. We would like to find an automatic method for learning the structure that ties together the features.
A week point of our automatic feature extraction algorithm from WordNet is that it does not find the relation between the concept to be described and the potential WordNet features extracted from glosses. Taking into account this observation we are exploring a better procedure for feature extraction, a procedure that exploits a parser for finding the correct relation between the focal concept and the concepts found in the glosses. We hope we will produce in the near future a graphical tool that will help the researchers working with feature norms to easily extract WordNet concept descriptions.

Acknowledgments

We like to thank professor Lawrence Barsalou from Emory University for providing us the paper discussing Wu and Barsalou taxonomy. We are also indebted to our colleagues Marco Baroni and Brian Murphy for some stimulating discussions.

References

Enhancing WordNets with Morphological Relations: A Case Study from Czech, English and Zulu

Sonja Bosch¹, Christiane Fellbaum², and Karel Pala³

¹ University of South Africa Pretoria, South Africa, boscse@unisa.ac.za
² Department of Psychology, Princeton University USA, fellbaum@princeton.edu
³ Faculty of Informatics, Masaryk University Brno, Czech Republic, pala@fi.muni.cz

Abstract. WordNets are most useful when their network is dense, i.e., when a given word of synsets is connected to many other words and synsets with lexical and conceptual relations. More links mean more semantic information and thus better discrimination of individual word senses. In the paper we discuss one kind of cross-POS relation for English, Czech and Bantu WordNets. Many languages have rules whereby new words are derived regularly and productively from existing words via morphological processes. The morphologically unmarked base words and the derived words, which share a semantic core with the base words, can be interlinked and integrated into WordNets, where they typically form "derivational nests", or subnets. We describe efforts to capture the morphological and semantic regularities of derivational processes in English, Czech and Bantu to compare the linguistic mechanisms and to exploit them for suitable computational processing and WordNet construction. While some work has been done for English and Czech already, WordNets for Bantu languages are still in their infancy ([2], [16]) and we propose to explore ways in which Bantu can benefit from existing work.

1 Introduction: Inflectional and Derivational Morphology

Many languages possess rules of word formation, whereby new words are formed from a base word by means of affixes. The derived words differ from the base words not only formally but also semantically, though the meanings of base and derivative words are closely related. These processes, referred to as morphology, fall into two major categories. Inflectional morphology, also called grammatical morphology, is concerned with affixes that have purely grammatical function. Thus, most Indo-European languages have (or once had) verbal morphology to mark person, number, tense and aspect as well as noun morphology to indicate categories like gender, number and case. Czech exploits what can be called a 'cumulation' of functions, i.e., one
inflectional suffix conveys as a rule several grammatical categories; for nouns, adjectives, pronouns (as well as numerals) the categories expressed by the affixes are gender, number and case. While Czech is a richly inflected language, English has developed characteristics of an analytic language where grammatical functions are assumed by free morphemes; for example, future tense, unlike past and present, is marked by will. As in Czech, a single morpheme can have several grammatical functions; -s marks both plural nouns and present tense third person verbs. Bantu languages are agglutinative and use affixes to express a variety of grammatical relations and meanings. These morphemes 'glue' onto stems or roots. The morphemes are not polysemous, as one of the principles that characterises agglutinating languages is the one-to-one mapping of form and meaning [11], and each morpheme therefore conveys one grammatical category or distinct lexical meaning.

The preparation of manuscripts which are to be reproduced by photo-offset requires special care. Papers submitted in a technically unsuitable form will be returned for retyping, or canceled if the volume cannot otherwise be finished on time.

Importantly, the inflected word belongs to the same form class (i.e., represents the same part of speech) as the base. By contrast, derivational morphology often yields words from a different form class. For example, the English verb soften is derived from the adjective soft by means of the suffix -en. Both inflectional and derivational morphology encompass regular and productive rules that are an important part of speakers' grammar. Given a new (or nonce) word like wug, even young children effortlessly produce the (inflected) plural form wugs [1]. Speakers avail themselves of the rules of derivational morphology to form and interpret tens of thousands of words.

A third productive mechanism to derive new words from existing ones is compounding. Examples are English flowerpot, bittersweet, and dry-clean. In Czech compounding is a regular word derivation procedure but it is considered rather marginal and not so productive. An example: česko+slovenský (Czecho-Slovak) or bratro+vrah (murderer of the brother).

In Bantu, compounding is also a productive and regular way of creating new words and it has its own rules. Examples are:

Northern Sotho
sekêpê (ship) + môya (air): sêkêpemôya (airship)

Zulu
abantu (people) + inyoni (bird): abantunyoni (astronauts)
unkhumbi (boat) + ingwenya (crocodile): umkhumbingwenya (submarine)

Venda
ngowa (mushroom/s) + mpengo (madman): ngowampengo (inedible mushroom/s)

In the remainder of this paper, we focus on derivational morphology. We ask how we can exploit its regularity to populate WordNets and to characterize both formal and semantic relations. We explore and formulate derivational rules (D-rules) allowing us to generate automatically as many word forms as possible in the three languages we focus on (English, Czech and Bantu) and to assign meaning to the output of these rules. Formulating D-rules rules would bypass the task of compiling and maintaining large lists of base forms (stems) and would allow us to generate automatically the core of the word stock of the individual languages.
When trying to write the formal D-rules that allow us to generate new words automatically we meet the problem of over- and undergeneration of derived forms. That is, the D-rules could either produce forms that are possible but not actually occurring forms (in corpora or dictionaries), or the rules could fail to generate all attested forms. To avoid errors as well as undergeneration, one currently relies primarily on the manual checking of the output, but we are developing procedures that can semi-automatize this process by comparing the output of the D-rules to corpora or dictionaries. Addressing the overgeneration problem requires re-inspection of the D-rules and correcting those that generate ill-formed strings.

2 Derivational morphology

Derivational affixes form new words with meanings that are related to, but distinct from the base to which they are attached. In this way they differ from inflectional affixes, which add grammatical specifications to a base word. Like inflectional morphology, derivational morphology tends to be regular and productive, i.e., speakers use the rules to form and understand words that they may never have encountered. This shows that derivational morphemes are associated with meanings.

However, the meanings may be polysemous, as in English and Czech, and speakers have to rely on the meaning of the base word or on world knowledge to understand the derived word.

In comparison to Czech and English, D-affixes in Bantu only acquire meaning by virtue of their connection with other morphemes (for example agent, result of an action, instrument of an action etc.) and cannot always be assigned an independent semantic value. This poses a challenge for the definition of “lexical unit” or “word”, which must be met when one constructs a WordNet.

2.1 Derivational relations in Czech

We discuss the two main mechanisms of Czech derivational morphology, suffixation and prefixation. We classify the suffixes and prefixes semantically.

2.1.1 The Czech morphological analyzer

The basic and most productive derivational relations expressed by affixes or, more precisely, the rules describing them were formulated and integrated into a Czech morphological analyzer, Ajka, resulting in a D-version. Ajka is an automatic tool that is based on the formal description of the Czech inflection paradigms [21 and that was developed at the NLP Centre at the Faculty of Informatics Masaryk University Brno. Ajka’s list of stems comprises approx. 400 000 items, up to 1600 inflectional paradigms and it is able to generate approx. 6 mil. Czech word forms. It can be used for lemmatization and tagging, as a module for syntactic analyzer, and other NLP applications.
A version of Ajka for derivational morphology, (D-Ajka), can generate new word forms derived from the stems using rules capturing suffix and prefix derivations. A Web Derivational Interface make it possible to further explore the semantic nature of the selected noun derivational suffixes as well as verb prefixes and establish a set of the semantic labels associated with the individual D-relations. For verbs, the work focused on exploring the derivational relations between selected prefixes and corresponding Czech verb stems or basic non-derived verbs for one verb semantic class (verbs of motion).

Using the analyzer Ajka and the D-interface allowed the addition of selected noun and verb D-relations to the Czech WordNet and its enrichment of approx. 31 000 new Czech synsets, using the DebVisdic editor and browser (see Fig. 1 screenshots).

2.1.2 The Czech data

The starting Czech data include 126 000 noun stems and 22 noun suffixes, 42 745 verb stems (or basic verb forms) and 14 verb prefixes. There are also alternations (infixes) in stems that are not considered here.

The complete inventory of the main noun suffixes is much larger (approx. 120) and the same holds for the set of verb prefixes (approx. 240); here we consider only the primary prefixes (14, and from them 4). The higher number of prefixes in Czech follows from the fact that for each primary prefix there are about 15 secondary (double) ones.

In Czech grammars [10] we can find the following main types (presently 14) of the derivational processes exploiting suffixes and prefixes:

1. mutation: noun -> noun derivation, e.g. ryba - ryb-ník (fish -> pond), semantic relation expresses location – between an object and its typical location,

2. transposition (the relation existing between different POS): noun -> adjective derivation, e.g. den -> den-ní (day ->daily), semantically the relation expresses property,

3. agentive relation (existing between different POS): verb -> noun e.g. myslit -> myslit-tel (think -> thinker), semantically the relation exists between action and its agent,

4. patient relation: verb -> noun, e.g. trestat -> trestanec (punish ->convict), semantically it expresses a relation between an action and the object (person) impacted by it,

5. instrument (means) relation: verb -> noun, e.g. držet -> držák (hold ->holder), semantically it expresses a tool (means) used when performing an action,

6. action relation (existing between different POS): verb -> noun, e.g. učit -> učen-í (teach -> teaching), usually the derived nouns are characterized as deverbatives, semantically both members of the relation denote action (process),

7. property-verbadj relation (existing between different POS): verb -> adjective, e.g. vypracovat -> vypracova-ný (work out -> worked out), usually the derived adjec-
tives are labelled as de-adjectives, semantically it is a relation between action and its property.

8. property-adjadv relation (existing between different POS): adjective -> adverb, e.g. *rychlý -> rychl-e* (*quick -> quickly*), semantically we can speak about property.

9. property-adjnoun (existing between different POS): adjective -> noun, e.g. *rychlý -> rychl-ost* (*fast -> speed*), semantically the relation expresses property in both cases,

10. gender change relation: noun -> noun, e.g. *inženýr -> inženýr-ka* (*engineer -> she engineer*), semantically the only difference is in sex of the persons denoted by these nouns,

11. diminutive relation: noun -> noun -> noun, e.g. *dům -> dom-ek -> dom-eèek* (*house -> small house -> very little house or a house to which a speaker has an emotional attitude*), in Czech the diminutive relation can be binary or ternary,

12. augmentative relation: noun -> noun, e.g. *dub -> dub-isko* (*oak tree -> huge, strong oak tree*), semantically it expresses different emotional attitudes to a person or object,

13. possessive relation (existing between different POS): noun -> adjective *otec -> otcùv* (*father -> father’s*), semantically it is a relation between an object (person) and its possession.

14. the last D-relation exploits prefixes, in fact, it represents a whole complex of D-relations holding between verbs only, i.e.: verb -> verb, e.g. *nést -> od-nést* (*carry -> carry away*), *tancovat -> dotancovat* (*dance -> finish dancing*). We will say more about them below.

The 25 selected suffixes in Table 1 express a number of semantic relations, particularly Action (deverbative nouns), Property, Possessive, Agentive, Instrument, Location, Gender Change and Diminutive. Result and Augmentative relation are not included in the Table 1.
Table 1. Selected D-relations with suffixes implemented in Czech WordNet

<table>
<thead>
<tr>
<th>Label</th>
<th>Parts of speech</th>
<th>Meaning</th>
<th>No of literals</th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>deriv-na</td>
<td>noun -&gt; adj</td>
<td>Property</td>
<td>641</td>
<td>-í,</td>
</tr>
<tr>
<td>deriv-pos</td>
<td>noun -&gt; adj</td>
<td>Possessive</td>
<td>4037</td>
<td>-ív, -in</td>
</tr>
<tr>
<td>deriv-an</td>
<td>adj -&gt; noun</td>
<td>Property</td>
<td>1930</td>
<td>-ost</td>
</tr>
<tr>
<td>deriv-aad</td>
<td>adj -&gt; adverb</td>
<td>Property</td>
<td>1416</td>
<td>-e, -l</td>
</tr>
<tr>
<td>deriv-dvrb</td>
<td>verb -&gt; noun</td>
<td>Action</td>
<td>5041</td>
<td>-í, -ní</td>
</tr>
<tr>
<td>deriv-ag</td>
<td>verb -&gt; noun</td>
<td>Agentive</td>
<td>186</td>
<td>-tel, -ík, -ák, -ec</td>
</tr>
<tr>
<td>deriv-instr</td>
<td>verb -&gt; noun</td>
<td>Instrument</td>
<td>150</td>
<td>-tko, -ík</td>
</tr>
<tr>
<td>deriv-loc</td>
<td>verb -&gt; noun</td>
<td>Location</td>
<td>340</td>
<td>-iště, -isko</td>
</tr>
<tr>
<td>deriv-ger</td>
<td>verb -&gt; adj</td>
<td>Property</td>
<td>1951</td>
<td>-icí, -ající, -ející</td>
</tr>
<tr>
<td>deriv-pas</td>
<td>verb -&gt; adj</td>
<td>Passive</td>
<td>9801</td>
<td>-en, -it</td>
</tr>
<tr>
<td>deriv-g</td>
<td>noun -&gt; noun</td>
<td>Gender</td>
<td>2695</td>
<td>-ka</td>
</tr>
<tr>
<td>deriv-dem</td>
<td>noun -&gt; noun</td>
<td>Diminutive</td>
<td>3695</td>
<td>-ek, -èèek, -ièka, -uška</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>31429</strong></td>
<td></td>
</tr>
</tbody>
</table>

The abbreviated labels used in Czech WordNet can be seen in the Table 1 and 2.
2.1.3 Prefixes

The core of the primary 14 prefixes contains the following ones: *do-* (to), *na-* (on, at), *nad-* (above, up), *od-* (from, away), *pro-* (for, because), *při-* (by, at), *pře-* (over), *roz-* (over), *s-/se-* (with, by), *u-* (at, near), *v-/ve-* (in, up), *vy-* (out, off) *z-/ze-* (of, off), *za-* (over, behind). The English equivalents are all phrasal verbs (though English does have verbal prefixes like out and over), reflecting the difference between an inflectional and an analytic language; English, unlike Czech, is undergoing a change from the former to the latter.

Prefix D-relations hold only among verbs, typically between a stem or basic form and the respective prefix. It can be seen that semantics of the prefix D-relations is different from the suffix ones because they hold between verbs that usually denote actions, processes, events and states.

Table 2 shows the analysis of Czech prefixes, indicates the semantic nature of the D-relations, and shows the number of literals generated by the individual D-relations:

The 4 selected prefixes in Table 2 denote a number of semantic relations such as location, time, intensity of action, various kinds of motion (see below), iterativity (repeated motion or action in general) and some others. It is obvious that they differ significantly from suffix based D-relations since they hold only between verbs. In the following we will show how they combine with the selected verbs of motion. Presently, we have explored the 4 following prefix D-relations:

**Table 2.** D-relations with prefixes implemented in Czech WordNet

<table>
<thead>
<tr>
<th>Label</th>
<th>Parts of speech</th>
<th>Meaning</th>
<th>No of literals</th>
<th>Prefixes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>prefix</em> do- (to, at)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deriv-act-t verb -&gt; verb finishing motion</td>
<td></td>
<td>173</td>
<td>do-</td>
<td></td>
</tr>
<tr>
<td>deriv-act-t-iter verb -&gt; verb finishing mot iterative</td>
<td></td>
<td>24</td>
<td>do-</td>
<td></td>
</tr>
<tr>
<td>prefix od- (from, off)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deriv-mot-from verb -&gt; verb motion from</td>
<td></td>
<td>187</td>
<td>od-</td>
<td></td>
</tr>
<tr>
<td>deriv-mot-from-iter verb -&gt; verb mot. from iterative</td>
<td></td>
<td>25</td>
<td>od-</td>
<td></td>
</tr>
<tr>
<td>deriv-oblig verb -&gt; verb Obligation</td>
<td></td>
<td>2</td>
<td>od-</td>
<td></td>
</tr>
</tbody>
</table>
prefix \textit{pře-} (over)

| deriv-mot-over | verb -> verb | motion over a place | 207 | pře- |
| deriv-mot-over-it | verb -> verb | motion over a place iteratively | 21 | pře- |

prefix \textit{při-} (to, at)

| deriv-mot-to | verb -> verb | Motion to a place | 171 | při- |
| deriv-mot-to-iter | verb -> verb | Motion to a place iteratively | 18 | při- |
| deriv-add | verb -> verb | Additivity | 3 | při- |
| Total | | | 743 | |

Note that the D-relation Iterative is a subset of the verbs of motion, thus we do not count iterative verbs here as a new group. We also deal only with verbs of motion that have one argument, i.e. the moving Agent (\textit{jít, walk/go}). Verbs of motion with two arguments like \textit{nést} (carry) are not included here though they represent quite a large number of the motion verbs. They are also not pure motion verbs but cross over into contact and transfer ("I bring you flowers").

2.1.4 Semantic classes of verbs and prefixes

The relation between semantic classes of verbs and verb prefixes should be mentioned here because in Czech WordNet we adduce for each verb the semantic class it belongs to.

The approaches to the semantic classes of verbs, particularly Levin’s classification of English verbs [12] and its extension by Palmer ([18]), are based on argument alternations whose nature is mostly syntactic. For instance, verbs that show a transitive-inchoative alternation (like \textit{break}) not only share this particular syntactic behavior but are semantically similar in that they denote changes of state or location.

Levin's list of the most frequent English verbs falls into over 50 classes (most of them with several subclasses); Palmer's VerbNet project has extended this work to 395 classes. These verb classes have been translated and adapted for the Czech language. Presently, we work with approximately 100 semantic verb classes in the VerbaLex database of Czech valency frames containing approx. 12 000 verbs.
In this approach to the verb classification in Czech we exploit the verb valency frames that contain semantic roles. It appears that the verb classes established using semantic roles can be well compared with the classes obtained by the alternations, however, according to our results the classes obtained by means of the semantic roles appear to be semantically more consistent.

The third approach is based on the meanings of prefixes. The function of prefixes in Czech is to classify verbs, yielding rather small and even more consistent semantic classes of verbs. Using prefixes as sorting criteria we obtain classes that are visibly closer to the real lexical data due to the fact that the prefixes are well established formal means. For example, let’s take prefix do- (it corresponds to the English preposition to or at) and apply it to the larger group of verbs of motion (approx. 1200). The result is a group containing 173 Czech verbs denoting finishing motion. The verb classes based on prefix criteria will be examined more thoroughly in future research.

2.2 Derivational relations in the English WordNet

Many traditional paper dictionaries include derivational word forms but list them as run-ons without any information on their meaning, relying on the user's knowledge of morphological rules. Dorr and Habash ([9]), recognizing the importance of morphology-based lexical nests for NLP, created "CatVar," a large-scale database of categorial variations of English lexemes. CatVar relates lexemes belonging to different syntactic categories (part of speech) and sharing a stem, such as hunger (n.), hunger (v.) and hungry (adj.). CatVar is a valuable resource containing some 100,000 unique English word forms; however, no information is given on the words' meanings.

2.2.1 Morphosemantics relations

Miller and Fellbaum ([15]) describe the addition of "morphosemantic links" to WordNet ([14], [6]), which connect words (synset members) that are similar in meaning and where one word is derived from the other by means of a morphological affix. For example, the verb direct (defined in WordNet as "guide the actors in plays and films") is linked to the noun director (glossed as "someone who supervises the actors and directs the action in the production of a show"). Another link was created for the verb-noun pair direct/director, meaning "be in charge of" and "someone who controls resources and expenditures," respectively. Most of these links connect words from different classes (noun-verb, noun-adjective, verb-adjective), though there are also noun-noun pairs like gang-gangster. English has many such affixes and associated meaning-change rules (Marchand,1969).

2.2.2 Adding semantics to the morphosemantic links

When the morphosemantic links were added to WordNet, their semantic nature was not made explicit, as it was assumed — following conventional wisdom — that the
meanings of the affixes are highly regular and that there is a one-to-one mapping between the affix forms and their meanings. But ambitious NLP tasks and automatic reasoning require explicit knowledge of the semantics of the links. Fellbaum, Osherson and Clark ([7]) describe on-going efforts to label noun-verb pairs with semantic "roles" such as Agent (direct-director) and Result (produce-product). The assumption was that there was a one-to-one mapping between affixes and meanings. Fellbaum et al. extracted all noun-verb pairs with derivational links from WordNet and grouped them into classes based on the affix. They manually inspected each affix class expecting to find only a limited number of exceptions in each class. Instead, they found that the affixes in each class were polysemous, i.e., a given affix yields nouns that bear different semantic relations to their base verbs.

Table 3 shows Fellbaum et al.'s [7] semantic classification of -er noun and verb pairs, with the number of pairs given in the right-hand column.

<table>
<thead>
<tr>
<th>Agent</th>
<th>2,584</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument</td>
<td>482</td>
</tr>
<tr>
<td>Inanimate agent/Cause</td>
<td>302</td>
</tr>
<tr>
<td>Event</td>
<td>224</td>
</tr>
<tr>
<td>Result</td>
<td>97</td>
</tr>
<tr>
<td>Undergoer</td>
<td>62</td>
</tr>
<tr>
<td>Body part</td>
<td>49</td>
</tr>
<tr>
<td>Purpose</td>
<td>57</td>
</tr>
<tr>
<td>Vehicle</td>
<td>36</td>
</tr>
<tr>
<td>Location</td>
<td>36</td>
</tr>
</tbody>
</table>

Examination of other morphological patterns showed that polysemy of affixes is widespread. Thus, nouns derived from verbs by -ion suffixation exhibit regular polysemy between Event and Result readings (the exam lasted two hours/the exam was lying on his desk, [20].

Fellbaum et al. [7] also found one-to-many mappings for semantic patterns and affixes: a semantic category can be expressed by means of several distinct affixes, though there seems to be a default semantics associated with a given affix. Thus, while many -er nouns denote Events, event nouns are regularly derived from verbs via -ment suffixation (bomb-bombardment, punish-punishment, etc.) Patterns are partly predictable from the thematic structure of the verb. Thus, nouns derived from
unergative verbs (intransitives whose subject is an Agent) are Agents, and the pattern is productive: *runner, dancer, singer, speaker, sleeper*, etc. Nouns derived from un-
accusative verbs (intransitives whose subject is a Patient/Undergoer) are Patients: 
*breaker* (wave), *streamer* (banner), etc. This pattern is far from productive: *faller, 
? ? ? arriver, leaver*, etc. Many verbs have both transitive (causative) and intransitive 
readings (cf. [12]):

(1) a. The cook roasted the chicken
b. The chicken was roasting

2.2.3 How many semantic relations are there?

For many such verbs, there are two corresponding readings of the derived nouns:
both the *host* in (1a) and the chicken in the (1b) can be referred to as a 
roaster. Other 
examples of Agent and Patient nouns derived from the transitive and intransitive 
readings of verbs are (best)seller, (fast) developer, broiler. But the pattern is not 
productive, as nouns like 
cracker, stopper, and 
freezer show.
For virtually all -er pairs that we examined, the default agentive reading of the 
noun is always possible, though it is not always lexicalized. Thus a person who plants 
trees etc. could well be referred to as a planter, but under this reading the noun seems 
infrquent enough not to deserve an entry in most lexicons. Speakers easily generate 
and process ad-hoc nouns like planter (gardener), but only in its (non-default) location 
reading (“pot”) is the noun part of the lexicon, as its meaning cannot be guessed 
from its structure.

We focused here on the –er class. But we note that the suffixes for Czech dis-
cussed earlier have close English correspondences.

The semantic relations that were identified by Fellbaum et al. are doubtless some-
what subjective. Other classifiers might well come up with more coarse-grained or 
finer distinctions. Nevertheless, it is encouraging to see that this classification over-
laps largely with that for Czech suffixes, which was arrived independently. In addition, 
the English relations are a subset of those identified by Clark and Clark ([3]), 
who examined the large number of English noun-verb pairs related by zero-affix 
morphology, i.e., homographic pairs of semantically related verbs and nouns (roof, 
lunch, Xerox, etc.) This is the largest productive verb-noun class in English, and 
Clark and Clark's relations include not only Agent, Location, Instrument and Body 
Part, but also Meals, Elements, and Proper Names.

2.2.4 Related work

In the context of the EuroWordNet project ([23]), Peters ([19], n.d.) manually 
established noun-verb and adjective-verb pairs that were both morphologically and 
semantically related. Of the relations that Peters considered, the following match the 
ones we identified: Agent, Instrument, Location, Patient, Cause. (Peters's 
methodology differed from that of Fellbaum et al., who proceeded from the 
previously classified morphosemantic links and assumed a default semantic relation
for pairs with a given affix. Peters selected pairs of word forms that were both morphologically related and where at least one member had only a single sense in WordNet. These were then manually disambiguated and semantically classified, regardless of regular morphosemantic patterns.)

2.3 Derivational relations in Bantu

Derivational morphology in Bantu constitutes a combination of morphemes, which may either produce a new word in a different word category or may leave the word category (class membership) unchanged. Firstly, types of derivation that produce another word class include nouns, verbs, adverbs and ideophones derived from other word categories. The derivation process of nouns from verbs (deverbatives) is the most productive, and is therefore singled out in this discussion. The Bantu language Zulu is used for illustrative purposes.

When nouns are derived from verb roots, a noun prefix as well as a deverbative suffix is required, as illustrated in the following examples of nouns formed from the verb root \textit{-fund-} `learn':

\begin{itemize}
  \item \textbf{u-m(u)-fund-}i `student' \hspace{1cm} (in Czech the corresponding root is \textit{uč-})
  \item \textbf{i-m-fund-o} `education' \hspace{1cm} (in Czech the corresponding root is \textit{uč-e-n-i})
  \item \textbf{i-si-fund-o} `lesson' \hspace{1cm} (no appropriate equivalent in Czech).
\end{itemize}

The deverbative suffixes in the above example are -i and -o. Such nouns may have more than one suffix if the deverbative noun is derived from a verb root that has been extended, e.g.

\begin{itemize}
  \item \textbf{u-m(u)-fund-is-i} `teacher' \hspace{1cm} (in Czech we have \textit{uč-i-t-el} (teach-er))
\end{itemize}

The suffix \textit{-is-} is a causative extension which changes the meaning of \textit{-fund-} `learn' to `cause to learn' i.e. `teach'. (Compare with English, where causatives are usually not morphologically derived, with very few exceptions like \textit{rise-raise} and \textit{fall-fell}; in most cases, causative and non-causatives are different morphemes: \textit{kill-die}, \textit{show-see}, etc.). The last suffix \textit{-i} is the deverbative suffix.

The following are general rules for the formation of nouns from verb stems, however not every verb can be treated in this way (cf. [5]):

### Table 4: D-Relations in Zulu

#### Personal deverbatives

<table>
<thead>
<tr>
<th>Prefix of personal class (i.e. noun class 1/2 or 7/8 or 9/10)</th>
<th>Verb Root</th>
<th>Suffix -i</th>
</tr>
</thead>
<tbody>
<tr>
<td>umu/aba (class 1/2) (personal class only) (most common)</td>
<td>fund (learn)</td>
<td>umfundi &quot;student&quot;</td>
</tr>
<tr>
<td></td>
<td>hamb (go, walk)</td>
<td>umhambi &quot;traveller&quot;</td>
</tr>
<tr>
<td></td>
<td>theng (buy)</td>
<td>umthengi &quot;customer&quot;</td>
</tr>
<tr>
<td></td>
<td>shumayel (preach)</td>
<td>umshumayeli &quot;preacher&quot;</td>
</tr>
<tr>
<td>isi/izi (class 7/8) (personal as well as impersonal class)</td>
<td>eb (steal)</td>
<td>isebi &quot;thief&quot;</td>
</tr>
<tr>
<td></td>
<td>thul (be silent)</td>
<td>isithuli &quot;a mute&quot;</td>
</tr>
<tr>
<td></td>
<td>gijim (run)</td>
<td>isigijimi &quot;runner, messenger&quot;</td>
</tr>
<tr>
<td>in/izin (class 9/10) (personal as well as impersonal class)</td>
<td>bong (praise)</td>
<td>imbongi &quot;royal praiser&quot;</td>
</tr>
</tbody>
</table>

#### Impersonal deverbatives

<table>
<thead>
<tr>
<th>Prefix of impersonal class (i.e. noun class 3/4 or 5/6 or 7/8 or 9/10 or class 11)</th>
<th>Verb root</th>
<th>Suffix -o</th>
</tr>
</thead>
<tbody>
<tr>
<td>umu/imi (class 3/4) (impersonal class only)</td>
<td>buz (ask)</td>
<td>umbuzo &quot;question&quot; (result)</td>
</tr>
<tr>
<td>(li)/ama (class 5/6) (personal as well as impersonal class)</td>
<td>ceb (devise, contrive)</td>
<td>icebo &quot;plan, scheme&quot; (result)</td>
</tr>
<tr>
<td>isi/izi (class 7/8) (personal as well as impersonal class)</td>
<td>aphul (break)</td>
<td>isaphulo &quot;rupture&quot; (result)</td>
</tr>
<tr>
<td>in/izin (class 9/10) (personal as well as impersonal class)</td>
<td>phuc (shave)</td>
<td>impuco &quot;razor&quot; (instrument)</td>
</tr>
<tr>
<td>u(lu) (class 11) (impersonal class only)</td>
<td>thand (love)</td>
<td>uthando &quot;love&quot; (abstract)</td>
</tr>
</tbody>
</table>
Impersonal deverbatives indicate the following semantic relations:

a) Instrument of the action signified by the verb
b) Result of an action is conveyed
c) Abstract idea conveyed by the verb

As can be seen in column 1 of the above table of impersonal deverbatives, there is overlap in the semantic content of classes (i.e. personal and impersonal), which makes the choice of correct class prefix rather unpredictable.

Exceptions to the general rule also occur, e.g. the impersonal noun umsebenzi (um-sebenz-i) “work” is derived from the verb root –sebenz- (work), but uses the “personal” suffix –i.

Secondly, derivations that produce a derived form of the same word class include diminutives, feminine gender, augmentatives and locatives, as illustrated in the following table:

<table>
<thead>
<tr>
<th>Noun</th>
<th>Prefix</th>
<th>Suffix</th>
<th>Derived form</th>
</tr>
</thead>
<tbody>
<tr>
<td>isitsha (dish)</td>
<td></td>
<td>ana (diminutive)</td>
<td>isitshana (small dish)</td>
</tr>
<tr>
<td>Intaba (mountain)</td>
<td></td>
<td>kazi (augmentative)</td>
<td>intabakazi (big mountain)</td>
</tr>
<tr>
<td>imvu (sheep)</td>
<td></td>
<td>kazi (feminine gender)</td>
<td>imvukazi (ewe)</td>
</tr>
<tr>
<td>ikhaya (home)</td>
<td>e</td>
<td>(locative)</td>
<td>ekhaya (at home)</td>
</tr>
<tr>
<td>indlu (house)</td>
<td>e</td>
<td>(locative)</td>
<td>endlini (in the house)</td>
</tr>
</tbody>
</table>

Although locativised nouns such as ekhaya (at home) may also be used to function as adverbs, they continue to exhibit certain characteristics of regular nouns, for instance functioning as subjects and objects and in the process triggering agreement.

3 Similarities and Differences for English, Czech and Bantu

A comparison of the D-relations in three languages indicates that Czech and a Bantu language such as Zulu are in a certain respect formally closer than Czech and English. This is due to rich system of affixes in both languages, though they are not exploited in the same way in Czech and Zulu. Similarity consists in highly developed prefixation and suffixation; in Zulu both are used in a way that is typical for agglutinative languages, in particular for noun prefixes. In Czech prefixation is typical mostly for verbs and deverbatives which are, in fact, verbs as well.

English also has verbal prefixes (e.g. out- prefixes to intransitive verbs and makes them transitive: I outran the bear) but makes regular use of separate particles to form phrasal verbs (look up/down/away, etc.).

What all three languages share is the small number of semantic relations expressed by morphemes that create new words. The analyses of Czech, English and Zulu presented here allow us to predict that these D-relations are likely to be universal. All three languages use morphological processes to regularly and productively
derive such semantic categories as Agent, Instrument, Location, Gender, Diminutiveness, Augmentation, Result as well as others.

4 D-relations in WordNet among literals (screenshots of Czech and Princeton WordNets)

The screenshot below indicates how D-relations are visually represented in Czech [17] and English WordNet using the browser and editor DebVisdic. The example shows the verb tancovat:tančit:1 –dance:1 in Czech WordNet and PWN 2.0. (We cannot show the verb dance in PWN 3.0 where the respective D-relations are more complete since it has not been converted yet for browsing in DebVisdic.)

5 Conclusions

We present an analysis of some basic and highly regular D-relations in English, Czech and Bantu. It is possible to enrich both Czech and English WordNets considerably with the derivational nests (subnets), and this kind of enrichment makes these resources more useful suitable for some applications involving searching. Finally, we tried to show how the Czech and English experience can be applied in building WordNets for Bantu languages.
Another motivation for our work comes from the hypothesis that the derivational relations and derivational subnetts reflect basic cognitive structures expressed in natural language. Such structures should be explored also in terms of ontological work.

We hope that the work reported here will stimulate similar work in other languages and allow insights into their morphological processes as well as facilitate the computational representation and treatment of crosslinguistic morphological processes and relations.

References

On the Categorization of Cause and Effect in WordNet

Cristina Butnariu and Tony Veale
School of Computer Science and Informatics,
University College Dublin, Dublin 4, Ireland
{Ioana.Butnariu, Tony.Veale}@ucd.ie

Abstract. The task of detecting causal connections in text would benefit greatly from a comprehensive representation of Cause and Effect in WordNet, since previous studies show that semantic abstractions play an important role in the linguistic detection of semantic relations, in particular the cause-effect relation. Based on these studies on causality, and on our own general intuitions about causality, we propose a cover-set of different WordNet categories to represent the ontological classes of Cause and Effect. We also propose a corpus-based approach to the population of these categories, whereby candidate words and senses are identified in a large corpus (such as the Google N-gram corpus) using specific syntagmatic patterns. We describe experiments using the Cause-Effect dataset from the 2007 SemEval workshop to evaluate the most effective combinations of WordNet categories and corpus data. Ultimately, we propose extending the WordNet category of Causal-Agent with the word-senses identified by this experimental exploration.

Keywords: semantic relations, WN categorization, cause, effect, causality, syntagmatic patterns.

1 Introduction

Causality plays a fundamental role in textual inference, not just because it is intrinsic to notions of cause and effect, but also because it is central to the meaning of artifacts, agents, products (whether physical or abstract) and even natural phenomena. Artifacts possess a purpose, or telicity, that is causally defined, while agents are often defined by the products that they cause to exist, and natural phenomena like storms and other acts of god are typically conceptualized as intentional processes. Since each of these notions – agents, artifacts, products and natural phenomena – are all explicitly represented and richly specialized in a lexical ontology like WordNet [4], one can ask whether the concepts of Cause and Effect can and should be as richly represented in WordNet. Of course, since these concepts correspond to the nouns “cause” and “effect”, they clearly are represented in WordNet. Indeed, WordNet represents different nuances of these concepts, distinguishing between cause-as-agent (or {causal-agent}) and cause-as-reason (or {cause, reason, grounds}) and effect-as-outcome and effect-as-symptom.

Nonetheless, these attempts at ontologizing causality are simultaneously too coarse-grained – insofar as they admit of too many specializations that are not
meaningfully represented as causes or effects – or too under-developed – insofar as they are little more than ontological place-holders that have few meaningful specializations. For instance, because WordNet defines the concept Causal-agent as a hypernym of Person, concepts like Victim, Martyr and Casualty will be seen indirectly as agents of their own state, even when this view is counter to their true meaning (these concepts are clearly better defined as causal-patients, though WordNet lacks such a concept). Likewise, WordNet categorizes antacids and other medicinally helpful substances (such as antacids) as causal agents but denies this classification to unhelpful substances such as poisons and allergens, as well as to harmful weather phenomena (such as storms and earthquakes) that are readily conceptualized as major causes by humans. Similarly, WordNet 2.1 only provides four possible specializations of the symptom meaning of Effect when any number of other WordNet concepts can, in the right circumstance, by seen as symptoms. Indeed, only 30% of the concepts whose WordNet 2.1 gloss contains the phrase “that causes” are categorized as causal agents in WordNet, even though all the concepts are valid examples of causal agency. WordNet would clearly benefit then from considerable house-cleaning under its categories of Cause (and Causal-Agent) and Effect. In this paper, we consider the effectiveness of WordNet in recognizing and capturing cause and effect relationships, by focusing on the cause-effect relation in the recent SemEval semantic-relations task (see [7]). While virtually all entrants in this task adopted a supervised machine-learning approach to the problem of detecting relations such as cause-effect between noun-pairs, we consider here how well WordNet, without training, can perform on this task when its basic causal repertoire is augmented with causally-indicative syntagmatic cues from a large corpus. In section 2 we briefly describe past-work on this topic, before presenting a purely WordNet-based approach to cause and effect in section 3. Causality is a highly contextual notion: a dinner plate is an effect (product) in the context of its construction, and a cause of pain when used as a projectile in the context of a domestic argument (see [12]). WordNet cannot hope to anticipate or reflect all of these contexts, but the language used in context-specific corpora may well reflect these causal nuances. In section 4 then, we present a corpus-based approach to identifying possible causes and effects in terms of lexico-syntactic patterns. Section 5 then presents an empirical evaluation of this corpus/WordNet combination. The paper concludes with some closing remarks in section 6.

2 Past Work

There have been many attempts in the computational linguistic communities to define and understand the Causality relation. Nastase in [11] defines causality as a general class of relations that describe how two occurrences influence each other. Further she proposes the following sub-relations of causality: cause, effect, purpose, entailment, enablement, detraction and prevention. She states that semantic relations can be expressed in different syntactic forms, at different syntactic levels. Hearst [8] states that “certain lexico-syntactic patterns unambiguously indicate certain semantic relations”. The key issue then is to discover the most efficient patterns that indicate a certain semantic relation. These patterns can be either manually specified by linguists
or discovered automatically from corpora. For instance, the subject-verb-object lexico-syntactic pattern (where subject and verb are noun-phrases) was used in [3] to detect causal relations in text, and from these patterns, automatically construct Bayesian for causal inference.

Girju proposes in [5] a classification of lexical patterns for mining instances of the causality relation from corpora, and describes a semi-automatic method to discover new patterns. She uses a general pattern \(<\text{NP1 verb NP2}>\) in combination with WordNet to impose semantic restrictions on NP1 (the Cause category) and NP2 (the Effect category). She defines the classes of Cause and Effect in WordNet terms as a patchwork of different synsets/categories. For Effect, she proposes a cover-set comprising the following synsets: \{'human\_action, human\_activity, act\}, \{'phenomenon\}, \{'state\}, \{'psychological\_feature\} and \{'event\}. However, she observes that the Cause class is harder to define in such terms of WordNet categories, since the notion of causality is frequently entwined with, and difficult to separate from, that of metonymy (e.g., does the poison cause death or the poisoner, or both? The gun or the gun-man?). She thus relies entirely on the intuitions already encoded in WordNet under the category of Causal-Agent. Girju then ranks the output patterns into five categories, according to their degree of ambiguity. She reports a precision of 68% when applying these patterns to a terrorism corpus.

The SemEval-2007 task 4 (see [7]) concerned itself with the classification of semantic relations between pairs of words in a given context. Seven semantic relations were proposed and a training dataset for each semantic relation (comprising positive and negative examples, the latter in the form of near misses) was collected from the web and classified by two human judges. The relation that interests us here is the Cause-Effect relation, which the task authors define as follows: "Cause-Effect(X,Y) is true for a sentence S if X and Y appear close in the syntactic structure of S and the situation described in S entails that X is the cause of Y." There are some restrictions imposed on X and Y: "X and Y can be a nominal denoting an event, state, activity or an entity, as a metonymic expression of an occurrence." The data-set for this relation comprises 220 noun pairs (with WordNet sense-tags and associated context fragments), of which 114 pairs are positive exemplars and 106 are negative "near-miss" exemplars.

## 3 Defining Cause and Effect in WordNet terms

Following Girju, we should intuitively expect a variety of high-level WordNet abstractions to encompass a range of concepts that play an enabling role in achieving certain ends, and thus to contribute to the cover-set that defines the class of Causes. Recall that Girju limits the definition of Cause to the WordNet category \{'causal\_agent\}, a snapshot of which is presented in Figure 1.
In contrast, we broaden the cover-set of Causes to include the following WordNet categories and their descendants: \{causal_agent\}, \{psychological_feature\}, \{attribute\}, \{substance\} (insofar as many are biological causal-agents), \{phenomenon\}, \{communication\} (insofar as they can drive agents to action), \{natural_action\} and \{organic_process\}. In contrast, the class of Effects should include: \{psychological_feature\}, \{attribute\}, \{physical_process\}, \{phenomenon\}, \{natural_action\}, \{possession\} and \{organic_process\}. The two cover-sets are similar because causes and effects typically interact as part of complex causal chains, so the causes of one effect are often themselves the effects of prior causes.

It is worth considering how well these WordNet-based cover-sets correspond to the exemplars of the SemEval dataset. Figure 2 reveals the coverage obtained for both the positive and negative exemplars by each WordNet category in the class of Causes. Note how the category Causal-Agent offers very little coverage for the positive exemplars (i.e., most of the actual causes in that data-set are not categorized as causal-agents in WordNet), and actually offers higher coverage for the negative exemplars (making it more likely to contribute to a classification error in the case of a near-miss).
Figure 2. The coverage (%) offered by different WordNet categories for SemEval positive and negative exemplars of the \textit{Cause} class.

Figure 3 presents a comparable analysis for the WordNet categories that comprise the cover-set for the class of Effects. Note how the category \{psychological\_feature\} looms large as both a Cause and an Effect in the SemEval data-set.

Figure 3. The coverage (%) offered by different WordNet categories for SemEval positive and negative exemplars of the \textit{Effect} class.
4 Defining Cause and Effect in Syntagmatic terms

Girju in [5] notes that certain lexico-syntactic patterns are indicative of causal relations in text, but that some patterns are more ambiguous than others. For instance, the patterns "NP2-causing NP1" and "NP1-caused NP2" are explicit and largely unambiguous cues to the interpretation of NP1 as a cause and NP2 as an effect. In contrast, Girju notes that "NP2-inducing NP1" and "NP2-generated NP1" are equally explicit but potentially more ambiguous patterns for identifying cause and effect in text. Nonetheless, the pattern "NP-induced NP" does occur quite frequently in large corpora, and does designate causes with high accuracy and low ambiguity. However, this triple of "NP-induced/inducing NP" produces a spare space of associations between different causes and effects, so it is more productive to consider each noun-phrase in isolation.

Thus, we look for the patterns "Noun-inducing" and "Noun-causing" in a large corpus to identify those nouns that can denote effects, as in the phrase "headache-inducing". Our corpus is the set of Google N-grams (see [1]), from which the above pairings can easily be mined. Similarly, we mine the patterns "Noun-induced" and "Noun-caused" from these n-grams to identify a large set of nouns that can denote causes, as in "caffeine-induced". In addition, we look to the patterns "-induced Noun" and "-caused Noun" to identify a further collection of possible effect nouns, and the patterns "-inducing Noun" and "-causing Noun" to identify further cause nouns. In this way, we obtain 3,500+ nouns as denoting potential causes, and 4,200+ nouns as denoting potential effects. Table 1 presents the top-ranked (by frequency) causes and effects in this data, as well as the top-ranked causality pairs (i.e., cause associated with specific effect).

<table>
<thead>
<tr>
<th>CAUSE-EFFECT pairs</th>
<th>CAUSE nouns</th>
<th>EFFECT nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>(organism, disease)</td>
<td>Drug</td>
<td>apoptosis</td>
</tr>
<tr>
<td>(laser, fluorescence)</td>
<td>stress</td>
<td>disease</td>
</tr>
<tr>
<td>(noise, hearing)</td>
<td>radiation</td>
<td>cancer</td>
</tr>
<tr>
<td>(chemical, cancer)</td>
<td>exercise</td>
<td>changes</td>
</tr>
<tr>
<td>(agent, cancer)</td>
<td>self</td>
<td>cell</td>
</tr>
<tr>
<td>(exercise, asthma)</td>
<td>laser</td>
<td>increase</td>
</tr>
<tr>
<td>(collagen, arthritis)</td>
<td>human</td>
<td>activation</td>
</tr>
<tr>
<td>(bacteria, disease)</td>
<td>acid</td>
<td>asthma</td>
</tr>
<tr>
<td>(pregnancy, hypertension)</td>
<td>light</td>
<td>inhibition</td>
</tr>
<tr>
<td>(human, climate)</td>
<td>virus</td>
<td>odor</td>
</tr>
</tbody>
</table>

Because the Google N-grams corpus is not sense-tagged, we can only guess at the senses of the nouns in Table 1. However, if we assume that each noun is used in one of its two most frequent senses, then we can assign these nouns to various WordNet categories, as we did for the SemEval nouns in Figures 2 and 3. Following this heuristic assignment of senses, Figure 4 presents the distribution of cause nouns to different WordNet cause categories.
A comparable distribution for effect nouns is displayed in Figure 5.

Because some noun senses belong to multiple categories, and because we use the two most frequent senses of each noun, the sum total of distributions in Figures 2 to 5 may exceed 100%. Note also that certain patterns are noisier than others. While "Noun-inducing" is a tight and rather unambiguous micro-context in which to recognize Noun as an effect, ":-induced Noun" is more prone to leakage. For instance, "drug-induced liver failure" yields "drug" as an unambiguous cause, but mistakenly suggests "liver" as an effect. Given that "Noun-induced" is a more frequent pattern than "Noun-inducing", the set of nouns designed as effects is noisier than the set of nouns designed as causes. For this reason, the Other category in Figure 5 is more populous than the Other category in Figure 4. The most frequently misclassified nouns in the Effect class are: protein, liver, gene, lung, acute, platelet, insulin, diabetic, skin, calcium, rat, cytotoxicity, genes, immune, and bone.
5 Empirical results

We can test the approaches of section 3 and 4 in a variety of guises and combinations:

The **WordNet-only approach** (as described in section 3): a word pair \(<X,Y>\) can be classified as a Cause-Effect pairing if and only if any of the two most frequent senses of X fall under a synset in the Cause cover-set and any of the two most frequent senses of Y fall under a synset in the Effect cover-set.

The **Corpus-only approach** (as described in section 4): a word pair \(<X,Y>\) can be classified as a Cause-Effect pairing if and only if X is found in the set of nouns that have been identified as cause nouns (e.g., because the pattern “X-induced” was found in the corpus) and Y is found in the set of effect nouns (e.g., because the pattern “Y-inducing” or “-induced Y” was found in the corpus). In our experiments we test two different sets of corpus-mining patterns: a minimal set based on just two causation verbs, *induce* and *cause*, and an extended set comprising variations of the verbs *induce, cause, power, fuel, activate, enable, control* and *operate*.

The **Hybrid approach** (WordNet used in combination with corpus-derived data): a word pair \(<X,Y>\) can be classified as a Cause-Effect pairing if any of the two most frequent senses of X fall under a synset in the Cause cover-set and a synonym of one these two senses (i.e., any word from the same two synsets) is found in the set of corpus-derived cause nouns, and if any of the two most frequent senses of Y fall under a synset in the Effect cover-set and a synonym of one these two senses of Y (or Y itself) is found in the set of effect nouns. The hybrid approach is thus a logical conjunction of the WordNet and corpus approaches, but one that includes synonyms of the words X and Y, so the corpus-data of the latter is effectively smoothed and made less sparse.

Table 2 presents empirical results for each of these approaches on the SemEval cause-effect data-set and the All-true baseline which always guesses “true” (and thereby maximizes recall). Interestingly, the WordNet-only approach has the best overall performance (F-score), which accords with the observations of the SemEval organizers: the statistics show that WordNet plays an important role in the task of relation classification.

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Total no</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. WordNet only approach</td>
<td>61.3</td>
<td>85</td>
<td>71.3</td>
<td>220</td>
</tr>
<tr>
<td>B. Corpus-only approach</td>
<td>54</td>
<td>60</td>
<td>62.3</td>
<td>220</td>
</tr>
<tr>
<td>Using {induce, cause} patterns</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Hybrid A+B approach</td>
<td>63.5</td>
<td>70</td>
<td>66.8</td>
<td>220</td>
</tr>
<tr>
<td>D. Corpus-only approach</td>
<td>51.6</td>
<td>83</td>
<td>63.6</td>
<td>220</td>
</tr>
<tr>
<td>using {induce, cause, power, fuel, activate, enable, control, operate} patterns</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. Hybrid A+D approach</td>
<td>60</td>
<td>85</td>
<td>70.3</td>
<td>220</td>
</tr>
<tr>
<td>All-true baseline</td>
<td>51.8</td>
<td>100</td>
<td>68.2</td>
<td>220</td>
</tr>
</tbody>
</table>
5.1 Analysis of Results

As the corpus yields a somewhat sparse and noisy data set of candidate cause and effect nouns, the corpus approach (B) that uses just *cause* and *induce* as causal markers achieves only 60% recall, with a low precision of 54%. The WordNet contribution in the Hybrid A+B approach boosts recall by 10% while also increasing precision. Recall is improved since the sparse corpus data is extrapolated by the use of WordNet synonyms; precision is also improved somewhat, over that of the WordNet-only approach (A) and the simple corpus approach (B) because WordNet’s category restrictions help to filter out some noisy and misclassified effect nouns. Nonetheless, there is need for more corpus data to increase the recall of the hybrid approach even further. In the second corpus approach (D), recall is boosted by using patterns based on a broader list of causative verbs (see [9]) to identify cause and effect nouns: {induce, cause, power, fuel, activate, enable, control, operate}. Note that when WordNet Cause and Effect categories of (A) are used to filter noisy classifications in the hybrid approaches, this imposes a WordNet-based ceiling of 85% (i.e., the recall of A) on the recall of the hybrid approaches: the tradeoff results in a lower precision but a better F-measure overall.

Each approach in Table 2 (WordNet-alone, corpus-alone, and the combination of both) is unsupervised and does not avail of the WN sense information provided for nouns in the SemEval data-set. Our best F-measure is 71.3% and is comparable with the 72% F-measure obtained by the best performing system in the corresponding SemEval category (i.e., category A, in which competing systems do not avail of WordNet sense tags). The relatively low precision is largely explained by the fact that SemEval’s negative examples are near misses rather than random examples of non-causal relationships. Our recorded precision is a lower-bound then for what one might expect on random word-pairings drawn from a real text.

6 Concluding Remarks

In this paper we presented three unsupervised approaches to the classification of causal-relations among noun-pairs: a corpus-based approach, an ontological WordNet-based approach, and a combination of both. The results achieved by these approaches on the SemEval dataset are encouraging, especially given the fact that these approaches do not apply machine-learning techniques to a training data-set. The WordNet categories which form the substance of the ontological approach, and which also contribute substantially to the combined approach, are hand-picked based on human intuitions about causality. However, a machine-learning approach to identifying these categories automatically is a topic of current research. As reflected in the superior performance of the WordNet-only approach, WordNet does have the capability to accurately represent high-level abstractions like Cause and Effect, and to do so in a non-trivial way that spans large numbers of more specific specializations.

Nonetheless, our results also bear out our initial observation that the WordNet category of Causal-agent is very weakly represented and in serious need of re-organization, at least if it is to properly serve its intended purpose. In the SemEval
data analyzed here, the \{causal\_agent\} category covers only 2% of the Cause instances in the positive exemplar set, and just 8% of the negative "near-miss" exemplars. Extension to this WordNet category can clearly be performed using intuition-guided ontological-engineering as well as corpus-based discovery. Based on our results then, we might ask which WordNet concepts should be included under the newly organized umbrella term of Causal-Agent, and under a new category, Causal-Patient? We suggest the word senses that satisfy approach E will make excellent candidates to populate these categories.

We next plan to extend the general approach described here to other classes of semantic relation, such as Content-Container, Part-Whole and Tool-Purpose, since these too combine a strong ontological dimension to their meaning with a strong usage-based (i.e., corpus-based) dimension. Overall, our results confirm that WordNet has a significantly useful role to play in the detection of semantic relations in text, but detection would be more efficient if WordNet could provide more insightful ontological classifications of the concepts underlying these relations. These ontological insights will come from using the existing structures of WordNet to hypothesize about, and filter, large quantities of relevant usage data in a corpus.

References

Evaluation of Synset Assignment to Bi-lingual Dictionary

Thatsanee Charoenporn¹, Virach Sornlertlamvanich¹, Chumpol Mokarat¹, Hitoshi Isahara², Hammam Riza³, and Purev Jaimai⁴

¹ Thai Computational Linguistics Lab., NICT Asia Research Center, Thailand Science Park, Pathumthani, Thailand
  {thatsanee, virach, chumpol}@tcllab.org

² National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho, soraku-gaun, Kyoto, Japan 619-0289
  isahara@nict.go.jp

³ IPTEKNET, Agency for the Assessment and Application of Technology, Jakarta Pusat 10340, Indonesia
  hammam@iptek.net.id

⁴ Center for Research on Language Processing, National University of Mongolia, Ulaanbaatar, Mongolia
  purev@num.edu.mn

Abstract. This paper describes an automatic WordNet synset assignment to the existing bi-lingual dictionaries of languages having limited lexicon information. Generally, a term in a bi-lingual dictionary is provided with very limited information such as part-of-speech, a set of synonyms, and a set of English equivalents. This type of dictionary is comparatively reliable and can be found in an electronic form from various publishers. In this paper, we propose an algorithm for applying a set of criteria to assign a synset with an appropriate degree of confidence to the existing bi-lingual dictionary. We show the efficiency in nominating the synset candidate by using the most common lexical information. The algorithm is evaluated against the implementation of Thai-English, Indonesian-English, and Mongolian-English bi-lingual dictionaries. The experiment also shows the effectiveness of using the same type of dictionary from different sources.

Keywords: synset assignment

1 Introduction

The Princeton WordNet (PWN) [1] is one of the most semantically rich English lexical databases that are widely used as a lexical knowledge resource in many research and development topics. The database is divided by part of speech into noun, verb, adjective and adverb, organized in sets of synonyms, called synset, each of which represents “meaning” of the word entry. PWN is successfully implemented in many applications, e.g., word sense disambiguation, information retrieval, text summarization, text categorization, and so on. Inspired by this success, many
languages attempt to develop their own WordNets using PWN as a model, for example\textsuperscript{1}, BalkaNet (Balkans languages), DanNet (Danish), EuroWordNet (European languages such as Spanish, Italian, German, French, English), Russnet (Russian), Hindi WordNet, Arabic WordNet, Chinese WordNet, Korean WordNet and so on.

Though WordNet was already used as a starting resource for developing many language WordNets, the constructions of the WordNet for languages can be varied according to the availability of the language resources. Some were developed from scratch, and some were developed from the combination of various existing lexical resources. Spanish and Catalan Wordnets \cite{2}, for instance, are automatically constructed using hyponym relation, a monolingual dictionary, a bilingual dictionary and taxonomy \cite{3}. Italian WordNet \cite{4} is semi-automatically constructed from definitions in a monolingual dictionary, a bilingual dictionary, and WordNet glosses. Hungarian WordNet uses a bilingual dictionary, a monolingual explanatory dictionary, and Hungarian thesaurus in the construction \cite{5}, etc.

This paper presents a new method to facilitate the WordNet construction by using the existing resources having only English equivalents and the lexical synonyms. Our proposed criteria and algorithm for application are evaluated by implementing them for Asian languages which occupy quite different language phenomena in terms of grammars and word unit.

To evaluate our criteria and algorithm, we use the PWN version 2.1 containing 207,010 senses classified into adjective, adverb, verb, and noun. The basic building block is a “synset” which is essentially a context-sensitive grouping of synonyms which are linked by various types of relation such as hyponym, hypernymy, meronymy, antonym, attributes, and modification. Our approach is conducted to assign a synset to a lexical entry by considering its English equivalent and lexical synonyms. The degree of reliability of the assignment is defined in terms of confidence score (CS) based on our assumption of the membership of the English equivalent in the synset. A dictionary from a different source is also a reliable source to increase the accuracy of the assignment because it can fulfill the thoroughness of the list of English equivalent and the lexical synonyms.

The rest of this paper is organized as follows: Section 2 describes our criteria for synset assignment. Section 3 provides the results of the experiments and error analysis on Thai, Indonesian, and Mongolian. Section 4 evaluates the accuracy of the assignment result, and the effectiveness of the complimentary use of a dictionary from different sources. And Section 5 concludes our work.

2 Synset Assignment

A set of synonyms determines the meaning of a concept. Under the situation of limited resources on a language, an English equivalent word in a bi-lingual dictionary is a crucial key to find an appropriate synset for the entry word in question. The synset assignment criteria described in this section relies on the information of

\textsuperscript{1} List of WordNets in the world and their information is provided at http://www.globalwordnet.org/gwa/wordnet_table.htm
Synset Assignment Criteria

Applying the nature of WordNet which introduces a set of synonyms to define the concept, we set up four criteria for assigning a synset to a lexical entry. The confidence score (CS) is introduced to annotate the likelihood of the assignment. The highest score, CS=4, is assigned to the synset that is evident to include more than one English equivalent of the lexical entry in question. On the contrary, the lowest score, CS=1, is assigned to any synset that occupies only one of the English equivalents of the lexical entry in question when multiple English equivalents exist.

The details of assignment criteria are: \( L_i \) denotes the lexical entry, \( E_j \) denotes the English equivalent, \( S_k \) denotes the synset, and \( \in \) denotes the member of a set.

**Case 1:** Accept the synset that includes more than one English equivalent with a confidence score of 4.

Fig. 1 simulates that a lexical entry \( L_0 \) has two English equivalents of \( E_0 \) and \( E_1 \). Both \( E_0 \) and \( E_1 \) are included in a synset of \( S_1 \). The criterion implies that both \( E_0 \) and \( E_1 \) are the synset for \( L_0 \) which can be defined by a greater set of synonyms in \( S_1 \). Therefore the relatively high confidence score, CS=4, is assigned for this synset to the lexical entry.

\[
\begin{align*}
L_0 & \quad E_0 & \quad S_0 \\
L_1 & \quad E_1 & \quad S_1 \\
& & \quad S_2
\end{align*}
\]

**Example:**
\( L_0: \) ʕ البحر
\( E_0: \) aim  \( E_1: \) target
\( S_0: \) purpose, intent, intention, aim, design
\( S_1: \) aim, object, objective, target
\( S_2: \) aim

In the above example, the synset, \( S_1 \), is assigned to the lexical entry, \( L_0 \), with CS=4.

**Case 2:** Accept the synset that includes more than one English equivalent of the synonym of the lexical entry in question with a confidence score of 3.

If Case 1 fails in finding a synset that includes more than one English equivalent, the English equivalent of a synonym of the lexical entry is picked up to investigate.
Fig. 2 shows an English equivalent of a lexical entry $L_0$ and its synonym $L_1$ in a synset $S_1$. In this case the synset $S_1$ is assigned to both $L_0$ and $L_1$ with CS=3. The score in this case is lower than the one assigned in Case 1 because the synonym of the English equivalent of the lexical entry is indirectly implied from the English equivalent of the synonym of the lexical entry. The newly retrieved English equivalent may not be distorted.

![Fig. 2. Synset assignment with CS=3](image)

**Example:**

$L_0$: นิ่ง
$L_1$: นิ่ง

$E_0$: stare
$E_1$: gaze

$S_0$: gaze, stare
$S_1$: stare

In the above example, the synset, $S_0$, is assigned to the lexical entry, $L_0$, with CS=3.

**Case 3:** Accept the only synset that includes only one English equivalent with a confidence score of 2.

![Fig. 3. Synset assignment with CS=2](image)

**Example:**

$L_0$: นิ่ง

$E_0$: obstetrician

$S_0$: obstetrician, accoucheur

In the above example, the synset, $S_0$, is assigned to the lexical entry, $L_0$, with CS=2.

**Case 4:** Accept more than one synset that includes each of the English equivalents with a confidence score of 1.
Case 4 is the most relaxed rule to provide some relation information between the lexical entry and a synset. Fig. 4 shows the assignment of CS=1 to any relations that do not meet the previous criteria but the synsets include one of the English equivalents of the lexical entry.

Example:

\begin{align*}
L_0 & \in S_0 \\
E_0 &: \text{hole} \\
S_0 &: \text{hole, hollow} \\
S_1 &: \text{hole, trap, cakehole, maw, yap, gop} \\
S_2 &: \text{canal, duct, epithelial duct, channel}
\end{align*}

In the above example, each synset, \( S_0 \), \( S_1 \), and \( S_2 \) is assigned to lexical entry \( L_0 \) with CS=1.

3 Experiment Results

We applied the synset assignment criteria to a Thai-English dictionary (MMT dictionary) [6] with the synset from WordNet 2.1. To compare the ratio of assignment for Thai-English dictionary, we also investigated the synset assignment of Indonesian-English and Mongolian-English dictionaries.

In our experiment, there are only 24,457 synsets from 207,010 synsets, which is 12% of the total number of the synsets that can be assigned to Thai lexical entries. Table 1 shows the successful rate in assigning synsets to the Thai-English dictionary. About 24% of Thai lexical entries are found with the English equivalents that meet one of our criteria.

Going through the list of unmapped lexical entries, we can classify the errors into three groups:

1. Compound
   The English equivalent is assigned in a compound, especially in cases where there is no appropriate translation to represent exactly the same sense. For example,
2. Phrase
Some particular words culturally used in one language may not be simply translated into one single word sense in English. In this case, we found it explained in a phrase. For example,

L: ร้านค้า
E: retail shop
L: จับ]
E: pull sharply

3. Word form
Inflected forms, i.e., plural, past participle, are used to express an appropriate sense of a lexical entry. This can be found in non-inflected languages such as Thai and most Asian languages. For example,

L: รำลึก
E: small pavilion for monks to sit on to chant
L: จปริาย
E: bouquet worn over the ear

The above English expressions cause an error in finding an appropriate synset.

Table 1. Synset assignment to Thai-English dictionary

<table>
<thead>
<tr>
<th>WordNet (synset)</th>
<th>TE Dict (entry)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td>Noun</td>
<td>145,103</td>
</tr>
<tr>
<td>Verb</td>
<td>24,884</td>
</tr>
<tr>
<td>Adjective</td>
<td>31,302</td>
</tr>
<tr>
<td>Adverb</td>
<td>5,721</td>
</tr>
<tr>
<td>Total</td>
<td>207,010</td>
</tr>
</tbody>
</table>

We applied the same algorithm to Indonesia-English and Mongolian-English [7] dictionaries to investigate how it works with other languages in terms of the selection of English equivalents. The difference in unit of concept is basically understood to affect the assignment of English equivalents in bi-lingual dictionaries. In Table 2, the size of the Indonesian-English dictionary is about half that of the Thai-English dictionary. The success rates of assignment to the lexical entry are the same, but the rate of synset assignment of the Indonesian-English dictionary is lower than that of the Thai-English dictionary. This is because the total number of lexical entries is about in the half that of the Thai-English dictionary.

A Mongolian-English dictionary is also evaluated. Table 3 shows the result of synset assignment.

These experiments show the effectiveness of using English equivalents and synonym information from limited resources in assigning WordNet synsets.
Table 2. Synset assignment to Indonesian-English dictionary

<table>
<thead>
<tr>
<th>WordNet (synset)</th>
<th>IE Dict (entry)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td>Noun</td>
<td>145,103</td>
</tr>
<tr>
<td>Verb</td>
<td>24,884</td>
</tr>
<tr>
<td>Adjective</td>
<td>31,302</td>
</tr>
<tr>
<td>Adverb</td>
<td>5,721</td>
</tr>
<tr>
<td>Total</td>
<td>207,010</td>
</tr>
</tbody>
</table>

Table 3. Synset assignment to Mongolian-English dictionary

<table>
<thead>
<tr>
<th>WordNet (synset)</th>
<th>ME Dict (entry)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td>Noun</td>
<td>145,103</td>
</tr>
<tr>
<td>Verb</td>
<td>24,884</td>
</tr>
<tr>
<td>Adjective</td>
<td>31,302</td>
</tr>
<tr>
<td>Adverb</td>
<td>5,721</td>
</tr>
<tr>
<td>Total</td>
<td>207,010</td>
</tr>
</tbody>
</table>

4 Evaluations

In the evaluation of our approach for synset assignment, we randomly selected 1,044 synsets from the result of synset assignment to the Thai-English dictionary (MMT dictionary) for manually checking. The random set covers all types of part-of-speech and degrees of confidence score (CS) to confirm the approach in all possible situations. According to the supposition of our algorithm that the set of English equivalents of a word entry and its synonyms are significant information to relate to a synset of WordNet, the result of accuracy will be correspondent to the degree of CS.

It took about three years to develop the Balkan WordNet on PWN 2.0 [8], [9]. Therefore, we randomly picked up some synsets that resulted from our synset assignment algorithm. The results were manually checked and the details of synsets to be used to evaluate our algorithm are shown in Table 4.
Table 5 shows the accuracy of synset assignment by part of speech and CS. A small set of adverb synsets is 100% correctly assigned irrelevant to its CS. The total number of adverbs for the evaluation could be too small. The algorithm shows a better result of 48.7% in average for noun synset assignment and 43.2% in average for all part of speech.

With the better information of English equivalents marked with CS=4, the assignment accuracy is as high as 80.0% and decreases accordingly due to the CS value. This confirms that the accuracy of synset assignment strongly relies on the number of English equivalents in the synset. The indirect information of English equivalents of the synonym of the word entry is also helpful, yielding 60.7% accuracy in synset assignment for the group of CS=3. Others are quite low, but the English equivalents are somehow useful to provide the candidates for expert revision.

**Table 4.** Random set of synset assignment

<table>
<thead>
<tr>
<th></th>
<th>CS=4</th>
<th>CS=3</th>
<th>CS=2</th>
<th>CS=1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>7</td>
<td>479</td>
<td>64</td>
<td>272</td>
<td>822</td>
</tr>
<tr>
<td>Verb</td>
<td>44</td>
<td>75</td>
<td>29</td>
<td>148</td>
<td>1044</td>
</tr>
<tr>
<td>Adjective</td>
<td>1</td>
<td>25</td>
<td>32</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>Adverb</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>552</td>
<td>143</td>
<td>334</td>
<td>1044</td>
</tr>
</tbody>
</table>

**Table 5.** Accuracy of synset assignment

<table>
<thead>
<tr>
<th></th>
<th>CS=4</th>
<th>CS=3</th>
<th>CS=2</th>
<th>CS=1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>5</td>
<td>306</td>
<td>34</td>
<td>55</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>(71.4%)</td>
<td>(63.9%)</td>
<td>(53.1%)</td>
<td>(20.2%)</td>
<td>(48.7%)</td>
</tr>
<tr>
<td>Verb</td>
<td>23</td>
<td>6</td>
<td>4</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(52.3%)</td>
<td>(8.0%)</td>
<td>(13.8%)</td>
<td>(22.3%)</td>
<td></td>
</tr>
<tr>
<td>Adjective</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(8.0%)</td>
<td>(3.4%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adverb</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>(100%)</td>
<td>(100%)</td>
<td>(100%)</td>
<td>(100%)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>335</td>
<td>44</td>
<td>60</td>
<td>451</td>
</tr>
<tr>
<td></td>
<td>(80.0%)</td>
<td>(60.7%)</td>
<td>(30.8%)</td>
<td>(18%)</td>
<td>(43.2%)</td>
</tr>
</tbody>
</table>

**Table 6.** Additional correct synset assignment by other dictionary (LEXiTRON)

<table>
<thead>
<tr>
<th></th>
<th>CS=4</th>
<th>CS=3</th>
<th>CS=2</th>
<th>CS=1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>2</td>
<td>22</td>
<td>29</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>Verb</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Adjective</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adverb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>28</td>
<td>33</td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

To examine the effectiveness of English equivalent and synonym information from a different source, we consulted another Thai-English dictionary (LEXiTRON) [10]. Table 6 shows the improvement of the assignment by the increased number of correct
assignment in each type. We can correct more in nouns and verbs but not adjectives. Verbs and adjectives are ambiguously defined in Thai lexicon, and the number of the remaining adjectives is too few, therefore, the result should be improved regardless of the type.

Table 7. Improved correct synset assignment by additional bi-lingual dictionary (LEXiTRON)

<table>
<thead>
<tr>
<th></th>
<th>CS=4</th>
<th>CS=3</th>
<th>CS=2</th>
<th>CS=1</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>14</td>
<td>337</td>
<td>72</td>
<td>93</td>
<td>516</td>
</tr>
<tr>
<td></td>
<td>(93.3%)</td>
<td>(61.1%)</td>
<td>(50.3%)</td>
<td>(27.8%)</td>
<td>(49.4%)</td>
</tr>
</tbody>
</table>

Table 7 shows the total improvement of the assignment accuracy when we integrated English equivalent and synonym information from a different source. The accuracy for synsets marked with CS=4 is improved from 80.0% to 93.3% and the average accuracy is also significantly improved from 43.2% to 49.4%. All types of synset are significantly improved if a bi-lingual dictionary from different sources is available.

5 Conclusion

Our synset assignment criteria were effectively applied to languages having only English equivalents and its lexical synonym. Confidence scores were proven efficiently assigned to determine the degree of reliability of the assignment which later was a key value in the revision process. Languages in Asia are significantly different from the English language in terms of grammar and lexical word units. The differences prevent us from finding the target synset by following just the English equivalent. Synonyms of the lexical entry and an additional dictionary from different sources can be complementarily used to improve the accuracy in the assignment. Applying the same criteria to other Asian languages also yielded a satisfactory result. Following the same process that we implemented for the Thai language, we are expecting an acceptable result from the Indonesian, Mongolian languages and so on.

References

Using and Extending WordNet to Support Question-Answering

Peter Clark¹, Christiane Fellbaum², and Jerry Hobbs³

¹Boeing Phantom Works, Seattle (USA)
²Princeton University, Princeton (USA)
³USC/ISI, Marina del Rey (USA)

peter.e.clark@boeing.com, fellbaum@clarity.princeton.edu, hobbs@isi.edu

Abstract. Over the last few years there has been increased research in automated question-answering from text, including questions whose answer is implied, rather than explicitly stated, in the text. WordNet has played a central role in many such systems (e.g., 21 of the 26 teams in the recent PASCAL RTE3 challenge used WordNet), and thus WordNet is being increasingly stretched to play more semantic tasks in applications. As part of our current research, we are exploring some of the new demands which question-answering places on WordNet, and how it might be further extended to meet them. In this paper, we present some of these new requirements, and some of the extensions that we are currently making to WordNet in response.

Keywords: WordNet, question answering, textual entailment, world knowledge

1 Introduction

Advanced question-answering is more than simply fact retrieval; typically, much of the knowledge that an author wishes to convey is never explicitly stated in text (by one estimate the ratio of explicit:implicit knowledge is 1:8, [1]). Rather, the reader fills in the missing pieces using his/her background knowledge, creating a "mental model" of the scenario the text is describing, allowing him/her to go beyond facts explicitly stated. For example, given:

"A soldier was killed in the gun battle"

a reader would infer that, plausibly, the solder was shot, even though this fact is never explicitly stated.

A key requirement for this task is access to a large body of world knowledge. However, machines are currently poorly equipped in this regard, and developing such resources is challenging. Typically, manual acquisition of knowledge is too slow, while automatic acquisition is too messy. However, WordNet [2,3] presents one avenue for making inroads into this problem: It already has broad coverage, multiple lexico-semantic connections, and significant knowledge encoded (albeit informally) in its glosses; it can thus be viewed as on the path to becoming an extensively
leveragable resource for reasoning. Our goal is to explore this perspective, and to accelerate WordNet along this path. The result we are aiming for is a significantly enhanced WordNet better able to support applications needing extensive semantic knowledge.

2 Semantic Requirements on WordNet

To assess WordNet’s strengths and limitations for supporting textual question-answering, have been working with the task of “recognizing textual entailment” (RTE) [4,5], namely deciding whether a hypothesis sentence, H, follows from an initial text T. For example, from:

(1.T) Satomi Mitarai bled to death.

the following hypotheses plausibly follow:

(1.H1) Satomi Mitarai died.
(1.H2) Mitari lost blood.

Similarly, from:

(2.T) Hanssen, who sold FBI secrets to the Russians, could face the death penalty.

it plausibly follows that:

(2.H1) The FBI had secrets.
(2.H2) Hanssen received money from the Russians.
(2.H3) Hanssen might be executed.
(2.H4) The Russians bought secrets from Hanssen.

Our methodology has been to define a test suite of such sentences, analyze the types of knowledge required to determine if the entailment holds or not, and then determine the extent to which WordNet can provide this knowledge already and where the gaps are. For these gaps, we are exploring ways in which they can be partially filled in.

The test suite we developed contains 244 T-H entailment pairs (122 of which are positive entailments) such as those shown above. The pairs are grammatically fairly simple, and were deliberately authored to focus on the need for lexico-semantic knowledge rather than advanced linguistic processing. Determining entailment is very challenging in many cases. Each positive entailment pair was analyzed to identify the knowledge required to answer them. For example, for the pair:

(3.T) Iran purchased plans for a nuclear reactor from A.Q.Khan.
(3.H) The Iranians bought plans for building a nuclear reactor.
the computer needs to know:
"Iranian" is a person from Iran (derivation link)
"buy" and "purchase" are approximately equivalent (synonyms)
"plans for X" can mean "plans for building X" (world knowledge)

This process was repeated for all 122 positive entailments. From this, we found the knowledge requirements could be grouped into approximately 15 major categories, namely knowledge of:

1. Synonyms
2. Hypernyms
3. Irregular word forms
4. Proper nouns
5. Adverb-adjective relations
6. Noun-adjective relations
7. Noun-verb relations and their semantics (e.g., a consumer is the AGENT of consume event)
8. Purpose of artifacts
9. Polysemy vs. homonymy (related vs. unrelated senses of a word form)
10. Typical/plausible behavior (planes fly, bombs explode, etc.)
11. Core world knowledge (e.g., time, space, events)
12. Specific world knowledge (e.g., bleeding involves loss of blood)
13. Knowledge about actions and events (preconditions, effects)
14. Paraphrases (linguistically equivalent ways of saying the same thing)
15. Other

Of these, WordNet already has rich coverage of synonyms, hypernyms, adverb-adjective relations, and noun-adjective relations. It also has knowledge of noun-verb relations, although it does not distinguish between the different semantic type of this relation (e.g., AGENT, INSTRUMENT, EVENT); and some knowledge about semantic similarity highly polysemous verbs. In addition, WordNet has some knowledge of irregular word forms and proper nouns, and additional information is easily obtainable from other existing resources. The remaining knowledge types are still lacking; our goal is to extend WordNet to help provide more of this kind of knowledge. Note that we do not view WordNet as the sole supplier of knowledge, rather we wish to increase its utility as a contributing knowledge resource of systems performing advanced question-answering.

3 Recent WordNet Extensions

Based on this analysis, we are making several extensions to WordNet, which we describe in the following sections.
3.1 Morphosemantic links

WordNet contains mostly paradigmatic relations, i.e., relations among synsets with words belonging to the same part of speech (POS). Version 2 introduced cross-POS links, so-called "morphosemantic links" among synsets that were not only semantically but also morphologically related [6]. There are currently tens of thousands of manually encoded noun-verb (sense) connections, linking derivationally related nouns and verbs, e.g.,:

- abandon#v1 - abandonment#n3
- rule#v6 - ruler#n1
- catch#v4 - catcher#n1

Importantly, the appropriate senses of the nouns and verbs are paired, e.g., "ruler" and "rule" refer to the measuring stick and the marking or drawing with a ruler, respectively, rather than to a governor and governing, which makes for a different pair. What WordNet does not currently inform about, however, is the nature of the relation. For example:

- abandonment#n3 is the EVENT of abandon#v1
- ruler#n1 is the INSTRUMENT of rule#v6
- catcher#n1 is the AGENT of catch#v4

Knowledge of the nature of such relations is essential for many question-answering tasks. For example, given

(4.T) "Dodge produces ProHeart devices",

it is needed to realize that "producer" refers to the AGENT ("Dodge"), "production" refers to the EVENT ("produces"), and "product" the RESULT ("ProHeart devices"), a prerequisite for correctly answering questions asking about the producer/production/product.

The scale of adding this information manually is somewhat daunting; there are approximately 21,500 noun-verb (sense) links needing to be typed in WordNet. (We have not yet considered morphosemantic links among synsets from other parts of speech, which could also contribute to WordNet's usefulness as a tool for automated question answering.)

We have devised the following semi-automated approach:

1. We extract the noun-verb pairs with a particular morphological relation, (e.g., "-er" nouns such as "builder"-"build")
2. We determine the default relation for these pairs (e.g., The noun is the AGENT of the action expressed by verb)
3. We manually go through the list of pairs, marking pairs not conforming to default relation.
4. We inspect and group the marked pairs, assigning the correct relations to them.
This methodology is substantially faster than simply labelling each pair one by one, as only exceptions to the default relation need to be manually classified. In addition, this method has revealed the surprisingly high degree to which generally accepted one-to-one mappings of morphemes with meanings is violated. Furthermore, it is interesting to see that across the morphological classes, a limited inventory of semantic relations applies (for details see [7]).

3.2 Purpose links

A second type of knowledge often needed in question-answering is the function or purpose of artifacts (natural entities like stones and trees do not have an inherent function). For example, given:

(5.T) "The soldier was killed in a gun fight"
(5.H) "The soldier was shot"

we need to know that a gun is for shooting in order to infer that 5.H plausibly follows from 5.T. Knowledge what an artifact is intended for and how it is typically used enables a computer to make a plausible guess about implicit events that are not overtly expressed in a text. So our goal is to add links among noun and verb synsets in WordNet such that the verbs denote the intended and typical function or purpose of the nouns.

The number of such links is potentially huge, as almost any object can be used for almost any function. Thus, one can kill someone with a stiletto shoe, using is as a weapon. Similarly, a tree stump could be sat on when no chair is available. Worse, just about any solid object of a certain size can be used for hitting. We try to limit our links to those expressing the intended function, similar to the Role qualia of Pustejovsky [8]. Corpus data, e.g., [9], can be used to identify the most frequent noun-verb cooccurrences and usually confirm one's intuition about which noun-verb synset pairs should be linked.

Manually adding the links is a daunting task. However, a semi-automated approach is possible, using existing morphosemantic links in WordNet. As noted by Clark and Clark [10], English has a productive rule and fairly regular rule whereby many nouns can be used as verbs, and in many cases, the verb denotes the noun's intended function (or, put differently, the noun is the Instrument for carrying out the action expressed by the verb). Examples are "gun"(n)-"gun"(v): A gun is for gunning; "pencil"(n)-"pencil"(v): A pencil is for penciling, a hyponym of writing. In cases where there is no corresponding verb, e.g., for "car"(n), we can search up the hypernym tree until a more general noun is found which does have a corresponding verb, e.g., "car"(n) is a "transport"(n), linked to "transport"(v), thus a "car" is for "transporting".

We are currently inspecting the list of so-called zero-derived (homographic) noun-verb pairs in WordNet and classifying them as described in 3.1. Those pairs where the noun is an Instrument will be encoded with purpose links. Similarly, all noun-verb pairs from the different morphological classes (-er, -al, -ment, -ion, etc.) that were classified as expressing an Instrument relation can be labeled as "Purpose."
The automatic extraction of pairs related via a specific affix (Step 1 in 3.1 above) generates a list of candidate pairs that is validated and corrected by the same lexicographer who manually inspects the pairs for their semantic relation. Most pairs that are generated are valid, but a few false hits must be discarded. For example, the noun synset \{coax, ethernet cable\} was paired with the verb "coax", which would lead to the statement "An ethernet cable is for coaxing". In the majority of cases the computer's guess is sensible, and hence construction of the database is much faster than working from scratch.

3.3 World Knowledge - WordNet Glosses

WordNet contains a substantial amount of knowledge within its glosses. In particular, note that knowledge about a word (sense) is not just contained in that sense's gloss and example sentences, but also in its use in other glosses and example sentences. For example, for the word "lawn", WordNet includes mention that a lawn:

- needs watering;
- can have games played on it;
- can be flattened, mowed;
- can have chairs on it and other furniture;
- can be cut/mowed;
- can have things growing on it;
- has grass;
- can have leaves on it; and
- can be seeded.

Despite this promise, this knowledge is largely locked up in informal English text, and difficult to extract in a machine usable form (although there has been some work on translating the glosses to logic, e.g., \cite{11,12}. The glosses were not originally written with machine interpretation in mind, and as a result the output of machine interpretation is often syntactically valid but semantically meaningless logic. To address this challenge, we are proceeding along two fronts: first, we are developing an improved language processor specifically designed for interpreting the WordNet glosses; second, we are manually rephrasing some of the glosses to create more regularity in their structure, so that the resulting machine interpretation is improved.

To scope this work, we are focusing on "Core WordNet" Because WordNet contains tens of thousands of synsets referring to highly specific animals, plants, chemical compounds, etc. that are less relevant to NLP, the Princeton WordNet group has compiled a CoreWordNet, consisting of 5,000 synsets that express frequent and salient concepts. These were selected as follows. First, a list with the most frequent strings from the BNC was automatically compiled and all WordNet synsets for these strings were pulled out. Second, two raters determined which of the senses of these strings expressed "salient" concepts \cite{13}. The resulting top 5000 concepts comprises the core that we are focusing on, and as a result of this method of data collection contains a mixture of general and (common) domains-specific terms. (CoreWordNet is downloadable from http://wordnet.cs.princeton.edu /downloads.html)
3.4 World Knowledge - Core Theories

In addition to the specific world knowledge that might be obtained from the glosses, question-answering sometimes requires more fundamental, "core" knowledge of the world, e.g., about space, time, events, cognition, people and activities. Because of its more general nature, such knowledge is less likely to come from the WordNet glosses, and instead we are encoding some of this knowledge by hand as a set of "core theories". Although these theories contain only a small number of concepts (synsets), these concepts are also often general, meaning that information about them can be applied to a large number of other WordNet concepts. For example, WordNet has 517 "vehicle" nouns, and so any general knowledge about vehicles in general is potentially applicable to all these subtypes; similarly WordNet has 185 "cover" verbs, so general knowledge about the nature of covering can potentially apply to all these subtypes. In general, the broad coverage of WordNet can be funneled into a much smaller defined core, which can then be richly axiomatized, and the resulting axioms applied to much of the wider vocabulary in WordNet.

To identify these theories, we sorted words in Core WordNet into groups based on (a somewhat intuitive notion of) coherence, resulting in 15 core theories (listed with a selection of the words in them):

- Composite Entities: perfect, empty, relative, secondary, similar, odd, ...
- Scales: step, degree, level, intensify, high, major, considerable, ...
- Events: constraint, secure, generate, fix, power, development, ...
- Space: grade, inside, lot, top, list, direction, turn, enlarge, long, ...
- Time: year, day, summer, recent, old, early, present, then, often, ...
- Cognition: imagination, horror, rely, remind, matter, estimate, idea, ...
- Communication: journal, poetry, announcement, gesture, charter, ...
- Persons and their Activities: leisure, childhood, glance, cousin, jump, ...
- Microsocial: virtue, separate, friendly, married, company, name, ...
- Material World: smoke, shell, stick, carbon, blue, burn, dry, tough, ...
- Geo: storm, moon, pole, world, peak, site, village, sea, island, ...
- Artifacts: bell, button, van, shelf, machine, film, floor, glass, chair, ...
- Food: cheese, potato, milk, break, cake, meat, beer, bake, spoil, ...
- Macrosocial: architecture, airport, headquarters, prosecution, ...
- Economic: import, money, policy, poverty, profit, venture, owe, ...

We are first focusing on Time and Event words. We have developed underlying ontologies of time and event concepts, explicating the key notions in these domains [14,15]. For example, the temporal ontology axiomatizes topological temporal concepts like before, duration concepts, and concepts involving the clock and calendar. The event ontology axiomatizes notions like subevent, and the internal structure of events and processes. We are then defining, or at least characterizing, the meanings of the various word senses in terms of these underlying theories. For example, to fix something is to bring about a state in which all the components of the thing are functional. This effort is of course a very labor intensive project, but since
we are concentrating on the synsets in the core WordNet, we believe we will achieve the maximum impact for the labor we put into it.

Because of the richness of WordNet's hypernym links, in principle these axioms can be heavily reused for reasoning about WordNet word senses. A number of the textual entailment problems in our test suite appeal directly to this knowledge, for example to judge the validity of this entailment:

(6.T) Baghdad has seen a spike in violence since the summer.
(6.H) There was greater violence in Baghdad since the summer.

requires reasoning about the core notion of change in a quantity ("spike", "rise"), rather than anything specific about Baghdad, violence, or summer. This kind of knowledge - namely the meaning of these core words and their relationships - is being encoded in these core theories.

4. Status and Summary

The work that we have described here is still a work in progress: To date, we have corrected/validated about half of the machine-generated database of morphosemantic links; made an initial start on the purpose links; have completed a first pass on logical forms for WordNet glosses and are focussing on improving both the phrasing and interpretation of Core WordNet; and have completed some of the core theories and are in the process of linking their core notions to WordNet word senses. Our goal is that these extensions will substantially improve WordNet's utility for language-based problems that require reasoning as well as basic lexical information, and we are optimistic that these will improve WordNet's ability to meet the increasingly strong requirements demanded by modern day language-based applications.

Acknowledgements

This work was supported by the AQUAINT Program of the Disruptive Technology Office under contract number N61339-06-C-0160.

References

An Evaluation Procedure for Word Net Based Lexical Chaining: Methods and Issues

Irene Cramer and Marc Finthammer

Faculty of Cultural Studies, University of Dortmund, Germany
irene.cramer|marc.finthammer@uni-dortmund.de

Abstract. Lexical chaining is regarded to be a valuable resource for NLP applications, such as automatic text summarization or topic detection. Typically, lexical chainers use a word net to compute semantically motivated partial text representations. However, their output is normally evaluated with respect to an application since generic evaluation criteria have not yet been determined and systematically applied. This paper presents a new evaluation procedure meant to address this issue and provide insight into the chaining process. Furthermore, the paper exemplarily demonstrates its application for a lexical chainer using GermaNet as a resource.

1 Project Context and Motivation

Converting linear text documents into documents publishable in a hypertext environment is a complex task requiring methods for the segmentation, reorganization, and linking. The HyTex project, funded by the DFG, aims at the development of conversion strategies based on text-grammatical features. One focus of our work is on topic-based linking strategies using lexical and thematic chains. In contrast to the lexical ones thematic chains are based on a selection of central words, so called topic anchors, which are e.g. words able to outline the content of a complete passage, and as in lexical chaining connected via semantically meaningful edges. An illustration is given in Fig. 1.

We intend to use lexical chaining for the construction of thematic chains: on the one hand as a feature for the extraction of topic anchors and on the other hand as a tool for the calculation of thematic structure, as shown in Fig. 1. For this purpose, we implemented a lexical chainer for German corpora based on GermaNet. In order to perform an in-depth analysis and evaluation of this chainer as well as to gain insight into the whole chaining process we developed a detailed evaluation procedure. We argue that this procedure is applicable to any lexical chainer regardless of the algorithm or resources used and helps to fine-tune the parameter setting ideal for a specific application. We also present a detailed evaluation of our own lexical chainer and illustrate the issues and challenges we encountered using GermaNet as a resource.

1 See our project web pages http://www.hytex.info/ for more information about the concept of thematic chains and the project context.
Paper plan: The remainder of this paper is structured as follows: Section 2 describes the basic aspects of lexical chaining and presents a detailed, new evaluation procedure. Section 3 presents the resources used for our lexical chainer and the evaluation. Section 4 discusses our preprocessing component necessary to handle the rather complex German morphology and well-known challenges, such as proper names, in lexical chaining. Section 5 discusses our chaining based disambiguation experiments. Section 6 presents a short overview of eight semantic relatedness measures and compares their values with the results of a human judgment experiment that we conducted. Section 7 outlines the evaluation of our chaining with respect to our application scenario and the project context. Section 8 summarizes and concludes the paper.

2 Lexical Chaining

Based on the concept of lexical cohesion [1] computational linguists e.g. [2] developed a method to compute partial text representations: lexical chains. To illustrate the idea an annotation is given as an example in Fig. 2. It shows that lexical chaining is achieved by the selection of vocabulary and significantly accounts for the cohesive structure of a text passage. The chains span over passages linking lexical items, where the linking is based on the semantic relations existing between them. Typical semantic relations considered in this context are synonymy, antonymy, hyponymy, hypernymy, meronymy and holonymy as well as complex combinations of these which are computed on the basis of lexical semantic resources such as WordNet [3]. In addition to WordNet, which has been used in the majority of cases e.g. [4], [5], [6], Roget’s Thesaurus [2] and GermaNet [7] have already been applied.
Jan sat down to rest at the foot of a huge beech-tree. Now he was so tired that he soon fell asleep; and a leaf fell on him, and then another, and then another, and before long he was covered all over with leaves, yellow, golden and brown.

Chain 1: sat down, rest, tired, fell asleep
Chain 2: beech-tree, leaf, leaves

Unsystematic relations not yet considered in resource for lexical chaining: foot / huge – beech-tree; yellow / golden / brown – leaves

Several natural language applications as text summarization e.g. [8], [9], malapropism recognition [4], automatic hyperlink generation e.g. [5], question answering e.g. [10] and topic detection/topic tracking e.g. [11] benefit from lexical chains as a valuable text representation.

In this paper we present the evaluation of our own implementation of a lexical chainer for German, GLexi, which is based on the algorithms described by [4] and [8] and was developed to support the extraction of thematic structures and topic development. As most systems, GLexi consists of the fundamental modules shown in Table 1, which reveals that preprocessing – thus, the selection of the so-called chaining candidates and determination of relevant information about these candidates, like text position and part-of-speech – play a major role in the whole process. A chaining candidate is the fundamental chain element; it is a token comprised of all bits of information belonging to it.

We argue that a sophisticated preprocessing may enhance coverage, which is acknowledged to be a crucial aspect in the development of a lexical chaining system e.g. [5], [8], and [4]. Accordingly, we address several ideas to improve the coverage of our system. At least two issues independent of language influence this aspect:

- limitations imposed on the whole process by the size and coverage of the lexical semantic resource used,
- and the presence of proper names in the text, which cannot be resolved without extensive preprocessing.
Table 1. Overview of chainer modules

<table>
<thead>
<tr>
<th>Module</th>
<th>Subtasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>preprocessing of corpora</td>
<td>chaining candidate selection:</td>
</tr>
<tr>
<td></td>
<td>determine chaining window,</td>
</tr>
<tr>
<td></td>
<td>sentence boundaries,</td>
</tr>
<tr>
<td></td>
<td>tokens, POS-tagging,</td>
</tr>
<tr>
<td></td>
<td>chunks etc.</td>
</tr>
<tr>
<td>core chaining algorithm</td>
<td>lexical semantic look-up resource (e.g. WordNet),</td>
</tr>
<tr>
<td>calculation of chains</td>
<td>scoring of relations,</td>
</tr>
<tr>
<td>or meta-chains</td>
<td>sense disambiguation</td>
</tr>
<tr>
<td>output creation</td>
<td>rating/scoring of chain strength</td>
</tr>
<tr>
<td></td>
<td>build application specific representation</td>
</tr>
</tbody>
</table>

However, it is even more critical for German coverage because
– of its complex morphology (e.g. inflection and word formation)
– and the smaller coverage of GermaNet in comparison to WordNet.

Both aspects as well as coverage in general are discussed in detail in the following sections.

In order to formally evaluate the performance – in terms of precision and recall – of GLexi for various parameter settings a (preferably standardized and freely available) test set would be required. To our knowledge there is no such resource – neither for English nor for German. Therefore, we have started to investigate the development of such a gold standard for German corpora. Initial results are discussed in [12]. Our experiments show that the manual annotation of lexical chains is a demanding task, which has also been emphasized in the work by [13], [14] and [15]. The rich interaction between various principles to achieve a cohesive text structure seems to distract annotators. We therefore argue that the evaluation of a lexical chainer might be best performed in four steps:

– **evaluation of coverage**: amount of chaining candidates the chainer is able to process,
– **evaluation of disambiguation quality**: number of chaining candidates correctly disambiguated with respect to lexical semantic resource,
– **evaluation of quality of semantic relatedness measures**: comparison with human judgment,
– **evaluation of chains** with respect to concrete application.

This procedure ensures that the most relevant parameters in the evaluation of our system, GLexi, can be judged separately and also enables us to gain the necessary insight into the chaining process.
We based the evaluation of our system and all experiments described in this paper on three main resources: GermaNet as the lexical semantic lexicon for our chainer, the HyTex project corpus and a set of word pairs compiled in a human judgment experiment for the evaluation steps discussed in Sect. 6.2.

3.1 GermaNet

GermaNet [16] is a machine readable lexical semantic lexicon for the German language developed in 1996 within the LSD Project at the Division of Computational Linguistics of the Linguistics Department at the University of Tübingen. Version 5.0 covers approximately 77,000 lexical units – nouns, verbs, adjectives and adverbs as well as some multi word units – grouped into approximately 53,500 so-called synonym sets. GermaNet contains approximately 4,000 lexical (between lexical units) and approximately 64,000 conceptual (between synonym sets) connections. Although it has much in common with the English WordNet [3] there are some differences; see [17] for more information about this issue. The most important difference in our opinion is the fact that GermaNet is much smaller than WordNet, which has a negative impact on the coverage. However, we found that none of the other differences, such as the presence of artificial concepts, have much influence over the results of our chainer.

3.2 Corpus

For the evaluation steps mentioned in Sect. 2 we used a part of the HyTex corpus, which contains 130 documents (approximately 3 million words). It was compiled and in parts manually annotated in project phase I; see [18] for more information. The HyTex corpus consists of 3 subcorpora: the so-called core corpus, supplementary corpus and statistics corpus. The corpora contain scientific papers, technical specifications, tutorials and textbook chapters, as well as FAQs about language technology and hypertext research.

In the core corpus logical text structure is marked, for example the organization of documents into chapters, sections, passages, figures, footnotes, tables etc. is annotated using DocBook-based XML tags; see [19] for more information. In order to split the documents into chainable sections, we used the core corpus and segmented the documents according to its annotation. The homogeneity and relevance of a chain largely depends on its length and thus on the length of the underlying text. We found the average length of a section to be adequate for chaining of our domain-specific corpus. We also decided to only select nouns and noun phrases as chaining candidates because our experiments revealed that terminology plays the key role in scientific and technical documents terminology.
3.3 Set of Word Pairs

In order to evaluate the quality of a relatedness measure, a set of pre-classified word pairs (in our case for German) is necessary. In previous work for English, most researchers used Rubenstein and Goodenough’s list [20] or Miller and Charles’s list [21]. For German there are – to our knowledge – three sets of word pairs: a translation of Rubenstein and Goodenough’s list by [22], a manually generated set of 350 word pairs by [23], and a semi-automatically generated set by [24]. Unfortunately, we could not find any of these German sets published. We also argue that the translation of a list constructed originally for English subjects might bias the results and therefore decided to compile our own set of word pairs as can be seen in Table 2. The goal was to cover a wide range of relatedness types, i.e. systematic and unsystematic relations, and relatedness levels, i.e. various degrees of relation strength. We also included nouns of diverse semantic classes, e.g. abstract nouns, such as das Wissen (Engl. knowledge), and concrete nouns, such as das Bügeleisen (Engl. flat-iron). We thus constructed a list of approximately 320 word pairs, picked 100 of these to evenly meet the constraints mentioned above and randomized them. We also included words which occur more than once (up to 8 times) in a word pair; these are grouped into consecutive blocks. We asked 35 subjects to rate the word pairs on a 5-level scale (0 = not related to 4 = strongly related). The subjects were instructed to base the rating on their intuition about any kind of conceivable relation between the two words. We used this list and the human judgment to evaluate the semantic relatedness measures described in Sect. 6.1.

4 Evaluation Phase I – Preprocessing Methods

We conducted several experiments to investigate the coverage of GermaNet and thus the coverage of GLexi. We found that GermaNet contains 56.42% of the 28,772 noun tokens mentioned in the corpus. We concluded from a sample analyzed that this coverage issue stems from the rich German morphology, domain-specific terminology and proper names, which are both not covered sufficiently by GermaNet. We therefore implemented the preprocessing architecture shown in Fig. 3. A document is first segmented into sections and then split into sentences and tokens. In addition, for each token a list of features is extracted, such as position in the document (with respect to sentence and section), part-of-speech, lemma, and morphology. On this basis the preprocessing component generates one or several alternative chaining candidates, e.g. the first alternative would be the singular instead of a plural, like for cats ⇒ cat. The second alternative considers compounds when applicable. Since our corpus is very rich in compounds this plays a major role in the implementation of our system and is discussed in more detail in Sect. 4.1. Technical terminology and proper names are also considered separately as alternatives.

For our study we used the Insight Discoverer Extractor Version 2.1. (cf. http://www.temis-group.com/). We thank the TEMIS group for kindly permitting us to use this technology in the framework of our project.
Table 2. Word pairs and human judgment mean value

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>Mean Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nahrungsmittel</td>
<td>Essen</td>
<td>3.94</td>
</tr>
<tr>
<td>Wasser</td>
<td>Flüssigkeit</td>
<td>3.89</td>
</tr>
<tr>
<td>Teller</td>
<td>Kind</td>
<td>3.86</td>
</tr>
<tr>
<td>Blume</td>
<td>Pflanze</td>
<td>3.86</td>
</tr>
<tr>
<td>Angst</td>
<td>Freude</td>
<td>3.86</td>
</tr>
<tr>
<td>Räume</td>
<td>Schornstein</td>
<td>3.80</td>
</tr>
<tr>
<td>Blume</td>
<td>Tulpe</td>
<td>3.80</td>
</tr>
<tr>
<td>Sonne</td>
<td>Strom</td>
<td>3.71</td>
</tr>
<tr>
<td>Nahrungsmittel</td>
<td>Tier</td>
<td>3.74</td>
</tr>
<tr>
<td>Wasser</td>
<td>Frisch</td>
<td>3.89</td>
</tr>
<tr>
<td>Mensch</td>
<td>Lebewesen</td>
<td>3.66</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Bademeister</td>
<td>3.63</td>
</tr>
<tr>
<td>Kiste</td>
<td>Gepäck</td>
<td>3.63</td>
</tr>
<tr>
<td>Mitarbeiter</td>
<td>Kollege</td>
<td>3.60</td>
</tr>
<tr>
<td>Behandlung</td>
<td>Therapie</td>
<td>3.54</td>
</tr>
<tr>
<td>Kampe</td>
<td>Leuchte</td>
<td>3.49</td>
</tr>
<tr>
<td>Sonde</td>
<td>Expedition</td>
<td>3.39</td>
</tr>
<tr>
<td>Ocean</td>
<td>Reise</td>
<td>3.46</td>
</tr>
<tr>
<td>Wahl</td>
<td>Dekorations</td>
<td>3.47</td>
</tr>
<tr>
<td>Schießkappe</td>
<td>Schwimmer</td>
<td>3.40</td>
</tr>
<tr>
<td>Würfel</td>
<td>Zufall</td>
<td>3.37</td>
</tr>
<tr>
<td>Wissen</td>
<td>Kenntnis</td>
<td>3.38</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Hecken</td>
<td>3.31</td>
</tr>
<tr>
<td>Würfel</td>
<td>Spiel</td>
<td>3.31</td>
</tr>
<tr>
<td>Nahrungsmittel</td>
<td>Hunger</td>
<td>3.31</td>
</tr>
<tr>
<td>Bewegung</td>
<td>Tanz</td>
<td>3.26</td>
</tr>
<tr>
<td>Kalte</td>
<td>Wärme</td>
<td>3.20</td>
</tr>
<tr>
<td>Mensch</td>
<td>Verstand</td>
<td>3.20</td>
</tr>
<tr>
<td>Nahrungsmittel</td>
<td>Restaurant</td>
<td>3.20</td>
</tr>
<tr>
<td>Wissen</td>
<td>Schule</td>
<td>3.17</td>
</tr>
<tr>
<td>Zufriedenheit</td>
<td>Freundlichkeit</td>
<td>3.17</td>
</tr>
<tr>
<td>Wissen</td>
<td>Quiz</td>
<td>3.09</td>
</tr>
<tr>
<td>Blume</td>
<td>Wasser</td>
<td>3.09</td>
</tr>
<tr>
<td>Heiß</td>
<td>Wärme</td>
<td>3.03</td>
</tr>
<tr>
<td>Kontinent</td>
<td>Landkarte</td>
<td>3.03</td>
</tr>
<tr>
<td>Sonne</td>
<td>Leben</td>
<td>3.00</td>
</tr>
<tr>
<td>Wissen</td>
<td>Intelligenz</td>
<td>3.00</td>
</tr>
<tr>
<td>Marschen</td>
<td>Geschäft</td>
<td>2.94</td>
</tr>
<tr>
<td>Sonne</td>
<td>Stern</td>
<td>2.91</td>
</tr>
<tr>
<td>Unterhaltung</td>
<td>Programm</td>
<td>2.91</td>
</tr>
<tr>
<td>Staat</td>
<td>Wohnung</td>
<td>2.94</td>
</tr>
<tr>
<td>Wasser</td>
<td>Feucht</td>
<td>2.90</td>
</tr>
<tr>
<td>Treppen</td>
<td>Aufzug</td>
<td>2.77</td>
</tr>
<tr>
<td>Heiß</td>
<td>Feucht</td>
<td>2.74</td>
</tr>
<tr>
<td>Blume</td>
<td>Lese</td>
<td>2.71</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Hundef</td>
<td>2.60</td>
</tr>
<tr>
<td>Mensch</td>
<td>Krankheit</td>
<td>2.60</td>
</tr>
<tr>
<td>Tür</td>
<td>Fenster</td>
<td>2.54</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>Mean Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sonne</td>
<td>Strom</td>
<td>2.51</td>
</tr>
<tr>
<td>Wasser</td>
<td>Nebel</td>
<td>2.43</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Ferien</td>
<td>2.40</td>
</tr>
<tr>
<td>Nahrungsmittel</td>
<td>Tier</td>
<td>2.34</td>
</tr>
<tr>
<td>Wasser</td>
<td>Fett</td>
<td>2.31</td>
</tr>
<tr>
<td>Würfel</td>
<td>Mathematik</td>
<td>2.23</td>
</tr>
<tr>
<td>Mensch</td>
<td>Hund</td>
<td>1.91</td>
</tr>
<tr>
<td>Wasser</td>
<td>Fett</td>
<td>1.89</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Ausflug</td>
<td>1.77</td>
</tr>
<tr>
<td>Würfel</td>
<td>Feucht</td>
<td>1.57</td>
</tr>
<tr>
<td>Würfel</td>
<td>Kugel</td>
<td>1.49</td>
</tr>
<tr>
<td>Wasser</td>
<td>Quelle</td>
<td>1.34</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Feucht</td>
<td>1.43</td>
</tr>
<tr>
<td>Würfel</td>
<td>Hof</td>
<td>1.30</td>
</tr>
<tr>
<td>Würfel</td>
<td>Hof</td>
<td>1.23</td>
</tr>
<tr>
<td>Mensch</td>
<td>Welt</td>
<td>1.26</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Eis</td>
<td>1.18</td>
</tr>
<tr>
<td>Wissen</td>
<td>Kenntnis</td>
<td>1.11</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Zahn</td>
<td>1.11</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Stadt</td>
<td>1.09</td>
</tr>
<tr>
<td>Wasser</td>
<td>Vergangen</td>
<td>1.03</td>
</tr>
<tr>
<td>Wasser</td>
<td>Lagerstrecke</td>
<td>1.03</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Hecken</td>
<td>1.17</td>
</tr>
<tr>
<td>Wissen</td>
<td>Ablage</td>
<td>1.11</td>
</tr>
<tr>
<td>Würfel</td>
<td>Honig</td>
<td>0.97</td>
</tr>
<tr>
<td>Pflanze</td>
<td>Klebe</td>
<td>0.54</td>
</tr>
<tr>
<td>Datum</td>
<td>Auskostung</td>
<td>0.54</td>
</tr>
<tr>
<td>Sonne</td>
<td>Arzt</td>
<td>0.31</td>
</tr>
<tr>
<td>Catambe</td>
<td>Restaurant</td>
<td>0.29</td>
</tr>
<tr>
<td>Mensch</td>
<td>Wolke</td>
<td>0.20</td>
</tr>
<tr>
<td>Sonne</td>
<td>Direktor</td>
<td>0.17</td>
</tr>
<tr>
<td>Nation</td>
<td>Hecken</td>
<td>0.17</td>
</tr>
<tr>
<td>Mittagessen</td>
<td>Hecken</td>
<td>0.17</td>
</tr>
<tr>
<td>Farbe</td>
<td>Richter</td>
<td>0.14</td>
</tr>
<tr>
<td>Vok</td>
<td>Punk</td>
<td>0.13</td>
</tr>
<tr>
<td>Rechte</td>
<td>Geld</td>
<td>0.11</td>
</tr>
<tr>
<td>Schleuder</td>
<td>Schallplatte</td>
<td>0.09</td>
</tr>
<tr>
<td>Löffel</td>
<td>Baum</td>
<td>0.09</td>
</tr>
<tr>
<td>Nahrungsmittel</td>
<td>Kabel</td>
<td>0.09</td>
</tr>
<tr>
<td>Hitze</td>
<td>Familie</td>
<td>0.09</td>
</tr>
<tr>
<td>Wasser</td>
<td>Rundfunk</td>
<td>0.09</td>
</tr>
<tr>
<td>Maus</td>
<td>Monat</td>
<td>0.09</td>
</tr>
<tr>
<td>Tasse</td>
<td>Motor</td>
<td>0.03</td>
</tr>
<tr>
<td>Bach</td>
<td>Wal</td>
<td>0.03</td>
</tr>
<tr>
<td>Schwimmbad</td>
<td>Kabel</td>
<td>0.03</td>
</tr>
<tr>
<td>Gardine</td>
<td>Bleichmittel</td>
<td>0.03</td>
</tr>
<tr>
<td>Oase</td>
<td>Hängeleisten</td>
<td>0.03</td>
</tr>
<tr>
<td>Wäscheleine</td>
<td>Toastbrötchen</td>
<td>0.03</td>
</tr>
<tr>
<td>Würfel</td>
<td>Wässer</td>
<td>0.02</td>
</tr>
<tr>
<td>Fluss</td>
<td>Drucker</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Table 3. Coverage of GermaNet

<table>
<thead>
<tr>
<th>Coverage of GermaNet</th>
<th>Coverage of GermaNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>The approximately 29,000 (noun) tokens in our corpus split into</td>
<td>56% in GermaNet</td>
</tr>
<tr>
<td></td>
<td>44% not in GermaNet, of these:</td>
</tr>
<tr>
<td></td>
<td>15% inflected</td>
</tr>
<tr>
<td></td>
<td>12% compounds</td>
</tr>
<tr>
<td></td>
<td>17% small, uncovered classes</td>
</tr>
</tbody>
</table>

Fig. 3. Preprocessing architecture
4.1 German Morphology

Compared to English, the German noun morphology is relatively complex: especially the presence of four cases and compounds, which are written as one word and not divided by blanks, plays a major role in our chaining system.

**Notes on German inflection:** In order to ensure that inflected nouns can be handled accurately we rely on lemmatization. Inflection in German means four cases and singular/plural forms.

**Coverage improvement on the basis of inflection processing:** On the basis of our lemmatization step, we were able to replace approximately 15% of the nouns by their lemmata and could thus increase the coverage to 71%.

**Open Issues:** However, we found that there are some cases in which the original (plural) form in the text should not be normalized to its singular form, e.g. the German word Daten (Engl. data or dates) can be lemmatized to Datum (Engl. date); the same holds for Medien (Engl. media) and Medium (Engl. psychic, data carrier). Thus, when lemmatized the words change their meaning. Moreover, the plural form is not included in GermaNet. Consequently, our system uses as a chaining element the first alternative of the original, e.g. Datum instead of Daten. Of course, in our domain specific corpus Daten (Engl. data) and Medien (Engl. media) are frequent words (Daten occurred 78 times in the corpus, Medien 41 times), which serve in the chains as glue for a list of other chaining elements and therefore need to be carefully considered. In addition, lemmatization is not very reliable for compounds. Nevertheless, we think that the results mentioned above emphasize that this preprocessing step is a necessary aspect to improve the coverage of a baseline chaining system.

**Notes on German compounds:** Compounds are frequent in our limited domain corpus. Two or more (free) morphemes are combined into one word, the compound, e.g. Druckerpatrone (components: Drucker and Patrone; Engl. ink cartridge). Sometimes, the components are additionally joined by a so-called Fugenelement (Engl. gap element), e.g. Liebeslied (components: Liebe and Lied, gap element: s; Engl. love song). Typically, the complete compound inherits the grammatical features, such as genus, of its last – so-called head – component, thus the one at the rightmost position, e.g. das Lied (genus: neutral; Engl. song) and das Liebeslied (genus: neutral), while it is die Liebe (genus: feminine; Engl. love). In addition to these grammatical features of compounds in German there are at least two semantically motivated classes: the semantically transparent and the intransparent compounds. Semantically transparent describes a compound for which the meaning of the whole can be deduced from the meaning of its parts, e.g. a Liebeslied (Engl. love song) is a kind of Lied (this component is the head of the compound; Engl. song), where the component Liebe (Engl. love) can be seen as the modifier of the head. In contrast, the meaning of a semantically intransparent compound cannot be deduced from its parts, e.g. Rotkehlchen (Engl. robbin; components: rot, Engl. red, and Kehlchen, which can be split into Kehle, Engl. throat and -chen diminutive suffix). An ideal lexical semantic resource would cover all intransparent compounds, whereas the transparent ones would not necessarily be included since it is possible to derive their meaning intel-
lectually or automatically. In principle GermaNet accounts for this rule, however, there are as always some compounds which are not included.

Coverage improvement on the basis of compound processing: On the basis of the morphological analysis we were able to include previously uncovered words, i.e. approximately 12% of the nouns could be replaced by their compound head word (e.g. Liebeslied would be replaced with Lied) and thus increase the coverage to 83%.

Open Issues: However, this step has at least two major drawbacks. First, the morphological analysis generated by the Insight Discoverer™ Extractor Version 2.1 contains all possible readings, e.g. the German word Agrarproduktion (Engl. agricultural production) might be split among other things into Agrar (Engl. agricultural), Produkt (Engl. artifact) and Ion (Engl. ion [chem.]). The automatic selection of a correct reading is in some cases demanding and the effect on the whole chaining process might be severe – e.g. given the word Produktion and the morphological analysis mentioned the chainer could decide to replace the word Produktion, given it cannot be found in GermaNet, with the word Ion, which could completely mislead the disambiguation of word sense in the chaining and thus the whole chaining process itself. Second, compounds containing more than two components could be split into several head-words, e.g. the head-word of the compound Datenbankbenutzerschnittstelle (Engl. data base user interface) could be Benutzerschnittstelle (Engl. user interface) or Schnittstelle (Engl. interface) or even only Stelle (Engl. position or area³). In our future work, we therefore plan to investigate which parameter settings might be ideal on the one hand to improve the coverage and on the other hand to account for semantic disambiguation performance. Nevertheless, we think that morphological analysis of compounds is a crucial aspect in the preprocessing of our lexical chainer.

4.2 Smaller Classes of Uncovered Material

As Table 3 shows, with our first preprocessing step we were able to include approximately 27% of the words, which we could initially not find in GermaNet, i.e. approximately 15% on the basis of lemmatization and approximately 12% on the basis of compound analysis. We examined a sample of the remaining 17%, the results are shown in Table 4. We found in the sample approximately 15% proper names, approximately 30% foreign words, especially technical terminology in English, approximately 25% abbreviations, and approximately 20% nominalized verbs, which are not sufficiently included in GermaNet and very prominent in German technical documents. The rest (not shown in Table 4) consists of incorrectly tokenized or POS-tagged material, such as broken web links.

No matter which language is considered, proper names are a well-known challenge in lexical chaining, e.g. [5]. They are semantically central items in most corpora and therefore need to be handled with care. The same holds for technical terminology, in

³ Note: This is the correct though in this context semantically inadequate translation.
Table 4. Detailed analysis of small classes not covered by GermaNet

| The small, uncovered classes (see Table 2) split into                     |
|-----------------------------|-----------------------------|
| 15% proper names            | 30% foreign words           | 25% abbreviations            | 20% nominalized verbs |

many cases multi-word units, which are obviously very frequent and relevant in technical and academic documents. We deal with both in the second phase of our preprocessing component. However, note that we only treat the classical named entities, i.e. names belonging to people, locations, and organizations. We do not yet cover other proper names.

We included the recognition of proper names and multi-word units in our preprocessing. After the basic preprocessing, such as sentence boundary detection, tokenization and lemmatization, which is accomplished by the Insight Discoverer\textsuperscript{TM} Extractor Version 2.1, we run the second preprocessing phase, which splits into the following two subtasks:

- Proper name recognition and classification: We use a simple named entity recognizer (NER) for German\textsuperscript{4}, which tags person names, locations, and organizations.
- Simple chunking of multi word units and simple phrases: We use the part-of-speech tags computed in the first preprocessing step by the Insight Discoverer\textsuperscript{TM} Extractor Version 2.1 to construct simple phrases.

Of course, these are interim solutions, and we plan to investigate strategies to improve the second preprocessing phase in our future work. Because we found names of conferences and product names to be relatively frequent, we intend to extend our NER system accordingly. Most of the technical terminology in our corpus is not included in GermaNet and could thus not be considered in the chaining. However, in the HyTex project we developed a terminological lexicon for our corpus (called TermNet), see [25] and [26], which we plan to use in addition to GermaNet. Ultimately, we hope this will again improve the coverage of our chainer. While it is thus far unclear how to handle nominalized verbs and abbreviations, the statistics shown in Table 4 emphasize their relevance, and they certainly need to be considered with care in our future work.

To conclude, without any preprocessing only 56% of the noun tokens in our corpus are chainable. Approximately 67% of the remaining nouns can be handled with morphological analysis and a very simple NER system. The remaining approximately 33% is comprised of abbreviations, foreign words, nominalized verbs and broken material as well as not yet covered proper names and technical terminology, which we intend to deal with in an expansion of our lexical semantic resource, i.e. in a combination of GermaNet and TermNet, statistical relatedness measures based on web counts and a refinement of our preprocessing components.

\textsuperscript{4} It is our own machine learning based implementation of a simple NER system.
5 Evaluation Phase II – Chaining-based Word Sense Disambiguation

In addition to the coverage issues described in Sect. 4 word sense disambiguation has a high impact on the performance of a lexical chainer. That is, if incorrectly disambiguated, a word with several word senses, such as *bank* or *mouse*, could mislead the complete chaining algorithm and cause the construction of inappropriate chains. As a matter of course, the disambiguation performance of a chainer is not able to outperform high-quality WSD systems, such as presented at the Senseval workshops, and it is not our purpose to compete against these systems but to locate potential sources of error in the chaining procedure. Consequently, the second step in our evaluation procedure is related to word sense disambiguation, in our case the selection of an appropriate synonym set in GermaNet. In principle, there are at least two different methods: the greedy selection of a word sense and the subsequent selection. Greedy word sense disambiguation means to choose the first matching synonym set which exhibits a suitable path or a semantic relatedness measure value. In contrast, subsequent disambiguation, see e.g. [9], means to first assemble all possible readings, i.e. all in principle suitable paths or semantic relatedness measure values, and then, given this information, select the best match. However, both methods have their pros and cons: the greedy selection is simple and straightforward, but it tends to pick the wrong word sense in cases in which the correct reading of a word cannot be determined until the rest of the potential chaining partners are examined. In addition to these two methods, there are several ranges between the greedy and the subsequent disambiguation: e.g. the appropriate synonym set of a word might be determined on the basis of a majority vote when all possible combinations containing this word are read. Alternatively, the information content (see Sect. 6.1) might be useful to pick a word sense.

Analysis of the chaining-based word sense disambiguation: In lexical chaining, the disambiguation is essentially based on the selection of a word sense with respect to a path or relatedness measure value between synonym sets. For example, a pair of words *A*, with three senses, and *B*, with two senses, has six possible readings: thus, the probability to pick the correct one is only 1/6. The more senses a word pair exhibits, the likelier it is to pick an incorrect reading for at least one of the two words. Table 5 shows the distribution of word senses for the (noun) tokens in our corpus. Obviously, almost every second token features more than one word sense in GermaNet. That means in the worst case every second token can in principle mislead the chainer in the case of an incorrect disambiguation.

---

5 We consider tokens instead of types because in principle every single occurrence of a word might exhibit a different word sense. We have such examples in our corpus, e.g. in one sentence the word *text* is used with three different senses.
Table 5. Overview of the number of word senses occurring in our corpus

<table>
<thead>
<tr>
<th>1 sense</th>
<th>2 senses</th>
<th>3 senses</th>
<th>4 senses</th>
<th>&gt; 4 senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>~ 53%</td>
<td>~ 22%</td>
<td>~ 15%</td>
<td>~ 7%</td>
<td>~ 3%</td>
</tr>
</tbody>
</table>

Fig. 4. Example ranking of the various readings

However, it is the basic idea of lexical chaining that lexicalized coherence in the text accounts for the mutually correct disambiguation of the words in a pair. In order to investigate the disambiguation quality, we randomly selected a corpus sample and computed the relatedness values. We then ranked the possible readings for each word pair according to their relatedness values. An example is shown in Fig. 4. We evaluated this against our manual annotation of word senses. The results are shown in Table 6. The three best relatedness measures in this context, Resnik, Wu-Palmer and Lin, correctly disambiguate approximately 50% of the word pairs in our sample. For all eight measures the correct reading is on the first four ranks in the majority of the cases. Although this disambiguation accuracy is only mediocre, it outperforms the baseline (approximately 39% correct disambiguation on rank 1), i.e. the performance of a chainer using the information content of a word to disambiguate its word sense. As mentioned above an additional alternative method to select the correct word sense is the majority voting: for a list of word pairs with one given word and all possible chaining partners in the text (e.g. mouse - computer, mouse - hardware, mouse - keyboard, mouse - etc.), the word sense, which is supported by most of the top-ranked relatedness measure values, is supposed to be the correct one. Our experiments showed that a majority voting is able to enhance the accuracy and bring the rate in some cases up to 63% correct disambiguation. We plan to investigate in our future work how we can again improve the disambiguation quality of our chainer. We especially plan to explore the method of meta-chaining proposed in [9] and to adapt it for a multiple relatedness measure chain-
ing framework. In addition, the integration of a WSD system might positively influence the performance of our chainer.

Table 6. Overview of semantic relatedness-based disambiguation performance

<table>
<thead>
<tr>
<th></th>
<th>correct disamb. on</th>
<th>Graph Path</th>
<th>Tree Path</th>
<th>Wu-Palmer</th>
<th>Leacock-Chodorow</th>
</tr>
</thead>
<tbody>
<tr>
<td>rank 1</td>
<td>34.93%</td>
<td>42.13%</td>
<td>50.67%</td>
<td>34.93%</td>
<td></td>
</tr>
<tr>
<td>rank 1 – 4</td>
<td>79.20%</td>
<td>80.80%</td>
<td>86.40%</td>
<td>79.20%</td>
<td></td>
</tr>
<tr>
<td>Hirst-StOnge</td>
<td>17.07%</td>
<td>57.60%</td>
<td>37.60%</td>
<td>50.13%</td>
<td></td>
</tr>
<tr>
<td>Resnik</td>
<td>19.20%</td>
<td>88.80%</td>
<td>77.87%</td>
<td>87.20%</td>
<td></td>
</tr>
<tr>
<td>Jiang-Conrath</td>
<td>17.07%</td>
<td>57.60%</td>
<td>37.60%</td>
<td>50.13%</td>
<td></td>
</tr>
<tr>
<td>Lin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6 Evaluation Phase III – Semantic Relatedness and Similarity

The third step in our evaluation procedure is related to the semantic measures, which are calculated on the basis of a lexical semantic resource (and word frequency counts) and used in the construction of lexical chains. A semantic measure expresses how much two words have to do with each other. The notion of semantic measure is controversially discussed in the literature e.g. [27]. The two most relevant terms in this context are semantic similarity and semantic relatedness, defined according to [27] as follows:

– Semantic similarity: Word pairs are considered to be semantically similar if any synonymy or hypernymy relations hold. (Examples: forest - wood ⇒ synonymy, flower - rose ⇒ hypernymy, rose - oak ⇒ common hypernym: plant)

– Semantic relatedness: Word pairs are considered to be semantically related if any systematic relation, such as synonymy, antonymy, hypernymy, holonymy, or any unsystematic relation holds. Compared to the semantic similarity measures this is the more general concept, as it includes any intuitive association or linguistically formalized relation between words. (Examples: flower - gardener or monkey - banana ⇒ intuitive association, tree - branch ⇒ holonymy, day - night ⇒ antonymy)

According to the definition by [27], semantic similarity is a subtype of semantic relatedness; in the following section we discuss various relatedness measures. In order to explore these measures and their relevant characteristics, we used the results of our human judgment experiment described in Sect. 3.3.

6.1 GermaNet-based Semantic Relatedness Measures

We expect that good lexical chains include systematic and unsystematic relations, a position which has also been stressed by the experiments reported in [13] and [14].
In fact, most of the established measures merely consider synonymy and hypernymy. Therefore, they actually fall under the notion of semantic similarity.

Figure 5 outlines how the calculation of the relatedness measures interacts with the chaining algorithm and the semantic resource. When the preprocessing is completed, the chaining algorithm selects chaining candidate pairs, in other words, word pairs, for which the relatedness needs to be determined (see Fig. 5 – Query 1: relatedness of word A and B?). Next, the relatedness measure component (RM component) performs a look-up in the semantic resource in order to extract all available features, such as shortest path length or information content of a word, which are necessary to calculate the relatedness value (see Fig. 5 – Query 2: semantic information about A and B?). On the basis of these features, the RM component computes a value which represents the strength of the semantic relation between the two words.

The various measures introduced in the literature use different features and therefore also cover different concepts or aspects of semantic relatedness. We have implemented eight of these measures, which are shortly sketched out below. All eight measures are based on a lexical semantic resource, in our case GermaNet, and some additionally utilize a word frequency list. The first four measures use a hyponym-tree induced from GermaNet. That means, given GermaNet represented as a graph, we exclude all edges except the hyponyms.

---

6 We used a word frequency list computed by Dr. Sabine Schulte im Walde on the basis of the Huge German Corpus (see http://www.schulteimwalde.de/resource.html). We thank Dr. Schulte im Walde for kindly permitting us to use this resource in the framework of our project.
Since this gives us a wood of nine trees, we then connect them to an artificial root and thus construct the required GermaNet hyponym-tree.

- **Leacock-Chodorow** [28]: Given a hyponym-tree, the Leacock-Chodorow measure computes the length of the shortest path between two synonym sets and scales it by the depth of the complete tree.

  \[
  \text{rel}_{\text{LC}}(s_1, s_2) = -\log_2 \frac{2 \cdot \text{sp}(s_1, s_2)}{2 \cdot D_{\text{Tree}}}
  \]  

  \(s_1\) and \(s_2\): the two synonym sets examined; \(\text{sp}(s_1, s_2)\): length of shortest path between \(s_1\) and \(s_2\) in hyponym-tree; \(D_{\text{Tree}}\): depth of the hyponym-tree

- **Wu-Palmer** [29]: Given a hyponym-tree, the Wu-Palmer measure utilizes the least common subsumer in order to compute the similarity between two synonym sets. The least common subsumer is the deepest vertex which is a direct or indirect hyponym of both synonym sets.

  \[
  \text{rel}_{\text{WP}}(s_1, s_2) = \frac{2 \cdot \text{depth}(\text{lcs}(s_1, s_2))}{\text{depth}(s_1) + \text{depth}(s_2)}
  \]  

  \(\text{depth}(s)\): length of the shortest path form root to vertex \(s\); \(\text{lcs}(s)\): least common subsumer of \(s\)

- **Resnik** [30]: Given a hyponym-tree and frequency list, the Resnik measure utilizes the information content in order to compute the similarity between two synonym sets. As typically defined in Information Theory, the information content is the negative logarithm of the probability. Here the probability is calculated on the basis of subsumed frequencies. A subsumed frequency of a synonym set is the sum of frequencies of the set of all words which are in this synonym set, or a direct or indirect hyponym synonym set.

  \[
  p(s) := \frac{\sum_{w \in W(s)} \text{freq}(w)}{\text{TotalFreq}}
  \]

  \[
  \text{IC}(s) := -\log p(s)
  \]

  \[
  \text{rel}_{\text{Res}}(s_1, s_2) = \text{IC}(\text{lcs}(s_1, s_2))
  \]  

  freq(w): frequency of a word within a corpus; \(W(s)\): set of the synonym set \(s\) and all its direct/indirect hyponym synonym sets; \(\text{TotalFreq}\): sum of the frequencies of all words in GermaNet; \(\text{IC}(s)\): information content of the synonym set \(s\)

- **Jiang-Conrath** [31]: Given a hyponym-tree and frequency list, the Jiang-Conrath measure computes the distance (as opposed to similarity) of two synonym sets. The information content of each synonym set is included separately in this distance value, while the information content of the least common subsumer of the two synonym sets is subtracted.

  \[
  \text{dist}_{\text{JC}}(s_1, s_2) = \text{IC}(s_1) + \text{IC}(s_2) - 2 \cdot \text{IC}(\text{lcs}(s_1, s_2))
  \]  

  dist\(_{\text{JC}}\)(\(s_1\), \(s_2\)): distance between synonym sets \(s_1\) and \(s_2\); \(\text{IC}(s)\): information content of the synonym set \(s\)
– **Lin** [32]: Given a hyponym-tree and a frequency list, the Lin measure computes the semantic relatedness of two synonym sets. As the formula clearly shows, the same expressions are used as in Jiang-Conrath. However, the structure is different, as the expressions are divided not subtracted.

\[
\text{rel}_{\text{Lin}}(s_1, s_2) = \frac{2 \cdot \text{IC(lcs}(s_1, s_2)))}{\text{IC}(s_1) + \text{IC}(s_2)}
\] (7)

– **Hirst-StOnge** [4]: In contrast to the four above-mentioned methods, the Hirst-StOnge measure computes the semantic relatedness on the basis of the whole GermaNet graph structure. It classifies the relations considered into 4 classes: *extra strongly related*, *strongly related*, *medium strongly related*, and *not related*. Two words are considered to be

  • *extra strongly related* if they are identical;
  • *strongly related* if they are synonym, antonym or if one of the two words is part of the other one and additionally a direct relation holds between them;
  • *medium strongly related* if there is a path in GermaNet between the two which is shorter than six edges and matches the patterns defined by [2].

In any other case the two words are considered to be unrelated. The relatedness values in the case of extra strong and strong relations are fixed values, whereas the medium strong relation is calculated based on the path length and the number of changes in direction.

– **Tree-Path** (Baseline 1): Given a hyponym-tree, the simple Tree-Path measure computes the length of a shortest path between two synonym sets. Due to its simplicity, the Tree-Path measure serves as a baseline for more sophisticated similarity measures.

\[
\text{dist}_{\text{Tree}}(s_1, s_2) = \text{sp}(s_1, s_2)
\] (8)

– **Graph-Path** (Baseline 2): Given the whole GermaNet graph structure, the simple Graph-Path measure calculates the length of a shortest path between two synonym sets in the whole graph, i.e. the path can make use of all relations available in GermaNet. Analogous to the Tree-Path measure, the Graph-Path measure gives us a very rough baseline for other relatedness measures.

\[
\text{dist}_{\text{Graph}}(s_1, s_2) = \text{sp}_{\text{Graph}}(s_1, s_2)
\] (9)

\text{sp}_{\text{Graph}}(s_1, s_2): \text{Length of a shortest path between } s_1 \text{ and } s_2 \text{ in the GermaNet graph}

**Differences and Challenges**: Most of the measures described in this section are completely based on the hyponym-tree. Therefore, many potentially useful edges of the word net graph structure are not considered, which affects the holonymy (in GermaNet approximately 3,800 edges), meronymy (in GermaNet approximately 900 edges) and antonymy\(^7\) (in GermaNet approximately 1,300 edges) relations. Some of the measures

\(^7\) Because antonyms are mostly organized as co-hyponyms, they are – in fact – not completely discarded in the hyponym-tree-based approaches.
additionally use the least common subsumer. Word pairs featuring potentially different levels of relation are thus subsumed\(^8\). One could also question if this is the only relevant information to be found in the hyponym-tree for a word pair. Interesting features such as network density or node depth are not included. Moreover, several measures rely on the concept of information content, for which a frequency list is required. Thus, the performance of experiments utilizing different lists as a basis is not directly comparable. Especially for lexical chaining, unsystematic relations are considered to be relevant, see e.g. \([21]\) and \([14]\). However, these are not in GermaNet and consequently cannot be considered in any of the measures mentioned above. We therefore expect them to produce many false negatives, i.e. low relation values for word pairs which are judged by humans to be (strongly) related.

**Interpretation of relatedness measure values:** Most of the relatedness measures mentioned in Sect. 6.1 are continuous, with the exception of Hirst-StOnge, Tree-Path and Graph-Path which are all discrete. All of the measures range in a specific interval between 0 (not related) and a maximum value, mostly 1. In any case, for each measure the interval could be normalized into a value ranging between 0 and 1. For the three distance measures, Jiang-Conrath, Tree-Path and Graph-Path, a concrete distance value can be converted into its corresponding relatedness value by subtracting it from the theoretical maximum distance. Suppose we plotted the empirically determined relatedness values\(^9\) against ideal relatedness measure values, we would get exemplary distribution functions as shown in Fig. 6a. For a specific empirically determined value, e.g. 0.5, we then obtained different values for the various measures considered, e.g. 0.27 for measure A and 0.94 for measure B. Thus, the values of a specific relatedness measure A range between 1 and approximately 0.94 for an empirically determined interval of relation strengths (e.g. the word pair is strongly related) whereas a relatedness measure B exhibits values between 1 and 0.27 for the same relations. In order to profitably use this information in our chaining system, we need to interpret the values and thus find intervals mapping between e.g. classes of relation strength and measure values\(^{10}\). In any case, the distribution functions should be noisy, as shown in Fig. 6b – at best indicating a trend function. However, as Figures 7a–c, 8a–c and 9a–b illustrate, the real values of our eight measures plotted against the empirically determined relatedness values do not display any kind of obvious trend function.

---

\(^8\) Given a pair of words \(w_A\) and \(w_B\) and their least common subsumer \(LCS_{AB}\), all pairs of a descendant of \(w_A\) and a descendant of \(w_B\) have \(LCS_{AB}\) as their least common subsumer.

\(^9\) These are the values deduced form our human judgment experiment mentioned in Sect. 3.3.

\(^{10}\) Note that we need to discriminate between the distribution functions (considering empirically determined values and measure values, as exemplarily shown in Fig. 6) and the relatedness functions (as mentioned in Sect. 6.1). Although the two are equal with regards to their output (concrete measure values), they differ with respect to their input dimension and type.
Fig. 6. Idealized (a) and noisy distribution (b) of semantic relatedness values
Fig. 7. Leacock-Chodorow (a), Wu-Palmer (b) and Resnik (c) each plotted against human judgment
Fig. 8. Jiang-Conrath (a), Lin (b) and Hirst-StOnge (c) each plotted against human judgment
Fig. 9. Tree-Path (a) and Graph-Path (b) each plotted against human judgment
6.2 Comparison of Human Judgment and GermaNet-based Measures

Figures 7a–c, 8a–c and 9a–b show values of the various measures for all word pairs of our human judgment experiment described in Sect. 3.3. Although the inter-annotator agreement in the human judgment experiment is relatively high (correlation: 0.76 +/- 0.04), the correlation between the various measures and the human judgment is relatively low (see Table 7). In addition, the trend functions potentially underlying the (very noisy) graphs in Figures 7a–c, 8a–c and 9a–b are not obvious at all.

<table>
<thead>
<tr>
<th></th>
<th>Graph Path</th>
<th>Tree Path</th>
<th>Wu-Palmer</th>
<th>Leacock-Chodorow</th>
</tr>
</thead>
<tbody>
<tr>
<td>correl. coeff.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hirst-StOnge</td>
<td>0.41</td>
<td>0.42</td>
<td>0.36</td>
<td>0.48</td>
</tr>
<tr>
<td>Resnik</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jiang-Conrath</td>
<td>0.47</td>
<td>0.44</td>
<td>0.45</td>
<td>0.48</td>
</tr>
<tr>
<td>Lin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In order to use one of these measures or a combination of them in GLexi, we need to determine the best measure(s) and, because a lexical chainer mostly works with classes of relatedness, a function, which maps these values into discrete intervals of relatedness. We question whether a relatedness measure used in a lexical chainer has to be continuous; a continuous value can misleadingly indicate an unrealistic grade of accuracy. Instead, a measure mapping from a list of features, such as relation type, network density or node depth etc., into three classes, such as not related, related and strongly related might be more adequate. The class distribution in our human judgment experiment shown in Fig. 10 confirms this idea. Because of the relatively low correlation between the measure values and the human judgment, the extreme noise in the distribution functions shown in Figures 7a–c, 8a–c and 9a–b, and the fact that interesting features of GermaNet are not yet considered in the calculation of the relatedness values, we assume that none of the measures presented in this paper is in fact appropriate for lexical chaining in German. In our future work we plan to integrate these findings into a Machine Learning based mapping between GermaNet-based features (and word counts, co-occurrence) and discrete classes of relatedness.

7 Evaluation Phase IV – Application-oriented Evaluation

The constraints imposed on our lexical chainer by the application scenario, i.e. the extraction of topic anchors and the topic chaining itself, are as follows: Firstly, we intend to utilize the structure and information about a specific text encoded in the lexical

---

11 The inter-annotator agreement in our study is slightly lower than those reported in the literature for English because we considered systematically and unsystematically related word pairs as well as abstract and tricky nouns.
An Evaluation Procedure for Word Net Based Lexical Chaining... 143

chained as input features for the extraction of topic anchors. Especially, the length of a chain, the density and strength of its internal linking structure should be of great importance. Admittedly, additional chaining of independent features could be necessary to ultimately determine the topic anchors of a text passage. Secondly, we plan to use the same algorithms and resources for the construction of both lexical and topic chains. Merely the chaining candidates, i.e. all noun tokens for lexical chaining and exclusively topic anchors for topic chaining, account for the difference between the two types of chaining. However, we assume that for both chaining types a net structure could be superior to linearly organized chains. This kind of structure for a passage of a newspaper article, which we computed on the basis of our lexical chainer, is shown in Fig. 11. The article covers child poverty in German society; accordingly, the essential concepts are Kind (Engl. child), Geld (Engl. money), Deutschland (Engl. Germany), and Staat (Engl. state). On the basis of, among other things, edge density and frequency, we calculated the most relevant words (especially, Kind, Geld, Deutschland, and Staat), which we then accordingly highlighted in the graph shown in Fig. 11. Finally, the parameter settings, which we found to be reasonable on the basis of the evaluation phases I–III, need to be integrated with the constraints imposed on our lexical chainer by our application in our future work.
8 Conclusions and Future Work

We explored the various components and aspects of lexical chaining for German corpora of technical and academic documents. We presented a detailed evaluation procedure and discussed the performance of our chaining system with respect to these aspects. We could show that preprocessing plays a major role due to of the complex morphology in German and furthermore that technical terminology and proper names are of great importance. Additionally, we discussed the performance of a simple chaining-based word sense disambiguation and outlined a method to enhance this aspect. We also presented a human judgment experiment which was conducted in order to evaluate the various semantic relatedness measures for GermaNet. We were able to show that it is thus far very difficult to determine the function mapping between the measure values and relatedness classes.

We now plan to continue this work on four levels: Firstly, we hope to further improve the preprocessing; i.e. we plan to enhance the compound analysis and the basic NER system. In addition, we intend to integrate components for the handling of abbreviations and technical terminology. Secondly, we aim to develop a sophisticated chaining-based disambiguation methodology which incorporates the idea of meta-chains and other potentially useful features. Thirdly, we plan to investigate alternative relatedness measures, especially Machine Learning based approaches, which map between sets of features and discrete classes of relatedness. Finally, we intend to further explore our lexical
chainer with respect to topic chaining and thus to evaluate our chainer in an application oriented manner.

References

On the Utility of
Automatically Generated WordNets

Gerard de Melo and Gerhard Weikum
Max Planck Institute for Informatics
Campus E1 4
66123 Saarbrücken, Germany
{demelo,weikum}@mpi-inf.mpg.de

Abstract. Lexical resources modelled after the original Princeton WordNet are being compiled for a considerable number of languages, however most have yet to reach a comparable level of coverage. In this paper, we show that automatically built WordNets, created from an existing WordNet in conjunction with translation dictionaries, are a suitable alternative for many applications, despite the errors introduced by the automatic building procedure. Apart from analysing the resources directly, we conducted tests on semantic relatedness assessment and cross-lingual text classification with very promising results.

1 Introduction

One of the main requirements for domain-independent lexical knowledge bases, apart from an appropriate data model, is a satisfactory level of coverage. WordNet is the most well-known and most widely used lexical database for English natural language processing, and is the fruit of over 20 years of manual work carried out at Princeton University [1]. The original WordNet has inspired the creation of a considerable number of similarly-structured resources for other languages (“WordNets”), however, compared to the original, many of these still exhibit a rather low level of coverage due to the laborious compilation process. In this paper, we argue that, depending on the particular task being pursued, one can instead often rely on machine-generated WordNets, created with translation dictionaries from an existing WordNet such as the original WordNet.

The remainder of this paper is laid out as follows. In Section 2 we provide an overview of strategies for building WordNets automatically, focusing in particular on a recent machine learning approach. Section 3 then evaluates the quality of a German WordNet built using this technique, examining the accuracy, coverage, as well as the general appropriateness of automatic approaches. This is followed by further investigations motivated by more pragmatic considerations. After considering human consultation in Section 4, we proceed to look more closely at possible computational applications, discussing our results in monolingual tasks such as semantic relatedness estimation in Section 5, and multilingual
ones such as cross-lingual text classification in Section 6. We conclude with final remarks and an exploration of future research directions in Section 7.

2 Building WordNets

In this section, we summarize some of the possible techniques for automatically creating WordNets fully aligned to an existing WordNet. We do not consider the so-called merge model, which normally requires some pre-existing WordNet-like thesaurus for the new language, and instead focus on the expand model, which mainly relies on translations [2]. The general approach is as follows: (1) Take an existing WordNet for some language $L_0$, usually Princeton WordNet for English (2) For each sense $s$ listed by the WordNet, translate all the terms associated with $s$ from $L_0$ to a new language $L_N$ using a translation dictionary (3) Additionally retain all appropriate semantic relations between senses in order to arrive at a new WordNet for $L_N$.

The main challenge lies in determining which translations are appropriate for which senses. A dictionary translating an $L_0$-term $e$ to an $L_N$-term $t$ does not imply that $t$ applies to all senses of $e$. For example, considering the translation of English “bank” to German “Bank”, we can observe that the English term can also be used for riverbanks, while the German “Bank” cannot (and likewise, German “Bank” can also refer to a park bench, which does not hold for the English term).

In order to address these problems, several different heuristics have been proposed. Okumura and Hovy [3] linked a Japanese lexicon to an ontology based on WordNet synsets. They considered four different strategies: (1) simple heuristics based on how polysemous the terms are with respect to the number of translations and with respect to the number of WordNet synsets (2) checking whether one ontology concept is linked to all of the English translations of the Japanese term (3) compatibility of verb argument structure (4) degree of overlap between terms in English example sentences and translated Japanese example sentences.

Another important line of research starting with Rigau and Agirre [4], and extended by Atserias et al. [5] resulted in automatic techniques for creating preliminary versions of the Spanish WordNet and later also the Catalan WordNet [6]. Several heuristic decision criteria were used in order to identify suitable translations, e.g. monosemy/polysemy heuristics, checking for senses with multiple terms having the same $L_N$-translation, as well as heuristics based on conceptual distance measures. Later, these were combined with additional Hungarian-specific heuristics to create a Hungarian nominal WordNet [7].

Pianta et al. [8] used similar ideas to produce a ranking of candidate synsets. In their work, the ranking was not used to automatically generate a WordNet but merely as an aid to human lexicographers that allowed them to work at faster pace. This approach was later also adopted for the Hebrew WordNet [9].

A more advanced approach that requires only minimal human work lies in using machine learning algorithms to identify more subtle decision rules that can
On the Utility of Automatically Generated WordNets

rely on a number of different heuristic scores with different thresholds. We will
briefly summarize our approach [10]. A classifier \( f \) is trained on labelled examples
\((x_i, y_i)\) for pairs \((t_i, s_i)\), where \( t_i \) is an \( L_N \)-term and \( s_i \) is a candidate sense for \( t_i \).
Each labelled instance consists of a real-valued feature vector \( x_i \), and an indicator
\( y_i \in \mathcal{Y} = \{0, 1\} \), where 1 denotes a positive example, which implies that linking
\( t_i \) with sense \( s_i \) is appropriate, and 0 characterizes negative examples. Based
on these training examples, \( f \) classifies new unseen test instances by computing
a confidence value \( y \in [0, 1] \) that indicates to what degree an association is
predicted to be correct. One may then obtain a confidence value \( y_{t,s} \) for each
possible pair \((t, s)\) where \( t \) is a \( L_N \)-term translated to an \( L_0 \)-term that is in turn
linked to a sense \( s \). These values can be used to create the new WordNet by
either maintaining all \( y_{t,s} \) as weights in order to create a weighted WordNet, or
alternatively one can use confidence thresholds to obtain a regular unweighted
WordNet. For the latter case, we use two thresholds \( \alpha_1, \alpha_2 \), and accept a pair
\((t, s)\) if \( y_{t,s} \geq \alpha_1 \), or alternatively if \( \alpha_1 > y_{t,s} \geq \alpha_2 \) and \( y_{t,s} > y_{t,s'} \) for all \( s' \neq s \).

The feature vectors \( x_i \) are created by computing a variety of scores based
on statistical properties of the \((t, s)\) pair as feature values. We mainly rely on a
multitude of semantic overlap scores reflecting the idea that senses with a high
semantic proximity to other candidate senses are more likely to be appropriate, as
well as polysemy scores that reflect the idea that a sense becomes more important
when there are few relevant alternative senses. The former are computed as
\[
\sum_{e \in \phi(t)} \max_{s' \in \sigma(e)} \gamma(t, s') \text{rel}(s, s') \tag{1}
\]
while for the latter we use
\[
\sum_{e \in \phi(t)} 1 + \sum_{s' \in \sigma(e)} \frac{1_{\sigma(e)}(s)}{\gamma(t, s')(1 - \text{rel}(s, s'))}. \tag{2}
\]
In these formulae, \( \phi(t) \) yields the set of translations of \( t \), \( \sigma(e) \) yields the set of
senses of \( e \), \( \gamma(t, s) \) is a weighting function, and \( \text{rel}(s, s') \) is a semantic relatedness
function between senses. The characteristic function \( 1_{\sigma(e)}(s) \) yields 1 if \( s \in \sigma(e) \)
and 0 otherwise. We use a number of different weighting functions \( \gamma(t, s) \) that
take into account lexical category compatibility, corpus frequency information,
etc., as well as multiple relatedness functions \( \text{rel}(s, s') \) based on gloss similarity
and graph distance (cf. Section 5.2).

This approach has several advantages over previous proposals: (1) Apart from
the translation dictionary, it does not rely on additional resources such as mono-
lingual dictionaries with field descriptors, verb argument structure information,
and the like for the target language \( L_N \), and thus can be used in many settings,
(2) the learning algorithm can exploit real-valued heuristic scores rather than
just predetermined binary decision criteria, leading to a greater coverage, (3) the
algorithm can take into account complex dependencies between multiple scores
rather than just single heuristics or combinations of two heuristics.
3 Analysis of a Machine-Generated WordNet

In the remainder of this paper, we will focus on a German-language WordNet produced using the machine learning technique described above, as it is the most advanced approach. The WordNet was generated from Princeton WordNet 3.0 and the Ding translation dictionary [11] using a linear kernel support vector machine [12] with posterior probability estimation as implemented in LIBSVM [13]. The training set consisted of 1834 candidate mappings for 350 randomly selected German terms that were manually classified as correct (22%) or incorrect. The values $\alpha_1 = 0.5$ and $\alpha_2 = 0.45$ were chosen as classification thresholds.

3.1 Accuracy and Coverage

In order to evaluate the quality of the WordNet generated in this manner, we considered a test set of term-sense mappings for 350 further randomly selected terms. We then determined whether the resulting 1624 mappings, which had not been involved in the WordNet building process, corresponded with the entries of our new WordNet. Table 1 summarizes the results, showing the precision and recall with respect to this test set.

Table 1. Evaluation of precision and recall on an independent test set

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>79.87</td>
<td>69.40</td>
</tr>
<tr>
<td>verbs</td>
<td>91.43</td>
<td>57.14</td>
</tr>
<tr>
<td>adjectives</td>
<td>78.46</td>
<td>62.96</td>
</tr>
<tr>
<td>adverbs</td>
<td>81.81</td>
<td>60.00</td>
</tr>
<tr>
<td>overall</td>
<td>81.11</td>
<td>65.37</td>
</tr>
</tbody>
</table>

The results demonstrate that indeed a surprisingly high level of precision and recall can be obtained with fully automated techniques, considering the difficulty of the task. While the precision might not fulfill the high lexicographical standards adopted by traditional dictionary publishers, we shall later see that it suffices for many practical applications. Furthermore, one of course may obtain a higher level of precision at the expense of a lower recall by adjusting the acceptance thresholds. For very high recall levels, an increased precision might not be realistic even using purely manual work, considering that Miháltz and Prőszéky [7] report an inter-annotator agreement of 84.73% for such mappings.

Table 2 shows that applying the classification thresholds to all terms in the dictionary leads to a WordNet with a considerable coverage. While smaller than GermaNet 5.0 [14], one of the largest WordNets, it covers more senses than any of the original eight WordNets delivered by the EuroWordNet project [2]. Table 3
Table 2. Quantitative Assessment of Coverage of the German WordNet

<table>
<thead>
<tr>
<th></th>
<th>sense mappings</th>
<th>terms</th>
<th>lexicalized synsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>53146</td>
<td>35089</td>
<td>28007</td>
</tr>
<tr>
<td>verbs</td>
<td>13875</td>
<td>5908</td>
<td>6304</td>
</tr>
<tr>
<td>adjectives</td>
<td>21799</td>
<td>13772</td>
<td>9949</td>
</tr>
<tr>
<td>adverbs</td>
<td>4243</td>
<td>2992</td>
<td>2593</td>
</tr>
<tr>
<td>total</td>
<td>93063</td>
<td>55522</td>
<td>46853</td>
</tr>
</tbody>
</table>

gives an overview of the polysemy of the terms as covered by our WordNet, with arithmetic means computed from the polysemy either of all terms, or exclusively from terms polysemous with respect to the WordNet.

Table 3. Polysemy of Terms and Mean Number of Lexicalizations (excluding unlexicalized senses)

<table>
<thead>
<tr>
<th></th>
<th>mean term polysemy</th>
<th>mean term polysemy excluding monosemous</th>
<th>mean no. of sense lexicalizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>1.51</td>
<td>2.95</td>
<td>1.90</td>
</tr>
<tr>
<td>verbs</td>
<td>2.35</td>
<td>4.36</td>
<td>2.20</td>
</tr>
<tr>
<td>adjectives</td>
<td>1.58</td>
<td>2.79</td>
<td>2.19</td>
</tr>
<tr>
<td>adverbs</td>
<td>1.42</td>
<td>2.52</td>
<td>1.64</td>
</tr>
<tr>
<td>total</td>
<td>1.68</td>
<td>3.07</td>
<td>1.99</td>
</tr>
</tbody>
</table>

A more qualitative assessment of the accuracy and coverage revealed the following issues:

- Non-Uniformity of Coverage: While even many specialized terms are included (e.g. “Kokarde”, “Vasokonstriktion”), certain very common terms were found to be missing (e.g. “Kofferraum”, “Schloss”). This seems to arise from the fact that common terms tend to be more polysemous, though frequently such terms also have multiple translations, which tends to facilitate the mapping process. One solution would be manually adding mappings for terms with high corpus frequency values, which due to Zipf’s law would quickly improve the relative coverage of the terms in ordinary texts.

- Lexical Gaps and Incongruencies: Another issue is the lack of terms for which there are no lexicalized translations in the English language, or which are not covered by the source WordNet, e.g. the German word “Feierabend” means the finishing time of the daily working hours. The solution could consist
in smartly adding new senses to the sense hierarchy based on paraphrasing translations (e.g. as a hyponym of “time” for our current example).

- Multi-word expressions in $L_N$: Certain multi-word translations in $L_N$ might be considered inappropriate for inclusion in a lexical resource, e.g. the Ding dictionary lists “Jahr zwischen Schule und Universität” as a translation of “gap year”. By generally excluding all multi-word expressions one would also likely drop a lot of lexicalized expressions, e.g. German “runde Klammer” (parenthesis). A much better solution is to automatically mark all multi-word expressions as possibly unlexicalized whenever no matching entry is found in monolingual dictionaries.

3.2 Relational Coverage

By producing mappings to senses of an existing source WordNet, we have the great advantage of immediately being able to import relations between those synsets. An excerpt of some of the relations we imported is given in Table 4.

Table 4. An excerpt of some of the imported relations. We distinguish full links between two senses both with $L_N$-lexicalizations, and outgoing links from senses with an $L_N$ lexicalization.

<table>
<thead>
<tr>
<th>relation</th>
<th>full links</th>
<th>outgoing</th>
</tr>
</thead>
<tbody>
<tr>
<td>hyponymy</td>
<td>26324</td>
<td>60062</td>
</tr>
<tr>
<td>hypernymy</td>
<td>26324</td>
<td>33725</td>
</tr>
<tr>
<td>similarity</td>
<td>10186</td>
<td>14785</td>
</tr>
<tr>
<td>has category</td>
<td>2131</td>
<td>2241</td>
</tr>
<tr>
<td>category of</td>
<td>2131</td>
<td>6135</td>
</tr>
<tr>
<td>has instance</td>
<td>641</td>
<td>5936</td>
</tr>
<tr>
<td>instance of</td>
<td>641</td>
<td>1131</td>
</tr>
<tr>
<td>part meronymy</td>
<td>2471</td>
<td>6029</td>
</tr>
<tr>
<td>part holonymy</td>
<td>2471</td>
<td>3408</td>
</tr>
<tr>
<td>member meronymy</td>
<td>400</td>
<td>734</td>
</tr>
<tr>
<td>member holonomy</td>
<td>400</td>
<td>1517</td>
</tr>
<tr>
<td>subst. meronymy</td>
<td>190</td>
<td>325</td>
</tr>
<tr>
<td>subst. holonomy</td>
<td>190</td>
<td>414</td>
</tr>
</tbody>
</table>

Lexical relations between particular terms cannot, in general, be transferred automatically, e.g. a region domain for a term in one language, signifying in what geographical region the term is used, will not apply to a second language. However, certain lexical relations such as the derivation relation still provide valuable information when interpreted as a general indicator of semantic relatedness, as can be seen in Table 5, which shows the results of a human evaluation for several different relation types. Incorrect relations are almost entirely due to incorrect term-sense mappings.
Table 5. Quality assessment for imported relations: For each relation type, 100 randomly selected links between two senses with $L_N$-lexicalizations were evaluated.

<table>
<thead>
<tr>
<th>relation</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>hyponymy, hypernymy</td>
<td>84%</td>
</tr>
<tr>
<td>similarity</td>
<td>90%</td>
</tr>
<tr>
<td>category</td>
<td>91%</td>
</tr>
<tr>
<td>instance</td>
<td>93%</td>
</tr>
<tr>
<td>part meronymy, holonymy</td>
<td>83%</td>
</tr>
<tr>
<td>member meronymy, holonymy</td>
<td>89%</td>
</tr>
<tr>
<td>subst. meronymy, holonymy</td>
<td>83%</td>
</tr>
<tr>
<td>antonymy (as sense opposition)</td>
<td>95%</td>
</tr>
<tr>
<td>derivation (as semantic similarity)</td>
<td>96%</td>
</tr>
</tbody>
</table>

3.3 Structural Adequacy

As mentioned earlier, our machine learning approach is very parsimonious with respect to $L_N$-specific prerequisites, and hence scales well to new languages. Some lexicographers contend that using one WordNet as the structural basis for another WordNet does not do justice to the structure of the new language’s lexicon.

The most significant issue is certainly that the source WordNet may lack senses for certain terms in the new language, as in the case of the German “Feierabend”. This point has already been addressed in Section 3.1.

Apart from this, it seems that general structural differences between languages rarely cause problems. When new WordNets are built independently from existing WordNets, many of the structural differences will not be due to actual conceptual differences between languages, but rather result from subjective decisions made by the individual human modellers [8].

Some of the rare examples of cultural differences affecting relations between two senses include perhaps the question of whether the local term for “guinea pig” should count as a hyponym of the respective term for “pet”. For such cases, our suggestion is to manually add relation attributes that describe the idea of a connection being language-specific, culturally biased, or based on a specific taxonomy rather than holding unconditionally.

A more general issue is the adequacy of the four lexical categories (parts of speech) considered by Princeton WordNet. Fortunately, most of the differences between languages in this respect either concern functional words, or occur at very fine levels of distinctions, e.g. genus distinctions for German nouns, and thus are conventionally considered irrelevant to WordNets, though such information could be derived from monolingual dictionaries and added to the WordNet.
4 Human Consultation

One major disadvantage of automatically built WordNets is the lack of native-language glosses and example sentences, although this problem is not unique to automatically-built WordNets. Because of the great effort involved in compiling such information, manually built WordNets such as GermaNet also lack glosses and example sentences for the overwhelming majority of the senses listed. In this respect, automatically produced aligned WordNets have the advantage of at least making English-language glosses accessible.

Another significant issue is the quality of the mappings. As people are more familiar with high-quality print dictionaries, they do not expect to encounter incorrect entries when consulting a WordNet-like resource.

In contrast, we found that machine-generated WordNets can instead be used to provide machine-generated thesauri, where users expect to find more generally related terms rather than precise synonyms and gloss descriptions. In order to generate such a thesaurus, we relied on a simple technique that looks up all senses of a term as well as certain related senses, and then forms the union of all lexicalizations of these senses (Algorithm 4.1 with \( n_h = 2, n_o = 2, n_g = 1 \)). Table 6 provides a sample entry from the German thesaurus resulting from our WordNet, and demonstrates that such resources can indeed be used for example as built-in thesauri in word processing applications.

### Algorithm 4.1 Thesaurus Generation

**Input:** a WordNet instance \( W \) (with function \( \sigma \) for retrieving senses and \( \sigma^{-1} \) for retrieving the set of all terms for a sense), number of hypernym levels \( n_h \), number of hyponym levels \( n_o \), number of levels for other general relations \( n_g \), set of acceptable general relations \( R \)

**Objective:** generate a thesaurus that lists related terms for any given term

```plaintext
1: procedure GenerateThesaurus(W, R)
2:   for each term t from W do \( \triangleright \) for every term t listed in the WordNet
3:     T ← ∅ \( \triangleright \) the list of related terms for t
4:     for each sense \( s \in \sigma(t) \) do \( \triangleright \) for each sense of t
5:       for each sense \( s' \in \text{Related}(W,s,n_h,n_o,n_g,R) \) do
6:         T ← T ∪ \( \sigma^{-1}(s') \) \( \triangleright \) add lexicalizations of \( s' \) to \( T \)
7:     output T as list of related terms for t
8: function Related(W, s, n_h, n_o, n_g, R)
9:   S ← {s}
10:  for each sense \( s' \) related to s with respect to \( W \) do \( \triangleright \) recursively visit related senses
11:    if \( (s' \text{ hypernym of } s) \wedge (n_h > 0) \) then
12:      S ← S ∪ Related(W, s', n_h - 1, 0, 0, 0)
13:    else if \( (s' \text{ hyponym of } s) \wedge (n_o > 0) \) then
14:      S ← S ∪ Related(W, s', 0, n_o - 1, 0, 0)
15:    else if \( \exists r \in R : (s' \text{ stands in relation } r \text{ to } s) \wedge (n_g > 0) \) then
16:      S ← S ∪ Related(W, s', 0, 0, n_g - 1, R)
17: return S
```

Table 6 provides a sample entry from the German thesaurus resulting from our WordNet, and demonstrates that such resources can indeed be used for example as built-in thesauri in word processing applications.
Table 6. Sample entries from generated thesaurus (which contains entries for 55522 terms, each entry listing 17 additional related terms on average)

<table>
<thead>
<tr>
<th>headword: Leseratte</th>
</tr>
</thead>
<tbody>
<tr>
<td>Büchgelehrte, Büchgelehrter, Bücherwurm, Geisteswissenschaftler, Gelehrte, Gelehrter, Stubengelehrte, Stubengelehrter, Student, Studentin, Wissenschaftler</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>headword: leserlich</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesbarkeit, Verständlichkeit, deutlich, entzifferbar, klar, lesbar, lesenswert, unlesbar, unleserlich, übersichtlich</td>
</tr>
</tbody>
</table>

5 Monolingual Applications

5.1 General Remarks

Although at first it might seem that having WordNets aligned to the original WordNet is mainly beneficial for cross-lingual tasks, it turns out that the alignment also proves to be a major asset for monolingual applications, as one can leverage much of the information associated with the Princeton WordNet, e.g. the included English-language glosses, as well as a wide range of third-party resources, incl. topic domain information [15], links to ontologies such as SUMO [16] and YAGO [17], etc.

For instance, for the task of word sense disambiguation, a preliminary study using an algorithm that maximizes the overlap of the English-language glosses [18] showed promising results, although we were unable to evaluate it more adequately due to the lack of an appropriate sense-tagged test corpus. One problem we encountered, however, was that the generated WordNet sometimes did not cover all of the terms and senses to be disambiguated, which means that it is not an ideal sense inventory for word sense disambiguation tasks.

Apart from that, generated WordNets can be used for most other tasks that the English WordNet is usually employed for, including text and multimedia retrieval, text classification, text summarization, as well as semantic relatedness estimation, which we will now consider in more detail.

5.2 Semantic Relatedness

Several studies have attempted to devise means of automatically approximating semantic relatedness judgments made by humans, predicting e.g. that most humans consider the two terms “fish” and “water” semantically related. Such relatedness information is useful for a number of different tasks in information retrieval and text mining, and various techniques have been proposed, many relying on lexical resources such as WordNet. For the German language, Gurevych [19] reported that Lesk-style similarity measures based on the similarity of gloss descriptions [20] do not work well in their original form because GermaNet features only very few glosses, and those that do exist tend to be rather short. With
machine-generated aligned WordNets, however, one can apply virtually any existing measure of relatedness that is based on the English WordNet, because English-language glosses and co-occurrence data are available.

We proceeded using the following assessment technique. Given two terms $t_1$, $t_2$, we estimate their semantic relatedness using the maximum relatedness score between any of their two senses:

$$\text{rel}(t_1, t_2) = \max_{s_1 \in \sigma(t_1)} \max_{s_2 \in \sigma(t_2)} \text{rel}(s_1, s_2)$$

(3)

For the relatedness scores, we consider three different approaches.

1. Graph distance: We consider the graph constituted by WordNet’s senses and sense relations, and compute proximity scores for nodes in the graph by taking the maximum of the products of relation-specific edge weights for any two paths between two nodes.

2. Gloss Similarity: For each sense in WordNet, extended gloss descriptions are created by concatenating the glosses and lexicalizations associated with the sense as well as those associated with certain related senses (senses connected via hyponymy, derivation/derived, member/part holonymy, and instance relations, as well as two levels of hypernyms). Each gloss description is then represented as a bag-of-words vector, where each dimension represents the TF-IDF value of a stemmed term from the glosses. For two senses $s_1$, $s_2$, one then computes the inner product of the two corresponding gloss vectors $c_1$, $c_2$ to determine the cosine of the angle $\theta_{c_1, c_2}$ between them, which characterizes the amount of term overlap for the two context strings:

$$\cos \theta_{c_1, c_2} = \frac{\langle c_1, c_2 \rangle}{||c_1|| \cdot ||c_2||}$$

(4)

3. Maximum: Since the two measures described above are based on very different information, we combined them into a meta-method that always chooses the maximum of these two relatedness scores.

For evaluating the approach, we employed three German datasets [19, 21] that capture the mean of relatedness assessments made by human judges. In each case, the assessments computed by our methods were compared with these means, and Pearson’s sample correlation coefficient was computed. The results are displayed in Table 7, where we also list the current state-of-the-art scores obtained for GermaNet and Wikipedia as reported by Gurevych et al. [22].

The results show that our semantic relatedness measures lead to near-optimal correlations with respect to the human inter-annotator agreement correlations. The main drawback of our approach is a reduced coverage compared to Wikipedia and GermaNet, because scores can only be computed when both parts of a term pair are covered by the generated WordNet.
Table 7. Evaluation of semantic relatedness measures, using Pearson’s sample correlation coefficient in %. We compare our three semantic relatedness measures based on the automatically generated WordNet with the agreement between human annotators and scores for two alternative measures as reported by Gurevych et al. [22], one based on Wikipedia, the other on GermaNet.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GUR65 Pearson r</th>
<th>Coverage</th>
<th>GUR350 Pearson r</th>
<th>Coverage</th>
<th>ZG222 Pearson r</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-Annot. Agreement</td>
<td>0.81</td>
<td>(65)</td>
<td>0.69</td>
<td>(350)</td>
<td>0.49</td>
<td>(222)</td>
</tr>
<tr>
<td>Wikipedia (ESA)</td>
<td>0.56</td>
<td>65</td>
<td>0.52</td>
<td>333</td>
<td>0.32</td>
<td>205</td>
</tr>
<tr>
<td>GermaNet (Lin)</td>
<td>0.73</td>
<td>60</td>
<td>0.50</td>
<td>208</td>
<td>0.08</td>
<td>88</td>
</tr>
<tr>
<td>Gen. WordNet (graph)</td>
<td>0.72</td>
<td>54</td>
<td>0.64</td>
<td>185</td>
<td>0.41</td>
<td>89</td>
</tr>
<tr>
<td>Gen. WordNet (gloss)</td>
<td>0.77</td>
<td>54</td>
<td>0.59</td>
<td>185</td>
<td>0.47</td>
<td>89</td>
</tr>
<tr>
<td>Gen. WordNet (max.)</td>
<td>0.75</td>
<td>54</td>
<td>0.67</td>
<td>185</td>
<td>0.44</td>
<td>89</td>
</tr>
</tbody>
</table>

One advantage of our approach is that it may also be applied without any further changes to the task of cross-lingually assessing the relatedness of English terms with German terms. In the following section, we will take a closer look at the general suitability of our WordNet for multilingual applications.

6 Multilingual Applications

6.1 General Remarks

We can distinguish the following two categories of applications with multilingual support.

- multilingual applications that need to support certain operations on more than just a single language, e.g. word processors with thesauri for multiple languages
- multilingual applications that perform cross-lingual operations

By creating isolated WordNets for many different languages one addresses only the first case. For the second case, one can use multiple WordNets for different languages where the senses are strongly interlinked. The ideal case is when there is no sense duplication, i.e. if two words in different languages share the same meaning, they should be linked to the same sense. The techniques described in Section 2 achieve this by producing WordNets that are strictly aligned to the source WordNet whenever appropriate.

Aligned WordNets thus can be used for various cross-lingual tasks, including cross-lingual information retrieval [23], and cross-lingual text classification, which will now be studied.
6.2 Cross-Lingual Text Classification

Text classification is the task of assigning text documents to the classes or categories considered most appropriate, thereby e.g. topically distinguishing texts about thermodynamics from others dealing with quantum mechanics. This is commonly achieved by representing each document using a vector in a high-dimensional feature space where each feature accounts for the occurrence of a particular term from the document set (a bag-of-words model), and then applying machine learning techniques such as support vector machines. For more information, please refer to Sebastiani’s survey [24].

Cross-lingual text classification is a much more challenging task. Since documents from two different languages obviously have completely different term distributions, the conventional bag-of-words representations perform poorly. Instead, it is necessary to induce representations that tend to give two documents from different languages similar representations when their content is similar.

One means of achieving this is the use of language-independent conceptual feature spaces where the feature dimensions represent meanings of terms rather than just the original terms. We process a document by removing stop words, performing part-of-speech tagging and lemmatization using the TreeTagger [25], and then map each term to the respective sense entries listed by the WordNet instance. In order to avoid decreasing recall levels, we do not disambiguate in any way other than acknowledging the lexical category of a term, but rather assign each sense $s$ a local score
\[
\sum_{s' \in \sigma(t)} w_{t,s'}
\]
whenever a term $t$ is mapped to multiple senses $s \in \sigma(t)$. Here, $w_{t,s}$ is the weight of the link from $t$ to $s$ as provided by the WordNet if the lexical category between document term and sense match, or 0 otherwise. We test two different setups: one relying on regular unweighted WordNets ($w_{t,s} \in \{0, 1\}$), and another based on a weighted German WordNet ($w_{t,s} \in [0, 1]$), as described in Section 2. Since the original document terms may include useful language-neutral terms such as names of people or organizations, they are also taken into account as tokens with a weight of 1. By summing up the weights for each local occurrence of a token $t$ (a term or a sense) within a document $d$, one arrives at document-level token occurrence scores $n(t, d)$, from which one can then compute TF-IDF-like feature vectors using the following formula:
\[
\log(n(t, d) + 1) \log \left( \frac{|D|}{|\{d \in D \mid n(t, d) \geq 1\}|} \right)
\]
where $D$ is the set of training documents.

This approach was tested using a cross-lingual dataset derived from the Reuters RCV1 and RCV2 collections of newswire articles [26, 27]. We randomly selected 15 topics shared by the two corpora in order to arrive at $\binom{15}{2} = 105$ binary classification tasks, each based on 200 training documents in one language, and 600 test documents in a second language, likewise randomly selected, however ensuring equal numbers of positive and negative examples in order to avoid
biased error rates. We considered a) German training documents and English test documents and b) English training documents and German test documents. For training, we relied on the SVMlight implementation [28] of support vector machine learning [12], which is known to work very well for text classification.

Table 8. Evaluation of cross-lingual text classification in terms of micro-averaged accuracy, precision, recall, and $F_1$-score for a German-English as well as an English-German setup. We compare the standard bag-of-words TF-IDF representation with two WordNet-based representations, one using an unweighted, the other based on a weighted German WordNet.

<table>
<thead>
<tr>
<th></th>
<th>acc.</th>
<th>prec.</th>
<th>rec.</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>German-English</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF-IDF</td>
<td>80.56</td>
<td>77.49</td>
<td>86.14</td>
<td>81.59</td>
</tr>
<tr>
<td>WordNet (unweighted)</td>
<td>87.09</td>
<td>85.27</td>
<td>89.68</td>
<td>87.42</td>
</tr>
<tr>
<td>WordNet (weighted)</td>
<td>87.98</td>
<td>85.48</td>
<td>91.51</td>
<td>88.39</td>
</tr>
<tr>
<td>English-German</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF-IDF</td>
<td>78.82</td>
<td>79.19</td>
<td>78.20</td>
<td>78.69</td>
</tr>
<tr>
<td>WordNet (unweighted)</td>
<td>85.39</td>
<td>87.38</td>
<td>82.74</td>
<td>84.99</td>
</tr>
<tr>
<td>WordNet (weighted)</td>
<td>87.47</td>
<td>87.73</td>
<td>87.07</td>
<td>87.40</td>
</tr>
</tbody>
</table>

The results in Table 8 clearly show that automatically built WordNets aid in cross-lingual text classification. Since many of the Reuters topic categories are business-related, using only the original document terms, which include names of companies and people, already works surprisingly well, though presumably not well enough for use in production settings. By considering WordNet senses, both precision and recall are boosted significantly. This implies that English terms in the training set are being mapped to the same senses as the corresponding German terms in the test documents. Using the weighted WordNet version further improves the recall, as more relevant terms and senses are covered.

7 Conclusions

We have shown that machine-generated WordNets are useful for a number of different purposes. First of all, of course, they can serve as a valuable starting point for establishing more reliable WordNets, which would involve manually extending the coverage and addressing issues arising from differences between the lexicons of different languages.

At the same time, machine-generated WordNets can be used directly without further manual work to generate thesauri for human use, or for a number of different natural language processing applications, as we have shown in particular for semantic relatedness estimation and cross-lingual text classification.
In the future, we would like to investigate techniques for extending the coverage of such statistically generated WordNets to senses not covered by the original Princeton WordNet. We hope that our research will aid in contributing to making lexical resources available for languages which to date have not been dealt with by the WordNet community.

References


Words, Concepts and Relations
in the Construction of Polish WordNet*

Magdalena Derwojedowa¹, Maciej Piasecki², Stanisław Szpakowicz³,⁴,
Magdalena Zawisławska¹, and Bartosz Broda²

¹ Institute of the Polish Language, Warsaw University,
{derwojed,zawisla}@uw.edu.pl

² Institute of Applied Informatics, Wrocław University of Technology,
{maciej.piasecki,bartosz.broda}@pwr.wroc.pl

³ School of Information Technology and Engineering, University of Ottawa,
szp@site.uottawa.ca

⁴ Institute of Computer Science, Polish Academy of Sciences

Abstract. A Polish WordNet has been under construction for two years.
We discuss the organisation of the project, the fundamental assumptions,
the tools and the resources. We show how our work differs from that
done on EuroWordNet and BalkaNet. In a year we expect the network
to reach 20000 lexical units. Some 12000 entries will have been completed
by hand. Work on others will be automated as far as possible; to that
end, we have developed statistics-based semantic similarity functions and
methods based on a form of chunking. The preliminary results show that
at least semi-automated acquisition of relations is feasible, so that the
lexicographers’ work may be reduced to revision and approval.

1 Organisation of the project

Ever since the initial burst of popularity of the original WordNet [1, 2], there
has been little doubt how useful WordNets are in Natural Language Processing.
For those who work with a language that lacks a WordNet, the question is not
whether, but how and how fast to construct such a lexical resource. The con-
struction is costly, with the bulk of the cost due to the high linguistic workload.
This appears to have been the case, in particular, in two multinational WordNet-
building projects, EuroWordNet [3] and BalkaNet [4]. The recent developments
in automatic acquisition of lexical-semantic relations suggest that the cost might
be reduced. Our project to construct a Polish WordNet (pWordNet) explores
this path as a supplement to a well-organized and well-supported effort of a team
of linguists/lexicographers.

* Work financed by the Polish Ministry of Education and Science, Project No. 3 T11C 018 29.
The three-year project started in November 2005. The Polish Ministry of Education and Science funds it with a very modest ca. 65000 euro (net). The stated main objective is the development of algorithms of automatic acquisition of lexical-semantic relations for Polish, but we envisage the manual, software-assisted creation of some 15000 to 20000 lexical units (LUs) as an important side-effect. The evolving network also plays an essential role in the automated acquisition of relations. We describe the current state of the project in Section 3.3.

We will automate part of the development effort. A core of about 7000 LUs has been constructed completely manually; in a form of bootstrapping, the remainder of the initial plWordNet will be built semi-automatically. Algorithms that generate synonym suggestions from a large corpus [5] will make suggestions for the linguists to act upon. The ultimate responsibility for every entry rests with its authors, in keeping with our general principle of high trustworthiness of the resource. We must, however, try to reduce the linguists’ workload and thus the time it takes to construct a network of a size comparable to several much more established European WordNets. We have allotted the funds approximately in the proportion 1:2 to manual work and to the software design and development work.

The remainder of the paper presents a more detailed overview of decisions made and work done till now, reviews the lessons learned, and sketches the plan for the last year of this project.

2 Fundamental assumptions

The backbone of any WordNet is its system of semantic relations. Two principles guided our design of the set of relations for Polish WordNet (plWordNet): we should — for obvious portability reasons — stay as close as possible to the Princeton WordNet (WN) set and the EuroWordNet (EWN) set, but we should also respect the specific properties of the Polish language, especially its very rich morphology. Tables 1 and 2 summarise our decisions.

In our description we have kept the division of lexemes into grammatical classes (parts of speech, as in WN): nouns, verbs and adjectives. Relations other than relatedness and pertains to connect lexemes in the same class. Some relations are symmetrical (e.g., if A is an antonym of B, then B is an antonym of A; the hyponymy-hypernymy pair is symmetrical, too), while others are not (e.g., holonymy: a spoke is part of a wheel, but not every wheel has spokes). We refer to this property of semantic relations as reversibility.

We consider it a more precise measure of WordNet size than the number of synsets.

Variously interconnected LUs – lexemes, generally speaking – are the basic building blocks of plWordNet.

EWN has introduced a number of other relations which are not relevant to the discussion in this paper.
Table 1. Semantic relations in WordNet, EuroWordNet and Polish WordNet

<table>
<thead>
<tr>
<th>relation</th>
<th>WordNet</th>
<th>EuroWordNet</th>
<th>Polish WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>synonymy</td>
<td>synonymy</td>
<td>synonymy</td>
<td>synonymy</td>
</tr>
<tr>
<td>antonymy</td>
<td>antonymy</td>
<td>antonymy</td>
<td>antonymy</td>
</tr>
<tr>
<td>hypo-/hypernymy</td>
<td>hypo-/hypernymy</td>
<td>hypo-/hypernymy</td>
<td></td>
</tr>
<tr>
<td>mero-/holonymy</td>
<td>mero-/holonymy</td>
<td>mero-/holonymy</td>
<td></td>
</tr>
<tr>
<td>entailment</td>
<td>entailment</td>
<td>entailment</td>
<td></td>
</tr>
<tr>
<td>troponymy</td>
<td>troponymy</td>
<td>troponymy</td>
<td></td>
</tr>
<tr>
<td>cause</td>
<td>caused/is caused by</td>
<td></td>
<td></td>
</tr>
<tr>
<td>derived form</td>
<td>derived</td>
<td>derived</td>
<td>-</td>
</tr>
<tr>
<td>pertainymy</td>
<td>pertainymy</td>
<td>relatedness</td>
<td>pertainymy</td>
</tr>
<tr>
<td>similar to</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>participle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>see also</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>attribute</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>role</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>has subevent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>in manner of</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>be in state</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fuzzynymy</td>
<td>fuzzynymy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Properties of the semantic relations in Polish WordNet

<table>
<thead>
<tr>
<th>relation</th>
<th>grammatical class</th>
<th>reversibility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>noun</td>
<td>verb</td>
</tr>
<tr>
<td>synonymy</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>hypo-/hypernymy</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>antonymy</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>conversion</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>mero-/holonymy</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>entailment</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>troponymy</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>relatedness</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>derived form</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>fuzzynymy</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
In plWordNet, relations hold between LUs — pairs of lexemes. For example, the adjective *mądry* ‘wise’ is antonymous with *głupi* ‘stupid’, but its synonym *inteligentny* ‘intelligent’ has a different antonym, *nieinteligentny* ‘unintelligent’; *mąż* ‘husband’ is a converse of *żona* ‘wife’, while its synonym *małżonek* ‘spouse’ has the converse *małżonka* ‘spouse’. A derived form has obviously one root.

From EWN, we adopted the *fuzzynymy* relation. It is meant for pairs of lexemes which are clearly connected semantically, but which the lexicographer cannot fit into the existing system of more sharply delineated relations. The practice bore out our decision. We found, even in the basic vocabulary of the core list of lexical units, numerous instances of fuzzynymy (przylądek - morze, ‘cape’ - ‘sea’, pacjent - przychodnia ‘patient’ - ‘walk-in clinic’). Future research includes a review of the fuzzynymy class to see if some subtypes of relations recur; this might be very interesting material for further linguistic investigation.

There is one relation unique to plWordNet: conversion (*narzeczony* - *narzeczona* ‘fiancé’ - ‘fiancée’, *rodzic* - *dziecko* ‘parent’ - ‘child’), *kupić* - *sprzedać* ‘to buy’ - ‘to sell’). Following Apresjan [6, pp. 242-265], we consider such cases to be different than antonymy.

Contrary to our initial expectation, hypo/hypernymy applies not only to nouns and verbs (samochoód - pojazd ‘car’ - ‘vehicle’, biec - poruszać się ‘to run’ - ‘to move’), but also to adjectives (turkusowy - niebieski ‘turquoise’ - ‘blue’). In fact, adjectival hypo/hypernymy has turned out to be relatively widespread, once we allowed the lexicographers to note it.

Neither WN nor EWN support relations that enable an effective rendition of the semantic variation carried by rich morphology and productive derivation. In Polish, we have verb aspect (*pisać* - *napisać* ‘to have written’), reflexivity (*golić* - *golić się* ‘to shave someone - to shave oneself’), subtle derivation via prefixes (*gnić* - *przegnić*, nadgnić, wygnić etc. ‘to rot - to rot through - to become partially rotten - to rot out’), diminutives (*kot* - *kotek*, koteczek, kocio, kotuś, kotunio; *maty* ‘small’ - *malutki*, *maluteńki*, *malusieńki*, *małuśki*), augmentatives (*dziewczyna* ‘girl’ - *dziewucha*, dziewczynisko, dziewuszysko), expressive names (*kobieta* ‘woman’ - *kobiecina* ‘a simple or poor woman’), gender pairs (*malarz* - *malarka* ‘painter’, *masc* - *fem* ‘painter’), names of offspring (*kot* - *kocię* ‘kitten’), names of action (*strzelać* ‘to shoot’ - *strzelanie* ‘shooting’, *strzelanina* ‘fusillade’), names of abstracts (*nienawidzić* ‘to hate’ - *nienawiść* ‘hatred’, *mądry* ‘wise’ - *madrość* ‘wisdom’), names of places (*jeść* ‘to eat’ - *jadalnia* ‘dining room’), names of carriers of attribute (*rudy* ‘red-haired’ - *rudzielec* ‘someone red-haired’), names of agents of action (*palić* ‘smoke’ - *palacz* ‘smoker’), relational adjectives (*uniwersytet* ‘university’ - *universytecki* ‘university (in noun-noun compounds)’).

Analogous phenomena were considered in Czech WordNet [7].

To account for this variety somehow, we decided to extend two relations, *relatedness* and *pertainymy*. In the former, we placed the most regular types of word formation: names of actions, abstract names, pure aspectual pairs (without any other semantic “surplus”, e.g., *pisać* ‘to write’ - *napisać* ‘to have written’,
kupić ‘to have bought’ - kupować ‘to buy habitually or to be buying’), causative verbs (martwić się ‘to worry’ - martwić (kogoś) ‘to worry someone’), relational adjectives and adjectival participles (which we do not consider as verb forms but as separate lexemes). The pertainymy relation accounts for the less regular word forms: names of places, carriers of attributes, agents of actions, offspring, augmentative, expressive and diminutive forms, gender pairs and names of nationalities. The prefixed verbs and “impure” aspectual pairs are captured by trponymy. Although we tried to fit as much as possible into the WN and EWN relation structure, we agree with the Czech WordNet team: it is necessary to go beyond that set of relation if we are to take into consideration the specificity of Slavic languages (Pala and Smrž 2004; (86).

It is perhaps unexpected that the most problematic lexical-semantic relation turned out to be the fundamental one: synonymy. It helped little that this semantic notion is so well explored. There are two approaches to synonymy. One approach defines synonyms as lexemes with the same lexical meaning but with different shades of meaning; the other requires synonyms to be substitutable in some contexts [6, pp. 205-207]. In our opinion, neither approach works well in a semantically motivated network. We sharpened the criterion by positing that synonyms have the same hypernym and the same meronym (if they have any). For example, the lexemes twarz, morda, gęba, ryj, pysk, faćjata, buzia, pyszczeć (all of them mean more or less ‘the face’) can be considered synonymous in a wide sense. There are valid substitutions in some contexts (e.g., dał mu w twarz/mordę/gębę ‘he hit him in the face’; pogłaskała go po twarzy/buzi/pyszcze ‘she stroked his face’). They do not, however, have the same hypernym and meronym: morda is an expressive name of a face, but not a body part. We regard such expressive names as hyponyms of the unmarked lexemes such as ‘face’; there is the same stance in [8]. One of the effects of this decision is that our synsets are very narrow, sometimes even with one element, but the hypo/hypernymy tree is much deeper.

The problem with synonymy definition also arose in Bulgarian WordNet (Ko-eva, Mihev, Tinchov 2004; 62): “In Princeton WordNet the substitution criteria for SYNONYMY is mainly adopted [...] The consequences from such an approach are at least two — not only the exact SYNONYMY is included in the data base (a context is not every context). Second, it is easy to find contexts in which words are interchangeable, but still denoting different concepts (for example hypernyms and hyponyms), and there are many words which have similar meanings and by definition they are synonyms but are hardly interchangeable in any context due to different reasons — syntactic, stylistic, etc. (for example an obsolete and a common word)”.

In our opinion, the vagueness of the synonymy definition and the lack of formal tools of establishing the synonymy of lexemes put in doubt the legitimacy of synonymy as the basic type of relation in lexical-semantics networks. It would appear that all relations link LUs. Suppose that B and D are (near-)synonyms
and B is a hypernym of synonymous A and C; in certain contexts D may be substituted for B, and is also a hypernym of A and C.

The plWordNet project is building the semantic network from scratch; we decided not merely to translate the WN trees (in WordNet 3.0), because that would reflect the structure of English rather than Polish. We did try to translate the higher levels of WN, only to discover a few serious problems. 1) Many lemmas in WN can hardly be considered to denote frequent, basic or most general concepts in Polish; examples include *skin flick* ‘film pornograficzny’, *party favour* ‘pamiątka z przyjęcia’, *butt end* ‘grubszy koniec’, *apple jelly* ‘galaretka jablkowa’. 2) WN glosses are not precise enough to let us find the Polish equivalent, or there may be no lexical Polish equivalent at all (other than calques of English words); examples of untranslatable entries include *changer*, *modifier*, *communicator*, *acquirer*, *banshee*, *custard pie*, *marshmallow*. 3) Translating WN would create nodes in the hypo/hypernymy structure that represent unnecessary or artificial concepts; examples include *emotional person* ‘osoba uczuciowa’, *immune person* ‘osoba uodporniona’, *large person* ‘duży człowiek’, *restor* ‘odpoczywający’, *smiler* ‘uśmiechający się’, *states’ rights* ‘prawa stanowe’.

Our fundamental design decision was corroborated by the experience of the Czech WordNet team [7, pp. 84-85]. The BalkaNet project systematically recorded concepts from other languages (mainly from English, based on WN), not lexicalized in the language at hand. […] The Czech team noticed problems with the translation of equivalents and the corresponding gaps with regard to English. They observed two types of cases where it was not possible to find synonyms (or
even near-synonyms). The Czech synsets had no lexical equivalents in English because of the difference in lexicalizations and conceptualization, or because of the typological differences between those two languages; there are, for example, no phenomena in English as the Czech as verb aspect, reflexive verbs or rich word formation. It is well known that concepts are not universal, nor are they expressed in the same way across languages (this is true even of so basic a notion as colour), although sometimes an ethnocentrism still can be observed — see Wierzbicka’s criticism on that approach [9, pp. 193]. We decided to describe the lexicalization and conceptualization in Polish as accurately as possible. We think that it is much more interesting to compare two semantic networks that reflect the real nature of two natural languages than to create a hybrid, which in fact would be just an English semantic network translated into Polish.

Near the end of year 2 of the plWordNet project, the noun network (the intended vocabulary) is ready. Work must be completed on verbs and adjectives. See Section 3.3 for more details.

3 Tools and resources

3.1 The linguist’s tool

We now discuss software support for the Polish WordNet enterprise: a dedicated editor and algorithms that support lexicographers’ decisions. Two years ago, all available tools – such as [10–12] – required editing the source format, not exactly linguist-friendly. A much more apt editor DEBVisDic [13] was not yet available\(^7\). We therefore chose to design our own WordNet editor, plWNApp, with tight coupling of the envisaged development procedure and the linguistic tasks. [14] present the implementation in some details; here, we focus on its use as a tool.

Linguists edit synsets and relations using plWNApp, which also supports verification and control by coordinators of the project’s linguistic side. Written in Java, so practically fully portable, plWNApp has a client-server architecture with a central database. Clients transparently connect to the database via the Internet, though a version that allows work on a local copy of the database is also maintained. Efficiency, even on low-end computers, was a priority. Network communication is efficient due to caching data exchanged with the database. While it might put screen data out of synch for up to two minutes, this has not happened in 1.5 year of use by a large, distributed group of linguists.

Linguists work via a Graphical User Interface and never edit source files. Every user downloads an appropriate current version of the WordNet from the server. Data are exported and archived in XML, in a special format that we plan to replace with a standard format once we have identified a fitting one. The

---

\(^7\) Early on, our project was also constrained by a commercial connection.
coordinators can edit source files; they did that during the initial assignment of lexical units (LUs) to domains. The coordinators’ stronger tool also supports definition of new lexical-semantic relations, invasive changes in the database and elements of group management. Both versions check on the fly such basic things as the existence of synsets/LUs or the appropriateness of relation instances to be added. More sophisticated diagnostic procedures have been designed, and some already installed.

Core plWordNet will have a complete description of selected LUs, so plWNApp distinguishes system LUs and user LUs. Only coordinators can add the former; other linguists introduce user LUs to complete synsets under construction.

Our linguistic assumptions suggested support for three main tasks:

1. construct an initial, broad synset for a given system LU;
2. correct and divide initial synsets into more cohesive, almost always smaller synsets;
3. link synsets by lexical-semantic relations.

To support these tasks, plWNAPP’s user interface features two perspectives: the LU perspective (2) and the synset perspective (3). The former is organised around selecting a LU and defining synsets and LU relations for it. A linguist would traverse the list of system LUs in the domain assigned to her and, for each LU, define all synsets to which it belongs. System LUs thus serve as starting points in synset construction.

The intended result of task 1 was to group LUs in broad sets of near-synonyms, but pairs of synset often overlapped because of lack of precision in the grouping criteria. In order to support coordinators in task 2, we added the comparison perspective, showing two lists of synsets that share at least \( k \) LUs. Coordinators can edit or merge synsets, or move LUs around. We soon discovered, however, that correction – task 2 – is only possible when done together with task 3, supported by the synset perspective. According to the definition of synsets and synset relations, a LU can participate in a synset only because of what we know about this synset’s relations. In the comparison perspective, synsets are isolated from the structure of synset relations, and coordinators find it very hard to determine the correctness of the overlap between two synsets. In the next version of plWNApp, we will enhance this perspective to a comparison of structures of synset relations around two synsets.

In the synset perspective, each user interaction was to begin by the selection of a source synset which either must be corrected or is chosen as the starting node of a relation instance. Next, the user was to divide the source synset into two or to select a target synset, and then to pick a relation between the two (hypo/hypernymy when dividing the source synset). The added relation instances appear in a table at the bottom of the synset perspective. Predictably, practice diverged significantly from the initial ideas. The relation table was used
most often, gradually becoming the central point of the synset perspective. Extracting a hypo/hypernym synset directly from the source synset was a very rare operation. Linguists preferred to create a new hypo/hypernym synset and move some LUs from the source synset, one by one. It may be easier to decide on one LU than on a group. In any event, the synset perspective is the basic tool in transforming the initial synsets into the deepened hierarchy of narrow synsets, in keeping with our fundamental assumptions. Also, the table shows only relations of the selected source synsets, so linguists suggested extending the table to a graph view. We plan to introduce the possibility of editing synsets and synset relations in combination with the enhanced comparison perspective.

Early on, we found that consistency among linguists was a concern. In order to increase consistency, we introduced substitution tests. For each relation in plWordNet – for synsets and for LUs – there is a morphologically generic test with slots for LUs from the linked synsets or for the linked LUs. (Coordinators can edit definitions.) Slots are filled with the appropriate morphological forms.
Fig. 3. The Synset perspective

Whenever a relation instance is to be added, plWNApp generates a test instance and shows it to the linguist.

The tool associates domains not only with LUs but also with synsets. A LU is assigned to some domains when added it is to the database. The domain of a synset is that of its first LU, usually the system LU that started this synset. Domains offer a simple, but useful way dividing work among linguists. It is the coordinators’ task to merge domain subsets. This is not trouble-free: occasionally, two linguists working on two close domains created a similar, overlapping structure of synsets and synset relation. An enhanced comparison perspective should help adjust such overlaps.
3.2 Toward automation

Work on extending plWNApp to support semi-automatic WordNet construction is under way. We will build software tools that:

– offer better corpus-browsing capability,
– criticize existing WordNet content,
– suggest possible instances of relations.

The browsing tools are based on the statistical analysis of a large corpus in search for distributional associations of LUs. One can identify potential collocations and extract a semantic similarity function (SSF), which for a pair of LUs returns a real-valued measure of their similarity. As our examples showed, the real LUs are the minority among the extracted collocations, and it would be very hard to add new multiword LUs automatically on the basis of a collocation list. A linguist, however, can easily spot possible new multiword LUs if shown a candidate list.

SSFs are based on Harris’s Distributional Hypothesis [15], aptly summarized in [16]: ‘The distributional hypothesis is usually motivated by referring to the distributional methodology developed by Zellig Harris (1909-1992). (...) Harris’ idea was that the members of the basic classes of these entities behave distributionally similarly, and therefore can be grouped according to their distributional behavior. As an example, if we discover that two linguistic entities, w1 and w2, tend to have similar distributional properties, for example that they occur with the same other entity w3, then we may posit the explanandum that w1 and w2 belong to the same linguistic class. Harris believed that it is possible to typologize the whole of language with respect to distributional behavior, and that such distributional accounts of linguistic phenomena are “complete without intrusion of other features such as history or meaning.”’

Many methods of SSF construction have been proposed. The serious problem is their comparison. A SSF produces real values. Manual inspection of even several real numbers is very hard on people. While all known SSF algorithms produce interesting result, how do we choose a SSF that distinguishes really similar LUs (synonyms or close hypo/hypernym) from other groupings? Core plWordNet, constructed manually, can serve as the basis for evaluation. Following [17], we evaluate a SSF by applying it in solving a version of WordNet-Based Synonymy Test (WBST; see also [18]): given a word and four candidates, separate the actual synonym from distractors. The test is automatically generated from plWordNet; for evaluation, different SSFs were extracted from the IPI PAN corpus8 [5] for the same set of LUs.

---

8 The IPI PAN Corpus contains about 254 millions of tokens and is rather unbalanced: most of the text in the corpus comes from newspapers, transcripts of parliamentary sessions and legal text, however also includes artistic prose and scientific texts.
We tested several versions of SSF, achieving the best result of 90.92% in WBST generated from plWordNet for a SSF based on the Rank Weight Function (RWF) [19]: on the basis of SSF\textsubscript{RWF} we can distinguish a synonym from three randomly selected words in some 90% cases. However, in a more difficult version of the WBST, called Extended WBST [18], in which decoys are chosen from LUs similar to the answer, the application of the same SSF\textsubscript{RWF} gave the accuracy of 53.52%. Though the ability of SSF\textsubscript{RWF} to distinguish among semantically related LUs is limited, it was added to plWNApp as a browsing tool. SSF\textsubscript{RWF} is used to produce lists of LUs k most similar to the given one. Such a list can help linguists look among the top positions in the list for possibly omitted synonyms and hypo/hypernyms.

SSF\textsubscript{RWF} is loosely correlated with similarity functions based on plWordNet but it is hard to find any threshold above which the similarity value guarantees the existence of the synonymy or hypo/hypernymy relation. In an experiment, we chose a value 0.2 as a threshold (on the basis of manual inspection). Next, one of the authors manually assessed a statistically significant sample of LU pairs with the similarity above the threshold, according to the synset relations: synonymy, hypo/hypernymy, meronymy and holonymy. Half of the pairs did not express any of these relations. The other half appeared to be worth browsing. In 7% of cases we found two synonyms already present in plWordNet, but only 1% of new synonym pairs. 20% of pairs were close hypo/hypernyms (not necessarily direct) already present in plWordNet, and 16% of new close hypo/hypernyms and co-hyponyms were discovered. 1% of known meronyms and holonyms were found and 5% of new ones were discovered.

SSFs are intended to extract more rather than fewer semantic relations between LUs. We will reintroduce restrictions by way of clustering of the results of SSF – constructing proto-synsets. We also want to apply statistical lexico-syntactic patterns – for example, in the style of [20] – to a large corpus, in order to extract candidate instances of plWordNet relations. The extracted instances will be used to combine the cluster resulting from grouping LUs into a network of synset relations. The results of automatic extraction will be always anchored to plWordNet, because we want to extend it gradually, at each step adding a small set of new LUs automatically suggested for inclusion. After each iteration of automatic acquisition, linguists will be asked to verify and correct the proposed proto-synsets and instances of relations. The proposals will be clearly marked in plWNApp.

3.3 The current state of the system

At the time of this writing, plWordNet contains 12483 LUs grouped in 8095 synsets, 6059 synset relations and 5379 LU relations. Table 3 shows more detailed facts. While we feel that the number of LUs is more important than the number
of synsets (Section 2), Table 3 separates relations between synsets and LUs — the former hold for every LU in a synset.

<table>
<thead>
<tr>
<th>LUs</th>
<th>LU relations</th>
<th>synset relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>8307</td>
<td>antonymy 1952</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hyper/hypernymy 4293</td>
</tr>
<tr>
<td>verbs</td>
<td>3317</td>
<td>converse 47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>holonymy 919</td>
</tr>
<tr>
<td>adjectives</td>
<td>3053</td>
<td>relatedness 1534</td>
</tr>
<tr>
<td></td>
<td></td>
<td>meronymy 847</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pertainymy 1175</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fuzzynymy 671</td>
</tr>
<tr>
<td>all</td>
<td>14677</td>
<td>all 5379</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 6059</td>
</tr>
</tbody>
</table>

Table 3. plWordNet in numbers, September 2007

The average rate of polysemy is 1.46 (calculated as the average number of synsets including the given homonymous LU, as in [1]), and the average size of a synset is 2.04 LUs. The detailed data appear in Tables 4 and 5.

<table>
<thead>
<tr>
<th>Synsets to which a homonymous LU belongs</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9 ≥ 10</th>
<th>Avg</th>
<th>WN</th>
</tr>
</thead>
<tbody>
<tr>
<td>All LUs [%]</td>
<td>73.45</td>
<td>16.10</td>
<td>5.98</td>
<td>2.32</td>
<td>1.02</td>
<td>0.58</td>
<td>0.30</td>
<td>0.09</td>
<td>0.05</td>
<td>0.12</td>
<td>1.47</td>
</tr>
<tr>
<td>Nouns LUs [%]</td>
<td>74.11</td>
<td>16.35</td>
<td>6.00</td>
<td>2.19</td>
<td>0.77</td>
<td>0.38</td>
<td>0.16</td>
<td>0.03</td>
<td>–</td>
<td>–</td>
<td>1.41</td>
</tr>
<tr>
<td>Verbs LUs [%]</td>
<td>79.40</td>
<td>14.73</td>
<td>4.04</td>
<td>1.34</td>
<td>0.28</td>
<td>0.18</td>
<td>0.04</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1.29</td>
</tr>
<tr>
<td>Adj. LUs [%]</td>
<td>64.61</td>
<td>17.10</td>
<td>8.23</td>
<td>3.84</td>
<td>2.53</td>
<td>1.56</td>
<td>0.97</td>
<td>0.34</td>
<td>0.25</td>
<td>0.55</td>
<td>1.79</td>
</tr>
</tbody>
</table>

Table 4. The level of LU polysemy in plWordNet, September 2007 (WN means the Princeton WordNet 3.0)

4 Observations and future work

Our work to date has taught us a few valuable lessons. Of much use, though less interest, is what we found about facilitating the linguists’ task. An important observation concerns the starting point of any properly conceived WordNet: it must be corpus-based. The core vocabulary should consist of words that are frequent in real-life text. We have learnt that, for that particular purpose, certain balance in the corpus is extremely important. In our case, a little too many formal text resulted in a shortage of everyday vocabulary, such as names of the edible plants and food in general, animals and so on, in exchange for a higher than average number of economic and legal terms.
Experiments with translating the Princeton WordNet indiscriminately clearly show that only the top levels of the hierarchy may carry over to other languages intact; it probably should work, because this hierarchy may well be universal. We must work out the lower level afresh, if we want a WordNet that represents the lexical system, or at least much of the lexical system, of the language at hand — see Section 2.

Last but not least, we feel that for a WordNet to cover as much vocabulary of a given language as possible, it would need its own set of relations — many of them derivational in nature. This, however, would make it hard to use WordNets for multilingual NLP tasks, a most likely “killer app” of the near future. In the end, then, one ought to keep balance between too few but rather universal relations (such as antonymy or hypernymy) and too many, too detailed language-specific derivational relations. We believe that any criterion for choosing a useful set relations should consider the feasibility of future NLP tasks and linguistic credibility.

On the computing side of the plWordNet project, we see further fine-tuning of semantic similarity functions as a major task for the near future. Although the results thus far are very promising, too much noise can be observed in the data (about 50% — see Section 3.2). One cannot keep naive thresholds as a means of constraining the output of SSFs. We must first of all take a look at multi-word expressions. We have already developed language-specific methods of extracting Polish multi-word expressions from a corpus, [21] but more work is necessary. We need to build more natural groupings of words based on SSFs. One approach that we will try is to use fuzzy clustering algorithms. The preliminary results are again promising. On the other hand, pattern-based methods are very accurate and have been widely used to extract relations for WordNet; an early example is [22]). We will try to combine pattern-based methods with clustering. One way to accomplish this is to do machine learning of patterns on the basis of statistical and cluster information provided by a SSF; it should at least be useful in disambiguating lexico-semantic relations from the output of a SSF, but it also might help build the WordNet up in a weakly supervised manner.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9≥10</th>
</tr>
</thead>
<tbody>
<tr>
<td>All synsets [%]</td>
<td>10.00</td>
<td>25.00</td>
<td>15.87</td>
<td>7.66</td>
<td>2.77</td>
<td>1.05</td>
<td>0.53</td>
<td>0.20</td>
<td>0.17</td>
</tr>
<tr>
<td>Noun synsets [%]</td>
<td>65.93</td>
<td>19.45</td>
<td>7.92</td>
<td>3.83</td>
<td>1.38</td>
<td>0.63</td>
<td>0.36</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>Verb synsets [%]</td>
<td>1.74</td>
<td>47.07</td>
<td>28.76</td>
<td>12.28</td>
<td>6.10</td>
<td>2.30</td>
<td>0.87</td>
<td>0.32</td>
<td>0.24</td>
</tr>
<tr>
<td>Adj. synsets [%]</td>
<td>15.69</td>
<td>26.17</td>
<td>33.04</td>
<td>17.29</td>
<td>4.87</td>
<td>1.47</td>
<td>0.87</td>
<td>0.33</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table 5. The number of LUs per synset in plWordNet.
References


Exploring and Navigating: Tools for GermaNet

Marc Finthammer and Irene Cramer
Faculty of Cultural Studies, University of Dortmund, Germany
marc.finthammer|irene.cramer@udo.edu

Abstract. GermaNet is regarded to be a valuable resource for German NLP applications, corpus research, and teaching. This demo presents three GUI-based tools meant to facilitate the exploration of and navigation through it.

1 Motivation

GermaNet [1], the German equivalent of WordNet [2], represents a valuable lexical-semantic resource for numerous German natural language processing (NLP) applications. However, in contrast to WordNet, only a few graphical user interface (GUI) based tools have been created up to now for the exploration of GermaNet. In principle, in order to get an idea of it, the user is left on his own with a bunch of XML-files and insufficient means for navigation or exploration\(^1\). Various sub-tasks of our research in the DFG (German Research Foundation) funded project HyTex\(^2\), such as lexical chaining, highly rely on the semantic knowledge represented in GermaNet. While intensively working with it, we accumulated a list of properties a GermaNet GUI should feature and accordingly implemented the GermaNet Explorer. In addition, during the course of our research on lexical chaining for German corpora [3], we also investigated semantic relatedness and similarity measures based on GermaNet as a resource. The results of this work led us to the implementation of eight GermaNet-based relatedness measures, which we provide as Java\(^\text{TM}\) API, the so-called GermaNet-Measure-API. In order to facilitate the use, we also developed a GUI for this API, the so-called GermaNet Pathfinder. We think that these three tools simplify the use of and work with GermaNet: they can be integrated into various NLP applications and can also be used as a resource for the visual exploration of and navigation through GermaNet. All three tools are freely available for download.

\(^1\) As a matter of course, the NLP community working with German data is much smaller than the one working with English data, consequently, the development of tools for German resources, such as GermaNet, takes more time.

\(^2\) The HyTex project aims at the development of text-to-hypertext conversion strategies based on text-grammatical features. Please, refer to our project web pages http://www.hytex.info/ for more information about our work.
2 GermaNet Explorer

Many researchers working with GermaNet have the same experience: they lose their way in the rich, complex structure of its XML-representation. In order to solve this problem, we implemented the GermaNet Explorer, of which a screenshot is shown in Figure 1. Its most important features are: the word sense retrieval function (Figure 1, region 1) and the structured presentation of all semantic relations pointing to/from the synonym set (synset) containing the currently selected word sense (Figure 1, region 2).

Fig. 1. Screenshot GermaNet Explorer

In addition, the GermaNet Explorer offers a visual, graph-based navigation function. A synset (in Figure 2 [Rasen, Grünfläche] Engl. lawn) is displayed in the center of a navigation graph surrounded by its direct semantically related synsets, such as hypernyms (in Figure 2 [Nutzfläche, Grünland]) above the current synset, hyponyms (in Figure 2 [Kunstrasen, Kunststoffrasen] and [Grüngürtel]) below, holonyms (in Figure 2 [Grünanlage, Gartenanlage, Eremitage]) to the left, and meronyms (in Figure 2 [Graspflanze, Gras]) to the right. In order to navigate the graph representation of GermaNet, one simply clicks on a related synset, in other words one of the rectangles surrounding the current synset shown in Figure 2. Subsequently, the visualization is refreshed: the selected synset moves into the center of the displayed graph and the semantically related synsets are updated accordingly.
Fig. 2. Screenshot GermaNet Explorer – Visual Graph Representation

Fig. 3. Screenshot GermaNet Explorer – Representation of the List of All GermaNet Synsets
In addition, the GermaNet Explorer features a representation of all synsets, which is illustrated in Figure 3, region 1. It also provides retrieval, filter, and sort functions (Figure 3, region 2). Further, the GermaNet Explorer exhibits the same function as shown in Figure 3 and a similar GUI for the list of all word senses. We found that these functions, both for the word senses and the synsets, provide a very detailed insight into the modeling and structure of GermaNet and thus helped us to understand its strengths and weak points.

We were already able to successfully utilize the GermaNet Explorer in various areas of our research and teaching. E.g. in experiments on the manual annotation of lexical chains in German corpora, our subjects used the GermaNet Explorer to find paths representing semantic relatedness between two words. This work is partially described in [4]. We also found it helpful for the visualization of lexical semantic concepts and thus for the training of our students in courses on e.g. semantics. We hence argue that the GermaNet Explorer represents a tool which is applicable in many scenarios.

3 GermaNet Pathfinder and Measure-API

Semantic relatedness measures express how much two words have to do with each other. This represents an essential information in various NLP applications and is extensively discussed in the literature e.g. [5]. Many measures have already been investigated and implemented for the English WordNet, however, there are only few publications addressing measures based on GermaNet e.g. [6] as well as [3]. The calculation of semantic relatedness is a subtask of our research in HyTex; we therefore implemented eight GermaNet and three GoogleTM based measures. Because of the–compared to WordNet–different structure of GermaNet, it was necessary to re-implement and adapt algorithms discussed in the literature and in parts already available for WordNet. The GermaNet-Measure-API is implemented as a JavaTM class library and consists of a hierarchically organized collection of measure classes, which provide methods to perform operations such as the calculation of specific relatedness values between words and synsets or the automated distance-to-relatedness conversion. In order to additionally facilitate the integration of these measures into user-defined applications and to allow the straightforward comparison and evaluation of the different measures, we also implemented a GUI, the GermaNet Pathfinder, shown in Figure 4. The most important features of these tools are: the calculation of the semantic relatedness between two words (or two synsets) with various adjustable parameter settings (Figure 4, region 1), the easy-to-apply JavaTM interface, which ensures the simple and fast integration of all measures into any application, and the visualization of the calculated relatedness.

3 For more information about the measures implemented as well as our research on lexical/thermantic chaining and the performance of our GermaNet based lexical chainer, please refer to [3] in this volume.

4 The three GoogleTM measures are based on co-occurrence counts and realize different algorithms to convert these counts into values representing semantic relatedness.
Marc Finthammer and Irene Cramer

Fig. 4. Screenshot GermaNet Pathfinder

as a path in GermaNet, which is shown in Figure 5. For a given word pair all possible readings—in other words all synsets—are considered to calculate the relatedness (or paths) with respect to GermaNet (Figure 4, region 2).

We already successfully used the GermaNet-Measure-API in our lexical chainer, called GLexi. We also found the GermaNet Pathfinder very helpful to explore GermaNet and retrace semantically motivated paths, which is illustrated in Figure 5 as the (shortest) path between Blume (Engl. flower) and Baum (Engl. tree). This path consists of three steps (hypernymy – hyponymy – hyponymy) and traverses two synsets; it thus represents the kind of (indirect) semantic relation relevant in e.g. lexical chaining.

4 Open Issues and Future Work

We have already used the GermaNet Explorer, GermaNet-Measure-API and Pathfinder in our research on thematic chaining, as a tool for the manual annotation of lexical chains and as a resource in our seminars. In our future work, we plan to further explore the possible fields of application, e.g. for training students and annotators. The research on relatedness measures both for GermaNet and WordNet among others [5] shows that the established algorithms are not yet able to satisfactorily represent the semantic relations between two words. Namely, human-judgement experiments stress that the correlation between the relatedness measures and the intuition of subjects is much too low. We therefore plan to investigate alternative relatedness measures, which we also
intend to integrate into the GermaNet Pathfinder. However, the usefulness of the GermaNet Explorer and Pathfinder is constrained by the coverage and modeling quality of the underlying semantic lexicon. Therefore, we also hope to hereby provide tools to see behind GermaNet’s curtain and to thus facilitate the user-centered work with this interesting and valuable resource.

![Screenshot GermaNet Pathfinder – Illustration of a Shortest Path Between Blume and Baum](image)

**Fig. 5.** Screenshot GermaNet Pathfinder – Illustration of a Shortest Path Between Blume and Baum

**References**

Using Multilingual Resources for Building SloWNet Faster

Darja Fišer
Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000, Ljubljana, Slovenia
darja.fiser@guest.arnes.si

Abstract. This project report presents the results of an approach in which synsets for Slovene WordNet were induced automatically from parallel corpora and already existing WordNets. First, multilingual lexicons were obtained from word-aligned corpora and compared to the WordNets in various languages in order to disambiguate lexicon entries. Then appropriate synset ids were attached to Slovene entries from the lexicon. In the end, Slovene lexicon entries sharing the same synset id were organized into a synset. The results were evaluated against a goldstandard and checked by hand.

Keywords: multilingual lexica, parallel corpora, word senses, word-alignment.

1 Introduction

Automated approaches for WordNet construction, extension and enrichment all aim to facilitate faster, cheaper and easier development. But they vary according to the resources that are available for a particular language. These range from Princeton WordNet (PWN) [7], the backbone of a number of WordNets [16, 14], to machine readable bilingual and monolingual dictionaries which are used to disambiguate and structure the lexicon [11], and taxonomies and ontologies that usually provide a more detailed and formalized description a domain [6].

For the construction of Slovene WordNet we have leveraged the resources at our disposal, which are mainly corpora. Based on the assumption that the translation relation is a plausible source of semantics we have used multilingual parallel corpora to extract semantically relevant information. The idea that senses of ambiguous words in SL are often translated into distinct words in TL and that all SL words that are translated into the same TL word share some element of meaning has already been explored by e.g. [13] and [10]. Our work is also closely related to what was been reported by [1], [3] and [17].

The paper is organized as follows: the methodology used in the experiment is explained in the next section. Sections 3 and 4 present and evaluate the results and the last section gives conclusions and work to be done in the future.
2 Methodology

2.1 Parallel Corpora

The experiment was conducted on two very different corpora, the MultextEast corpus [2] and the JRC-Acquis corpus [14].

The former is relatively small (100,000 words per language) and it only contains a single text, the novel “1984” by George Orwell. Although the corpus contains a single literary text but is written in a plain and contemporary style and contains general vocabulary. But because it had already been sentence-aligned and tagged, as many as five languages could be used (English, Czech, Romanian, Bulgarian and Slovene).

The latter, by contrast, contains EU legislation and is very domain-specific. It is also the biggest parallel corpus of its size in 21 languages (about 10 million words per language). However, the JRC-Acquis is paragraph-aligned with HunAlign [18] but is not tagged, lemmatized, sentence- or wordaligned. This means that the pre-processing stage was a lot more demanding than with the 1984 corpus. We were therefore forced to initially limit the languages involved to English, Czech and Slovene with the aim of extending it to Bulgarian and Romanian as soon as tagging information becomes available for these languages.

The English and Slovene part JRC-Acquis of the corpus were first tokenized, tagged lemmatised with totale [4] while the Czech part was kindly tagged for us with Ajka [14] by the team from the Faculty of Informatics from the Masaryk University in Brno. We included the first 2000 documents from the corpus in the dataset and filtered out all function words.

Both corpora were sentence- and word-aligned with Uplug [15] for which the slowest but best performing ‘advanced setting’ was used. It first creates basic clues for word alignments, then runs GIZA++ [13] with standard settings and aligns words with the existing clues. Alignments with the highest confidence measure are learned and the last two steps are repeated three times. The output of the alignment process is a file containing word links with information on word link certainty between the aligned pair of words and their unique ids.

2.2 Extracting Translations of One-Word Literals

Word-alignments were used to create bilingual lexicons. In order to reduce the noise in the lexicon as much as possible, only 1:1 links between words of the same part of speech were taken into account. All alignments occurring only once were discarded.

In this experiment, synonym identification and sense disambiguation were performed by observing semantic properties of words in several languages. This is why the information from bilingual word-alignments was combined into a multilingual lexicon. The lexicon is based on English lemmas and their word ids, and it contains all their translation variants found in other languages. The obtained multilingual lexicon was then compared to the already existing WordNets in the corresponding languages.
For English, PWN was used while for Czech, Romanian and Bulgarian WordNets from the BalkaNet project [16] were used. There were two reasons for using BalkaNet WordNets: (1) the languages included in the project correspond to the multilingual corpus we had available; and (2) the WordNets were developed in parallel, they cover a common sense inventory and are also aligned to one another as well as to PWN, making the intersection easier.

If a match was found between a lexicon entry and a literal of the same part of speech in the corresponding WordNet, the synset id was remembered for that language. If after examining all the existing WordNets there was an overlap of synset ids across all the languages for the same lexicon entry, it was assumed that the words in question all describe the concept marked with this id. Finally, the concept was extended to the Slovene part of the multilingual lexicon entry and the synset id common to all the languages was assigned to it. All the Slovene words sharing the same synset id were treated as synonyms and were grouped into synsets.

### 2.3 Extracting Translations of Multi-Word Literals

The automatic word-alignment used in this experiment only provides links between individual words, not phrases. However, simply ignoring all the expressions that extend beyond word boundaries would be a serious limitation of the proposed approach, especially because so much energy has been invested in the preparation of the resources. The second part of the experiment is therefore dedicated to harvesting multi-word expressions from parallel corpora.

The starting point was a list of multi-word literals we extracted from PWN. It contains almost 67,000 unique expressions. A great majority of those (almost 61,000) are from nominal synsets. Another interesting observation is that most of the expressions (more than 60,000) appear in only one synset and are therefore are monosemous. Again, most nouns are monosemous (almost 57,000) and there are only about 150 nouns that have more than three senses. The highest number of senses for nouns is 6, much lower than for verbs which can have up to 19 senses. We therefore concluded that sense disambiguation of multi-word expressions will not be a serious problem, and limited the approach only to English and Slovene. Bearing in mind the differences between the two languages, we also assumed that we would not be very successful in finding accurate translations of e.g. phrasal verbs automatically, which is why we decided to first look for two- and three-word nominal expressions only.

First, the Orwell corpus was searched for the nominal multi-word expressions from the list. If an expression was found, the id and part of speech for each constituent word was remembered. This information was then used to look for possible Slovene translations of each constituent word in the file with word alignments. In order to increase the accuracy of the target multi-word expressions, translation candidates had to meet several constraints:
(1) a Det-Noun phrase could only be translated by a single Noun (example: ‘a people’ – ‘narod’);

(2) a Det-Adj phrase could only be translated by a single Noun or by a single Adj_Pl (example: ‘the young’ – ‘mladina’ or ‘mladi’);

(3) an (Adj-)Adj-Noun phrase could only be translated by an (Adj-)Adj-Noun phrase (example: ‘blind spot’ – ‘slepa pega’);

(4) a (Adj-)Noun-Noun phrase could be translated either by an (Adj-)Adj-Noun or by a Noun-Noun_gen phrase (examples: ‘swing door’ – ‘nihajna vrata [Adj-N]’, ‘death rate’ – ‘stopnja umrlih [N-N_gen]’, exceptions: ‘cloth cap’ which is translated into Slovene as ‘pokrivalo iz blaga [a cap made of cloth]’, ‘chestnut tree’ – ‘kostanj’);


Because word-alignment is far from perfect, alignment errors were avoided by checking whether translation candidates actually appear as a phrase in the corresponding sentence in the corpus. If a translation was not found for all the parts of the multi-word expression in the file with alignments, an attempt was made to recover the missing translations by first locating the known translated word in the corpus and then using the above-mentioned criteria to guess the missing word from the context. In the end, the canonical word forms for the successfully translated were extracted from the corpus and all phrases sharing the same synset id were joined into a single synset.

3 Results

3.1 Word-Based Approach

The first version of the Slovene WordNet (SLOWN0) was created by translating Serbian synsets [12] into Slovene with a Serbian-Slovene dictionary [5]. The main disadvantage of that approach was the inadequate disambiguation of polysemous words, therefore requiring extensive manual editing of the results. In the current approach we tried to use multilingual information to improve the disambiguation stage and generate more accurate synsets.

In the experiment with the Orwell corpus, four different settings were tested, each of them using one more language [8]. Table 1 shows the number of nominal one-word synsets generated from the Orwell corpus, depending on the number of languages involved. Recall drops significantly when a new language is added. On the other hand, the average number of literals per synset is not affected.
The same approach was also tested on the JRC-Acquis corpus that is from an entirely different domain and is much larger [9]. It is interesting to observe the change in synset coverage and quality resulting from the different dataset.

However, because the corpus is not annotated with the linguistic information needed in this experiment, we could only implement the approach on English, Czech and Slovene in this setting. Note that although the corpus used was much larger, the number of the generated synsets is only slightly higher. This could be explained by the high degree of repetition and domain-specificity of texts from the dataset.

Table 1. Nominal synsets generated by leveraging existing multi-lingual resources (one-word literals only).

<table>
<thead>
<tr>
<th></th>
<th>SLOWN0</th>
<th>SLOWN1</th>
<th>SLOWN2</th>
<th>SLOWN3</th>
<th>SLOWN4</th>
<th>SLOWNJRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>3,210</td>
<td>2,964</td>
<td>870</td>
<td>671</td>
<td>291</td>
<td>3,528</td>
</tr>
<tr>
<td>max l/s</td>
<td>40</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>avg l/s</td>
<td>4.8</td>
<td>1.4362</td>
<td>1.4</td>
<td>1.4</td>
<td>1.7</td>
<td>2.6</td>
</tr>
</tbody>
</table>

3.2 Phrase-Based Approach

Nominal multi-word literals were extracted from PWN and then translated into Slovene based on word-alignments. In order to avoid the alignment errors some restrictions on the translation patterns were introduced and phrase candidates were checked in the Slovene corpus as well. This simple approach to match phrases in word-aligned parallel corpora yielded more synsets than was initially expected. If it was extended to other patterns, even more multi-word literals could be obtained. Another approach would be to use statistical co-occurrence measures to check the validity of more elusive patterns.

Table 2. Nominal synsets generated from parallel corpora (two-word and three-word literals only).

<table>
<thead>
<tr>
<th></th>
<th>ORWELL</th>
<th>JRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>mwe’s found</td>
<td>163</td>
<td>5,652</td>
</tr>
<tr>
<td>mwe’s translated</td>
<td>121 (73%)</td>
<td>1,984 (34%)</td>
</tr>
<tr>
<td>max l/s</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>avg l/s</td>
<td>1.29</td>
<td>1.13</td>
</tr>
</tbody>
</table>
4 Evaluation

4.1 Synset Quality

Automatic evaluation was performed against a manually created goldstandard. Its literals were compared to literals in the automatically induced WordNets with regard to which synsets they appear in. This information was used to calculate precision, recall and f-measure.

Precision gives the proportion of retrieved and relevant synset ids for a literal to all synset ids for that literal. Recall is the proportion of relevant synset id retrieved for a literal out of all relevant synset ids available for that literal. Finally, precision and recall were combined in the traditional f-measure: \( (2 \times P \times R) / (P + R) \). This seems a fairer alternative to simply evaluating synsets because of the restricted input vocabulary.

![Graph showing precision, recall, and f-measure for nominal synsets according to the number of languages used in the disambiguation stage of automatic synset induction from the Orwell corpus.](image)

**Fig. 1.** A comparison of precision, recall and f-measure for nominal synsets according to the number of languages used in the disambiguation stage of automatic synset induction from the Orwell corpus.

Figure 1 shows the drop in recall and increase in precision and f-measure each time a new language is added to the disambiguation stage, peaking at 77.37% for precision, 75.88% for recall and 76.62% for f-measure. The results for the JRC-Acquis corpus are worse due to fewer languages involved and less accurate word-alignment (precision: 67.0%, recall: 72.0% and f-measure: 69.4%).
4.2 Multi-Word Expressions

Because there was virtually no overlap between the goldstandard and synsets containing automatically translated multi-word expressions, all the synsets obtained from the Orwell corpus was checked by hand. As can be seen in Table 3, about a third of the generated literals were completely wrong.

The errors were analyzed and grouped into categories. Most errors (17 synsets) occurred because an English multi-word expression should be translated into Slovene with a single word (e.g. ‘top hat’ – ‘cilinder’). The next category (12 synsets) contains alignment errors in which one of the constituent words is mistranslated or a translation is missing (e.g. ‘mortality rate’ – ‘umrljivost otrok’, should be ‘stopnja smrtnosti’). In the next category there are 8 expressions that have been translated correctly but can not be included in the synset because the senses of the translation and the original synset are not the same (e.g. ‘white knight’ – ‘beli tekač’ as in chess, should be ‘beli vitez’ as in business takeovers). And finally, there are 12 borderline cases that contain a correct translation but also an error (e.g. ‘black hole’ – ‘črna odprtina[wrong]’ and ‘črna luknja[correct]’).

Table 3. Manual evaluation of multi-word expressions obtained from the Orwell corpus.

<table>
<thead>
<tr>
<th></th>
<th>ORWELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>completely wrong</td>
<td>39 (32%)</td>
</tr>
<tr>
<td>contain some errors</td>
<td>12 (10%)</td>
</tr>
<tr>
<td>fully correct</td>
<td>70 (58%)</td>
</tr>
<tr>
<td>total no. of synsets</td>
<td>121</td>
</tr>
</tbody>
</table>

A larger-scale evaluation of multi-word expressions harvested from the JRC-Acquis has not been carried out but is planned for the near future. A quick overview of the results suggests that the quality of the generated synsets is comparable to the ones obtained from the Orwell corpus.

5 Conclusions

In this paper we have presented an approach to automatically generate WordNet synsets from the two parallel corpora. The method works best on nouns which are disambiguated against several languages. The limitation of the word-alignment based approach was successfully overcome by using the alignment information to form multi-word expressions.

However, the issue of adding multi-word units to WordNet is far from exhausted. More sophisticated statistic-based methods could be used to find more reliable translations of multi-word units. Another possibility to get even more added value from parallel corpora would be an attempt to identify (domain-specific) multi-word expressions that are not part of PWN and add them to Slovene WordNet.
Acknowledgements

I would like to thank Aleš Horák from the Faculty of Informatics, Brno Masaryk University, for POS-tagging and lemmatizing the Czech part of the JRC-Acquis corpus.

References

Abstract. In the presented paper we show how the Global WordNet Grid software is designed. The goal of the Grid is to provide a free network of WordNets linked together through interlingual indexes. We have set as our goal to work on the Grid preparation in the Masaryk University NLP Centre and design its software background. All participating WordNets will be encapsulated by a DEB (Dictionary Editor and Browser) server established for this purpose. The following text presents design details of the new DEBGrid application with possibilities of three types of public and authenticated user access to the Grid WordNet data.

Key words: WordNet; DEB platform; DEBVisDic; Global WordNet Grid

1 Introduction

In June 2000, the Global WordNet Association (GWA [1]) was established by Piek Vossen and Christiane Fellbaum. The purpose of this association is to “provide a platform for discussing, sharing and connecting WordNets for all languages in the world.” One of the most important actions of GWA is the Global WordNet Conference (GWC) that is being held every two years on different places all over the world. The second GWC was organized by the MU NLP Centre in Brno and the NLP Centre members are actively participating in GWA plans and activities. A new idea that was born during the third GWC in Korea is called the Global WordNet Grid with the purpose of providing a free network of smaller (at the beginning) WordNets linked together through ILI. The Grid preparation is currently just starting and the MU NLP Centre is going to secure its software background.

The idea of connecting WordNets has been suggested during the Balkanet project (2001–2004 [2]) in which Patras team developed the core of the WordNet Management System designed to link all the WordNets developed in the course of the project (Deliverable 9.1.04, September 2004).
It was tested successfully on Greek and Czech WordNets. However, the Patras team did not proceed with it and the system remained only as a partial result of the research that was not pursued further. Before the end of Balkanet project, the Czech team decided to re-implement the local version of the VisDic browser and editor using client/server architecture. This was the origin of the DEBVisDic tool that was fully implemented only after finishing Balkanet project. Fully operational version of DEBVisDic was presented at the 3rd Global WordNet Conference 2006 in Korea [3]. In our view this client/server tool will become a software background for the Grid preparation mentioned above (see below in the Section 3.2).

2 The Global WordNet Grid

Since the first publicly available WordNet, the Princeton WordNet [4], more than fifty national WordNets have been developed all over the world. However, the availability of the WordNets is limited – that is why the idea of a completely free Global WordNet Grid has appeared.

It is a known fact that, for instance, the results of the EuroWordNet are not freely accessible though the participants of the project have developed (and are developing) more complete and larger WordNets for the individual languages. Practically the same can be said also about the results of the Balkanet project. If one wants to exploit WordNets for different languages it is always necessary to get in touch with the developers and ask them for the permission to use the WordNet data.

Another reason for building and having the completely free Global WordNet Grid is the fact that the particular WordNets can be linked to the selected ontologies (e.g. Sumo/Milo) and domains. This has already took place with the WordNets developed in the Balkanet project. The links to the ontologies should be provided for all WordNets included in the Global WordNet Grid.

The Grid also provides a common core of 4,689 synsets serving as a shared set of concepts for all the Grid’s languages. These synsets are selected from the EuroWordNet Common Base Concepts used in many WordNet projects.

3 DEBGrid – the DEB Application for the Global WordNet Grid

The DEBGrid application will be built over the DEBVisDic application with the DEB server either set up at the NLP Centre of Masaryk University in Brno or it will be set up by the Global WordNet Association. The DEB platform provides important backgrounds for the WordNet Grid universal features.
3.1 The DEB Architecture

The Dictionary Editor and Browser (DEB) platform [3, 5, 6] has been developed as a general framework for fast development of wide range of dictionary writing applications. The DEB platform provides several very important foundations that are common to most of the intended dictionary systems. These foundational features include:

– a strict separation to the client and server parts in the application design. The server part provides all the necessary data manipulation functions like data storage and retrieval, data indexing and querying, but also various kinds of data presentations using templates. In DEB, the dictionary entries are stored using a common XML format, which allows to design and implement dictionaries and lexicons of all types (monolingual, translational, thesauri, ontologies, encyclopaedias etc.). The client part of the application concentrates on the user interaction with the server part, it does not produce any complicated data manipulation. The client and server parts communicate by means of the standard HTTP (or secured HTTPs) protocol.

– a common administrative interface that allows to manage user accounts including user access rights to particular dictionaries and services, dictionary schema definitions, entry locking administration or entry templates definitions.

– XML database backend for the actual dictionary data storage. Currently, we are working with the Oracle Berkeley DB XML [7, 8] database, which provides a flexible XML database with standard XPath and XQuery interfaces. The DB XML database is well suited for processing complicated XML structures, however, we (and according to private discussions other DB XML users as well) have encountered efficiency problems when processing certain kinds of queries that result in large lists of answers. Simple processing of the data (like export or import of the whole dictionary) is not a problem as the whole English WordNet export (over 100,000 entries) takes less than 1 minute, but searching for values of specific subtags can take several seconds in such large dictionary even when indexes are used. We are currently working on several solutions for this, which include link caching, specific DB XML indexing and also trying a completely different database backend. The key advantage for all the DEB applications is that a replacement of the DB XML backend with another database will be a completely transparent process which does not need any change in the applications themselves.

Based on these common features several developed and widely used dictionary applications have been implemented, including the well-known WordNet editor DEBVisDic that has been used in several national WordNets development recently (Czech, Polish, Hungarian or South African languages). With this evidence, we believe that DEB is the right concept for the Global WordNet Grid data provision.
3.2 The DEBGrid Design and Implementation

In the DEB platform environment, all the WordNets are usually stored on single DEBVisDic server. In the Grid, most of the WordNets will be also stored in this way, however, since the Grid could be finally composed of large number of WordNet dictionaries developed by different organizations, this solution may not be always the best option (for example because of licensing issues). Thanks to the client-server nature of the DEB platform, DEBGrid can offer two possible types of encapsulating WordNets in the server:

- a WordNet can be physically stored on the central server. This is the traditional DEBVisDic setup and offers the best performance.
- a WordNet can be stored on a DEBVisDic server located at the WordNet owner’s institution. All servers in the Grid can then communicate with each other (depending on the server setup). The Central Grid server for this WordNet has only the knowledge of which server to contact, instead of having the full WordNet database stored locally, and all queries are dynamically resolved over the Internet. This option may be slower as it depends on the quality of connection to different servers and their performance. On the other hand, the WordNet owner has full control over the displayed data and access permissions.
- a mixed solution – some WordNets are stored on central server and some are stored on their respective owners’ servers. This is just an extension of the previous option. Again, the performance of the whole Grid depends on the performance of single servers, but the speed can be improved if the most used WordNets are stored on the central server.

The DEB framework provides several possibilities of working with the WordNet data. All the types of the Grid access undergo the same control of service and user management with the option to provide information for public (anonymous) access as well as authenticated access for registered users.

Basically, each WordNet in the Grid can be presented to the Grid users in one of the following forms:

a) by means of a simple purely HTML interface working in any web browser. This interface is able to display one WordNet dictionary or the same synset in several WordNets. Synsets are displayed using XSLT templates – the server can provide several view of the synset data ranging from a terse view up to a detailed view. The view can be even different for each dictionary. An example of such presentation of one synset in three WordNets is displayed in the Figure 1. This type of WordNet view is probably the best for public anonymous access to the Grid, since it does not need any installation of user software or packages.

a) using the full DEBVisDic application. This application needs to be installed as an extension of the freely available Firefox web browsers, but it offers
Fig. 1. The web interface of DEBGrid with three interlinked WordNets.

much complex functionality than the web access. Each WordNet is opened in its own window which offers several views of the WordNet data (a textual preview, hypernym/hyponymic tree structures, user query lists or XML) and also the possibility to edit the data (for users with the write permissions).

With this type of the Grid access, the user would have the most advanced environment for working with the Grid WordNets.

a) by means of a defined interface of the DEBVisDic server. This way any external application may query the server and receive WordNet entries (in XML or other form) for subsequent processing. In this way, local external applications can easily process the Grid data in standard formats.

In all cases, users (or external applications) could authenticate with a login and password over secure HTTP connection. Each user can be given a read-only or read-write access to particular WordNets.

For some applications it is useful to use a visualization tool that allows to view synsets and their links as graphs. Such tool is under development at the MU NLP Centre, it is called Visual Browser [9]. Its important feature is the ability to process WordNet synsets from a DEB server storage and convert them into the RDF notation for visualization. Visual Browser is also suitable for representing ontologies that can and will be integrated within Global WordNet Grid.
4 Conclusions

In this article, we have presented a report of the design and implementation of the Global WordNet Grid software background. The basic idea of the WordNet Grid introduced by P. Vossen, Ch. Fellbaum and A. Pease at GWC 2006 includes establishing of an interlinked network of national WordNets connected by means of the interlingual indexes. In the starting phase the Grid contains only a subset of the EuroWordNet Base Concepts with nearly 5,000 synsets.

The management and intelligent processing of the included WordNets is driven by the DEB development platform tool called DEBGrid. This tool is built on top of the DEBVisDic WordNet editor and allows thus a versatile environment for working with large number of WordNets in one place and style.

Acknowledgements

This work has been partly supported by the Academy of Sciences of Czech Republic under the project T100300419, by the Ministry of Education of CR in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project 201/05/2781.

References

The Development of a Complex-Structured Lexicon based on WordNet

Aleš Horák1, Piek Vossen2, and Adam Rambousek1

1 Faculty of Informatics
Masaryk University
Botanická 68a, 60200 Brno
Czech Republic
{hales,xrambous}@fi.muni.cz
2 Faculteit der Letteren
Vrije Universiteit van Amsterdam
e Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
Piek.Vossen@irion.nl

Abstract. The Cornetto project develops a new complex-structured lexicon for the Dutch language. The lexicon comprises information from two current electronic dictionaries – the Referentie Bestand Nederlands (RBN), which contains FrameNet-like structures, and the Dutch WordNet (DWN) with the usual WordNet structures. The Cornetto lexicon (stored in the Cornetto database) will be linked to English WordNet synsets and have detailed descriptions of lexical items in terms of morphologic, syntactic, combinatoric and semantic information. The database is organized in four data collections – lexical units, synsets, ontology terms and the Cornetto identifiers. The Cornetto identifiers are specifically used for managing the relations between lexical units on the one hand and synsets on the other hand. The mapping is first created automatically, but then revised manually by lexicographers. Special interfaces have been developed to compare the different perspectives of organizing concepts (lexical units versus synsets versus ontology terms).

In this article, we describe the background information about the Cornetto project and the implementation of necessary project tools that are based on the DEBVisDic tool for WordNet editing. The development of the Cornetto clients is a joint project of the Masaryk University in Brno and the University of Amsterdam.

Key words: Cornetto project; WordNet; DEB platform; DEBVisDic

1 Introduction

Cornetto is a two-year Stevin project (STE05039) in which a lexical semantic database is built that combines WordNet with FrameNet-like information [1]
for Dutch. The combination of the two lexical resources will result in a much richer relational database that may improve natural language processing (NLP) technologies, such as word sense-disambiguation, and language-generation systems. In addition to merging the WordNet and FrameNet-like information, the database is also mapped to a formal ontology to provide a more solid semantic backbone.

The database will be filled with data from the Dutch WordNet [2] and the Referentie Bestand Nederlands [3]. The Dutch WordNet (DWN) is similar to the Princeton WordNet for English, and the Referentie Bestand (RBN) includes frame-like information as in FrameNet plus additional information on the combinatoric behaviour of words in a particular meaning.

Both DWN and RBN are semantically based lexical resources. RBN uses a traditional structure of form-meaning pairs, so-called Lexical Units [4]. Lexical Units contain all the necessary linguistic knowledge that is needed to properly use the word in a language. The Synsets are concepts as defined by [5] in a relational model of meaning. Synsets are mainly conceptual units strictly related to the lexicalization pattern of a language. Concepts are defined by lexical semantic relations. For Cornetto, the semantic relations from EuroWordNet are taken as a starting point [2].

Within the project, we try to clarify the relations between Lexical Units and Synsets, and between Synsets and an ontology. DEBVisDic is specifically adapted for this purpose.

In the next section we give a short overview of the structure of the database. The following sections give some background information on DEBVisDic and explain the specific adaptations and clients that have been developed to support the work of mapping the three resources.

2 The Cornetto Lexical Database

The Cornetto database (CDB) consists of 3 main data collections:

- Collection of Lexical Units, mainly derived from the RBN
- Collection of Synsets, mainly derived from DWN
- Collection of Terms and axioms, mainly derived from SUMO and MILO

In addition to the 3 data collections, a separate table of so-called Cornetto Identifiers (CIDs) is provided. These identifiers contain the relations between the lexical units and the synsets in the CDB but also to the original word senses and synsets in the RBN and DWN.

DWN was linked to WordNet 1.5. WordNet domains are mapped to WordNet 1.6 and SUMO is mapped to WordNet 2.0 (and most recently to WordNet 2.1). In order to apply the information from SUMO and WordNet domains
to the synsets, we need to exploit the mapping tables between the different versions of WordNet. We used the tables that have been developed for the MEAN-ING project [6, 7]. For each equivalence relation to WordNet 1.5, we consulted a table to find the corresponding WordNet 1.6 and WordNet 2.0 synsets, and via these we copied the mapped domains and SUMO terms to the Dutch synsets.

The structure for the Dutch synsets thus consists of:

- a list of synonyms
- a list of language internal relations
- a list of equivalence relations to WordNet 1.5 and WordNet 2.0
- a list of domains, taken from WordNet domains
- a list of SUMO mappings, taken from the WordNet 2.0 SUMO mapping

The structure of the lexical units is fully based in the information in the RBN. The specific structure differs for each part of speech. At the highest level it contains:

- orthographic form
- morphology
- syntax
- semantics
- pragmatics
- examples

The above structure is defined for single word lexical units. A separate structure will be defined later in the project for multi-word units. It will take too much space to explain the full structure here. We refer to the Cornetto website [8] for more details.
The Dictionary Editor and Browser (DEB) platform [9, 10] offers a development framework for any dictionary writing system application that needs to store the dictionary entries in the XML format structures. The most important property of the system is the **client-server** nature of all DEB applications. This provides the ability of distributed authoring teams to work fluently on one common data source. The actual development of applications within the DEB platform can be divided into the server part (the server side functionality) and the client part (graphical interfaces with only basic functionality). The server part is built from small parts, called **servlets**, which allow a modular composition of all services. The client applications communicate with servlets using the standard HTTP web protocol.

For the server data storage the current database backend is provided by the Berkeley DB XML [11], which is an open source native XML database providing XPath and XQuery access into a set of document containers.

The user interface, that forms the most important part of a client application, usually consists of a set of flexible forms that dynamically cooperate with the server parts. According to this requirement, DEB has adopted the concepts of the Mozilla Development Platform [12]. Firefox Web browser is one of the many applications created using this platform. The Mozilla Cross Platform Engine provides a clear separation between application logic and definition, presentation and language-specific texts.
3.1 New DEB Features for the Cornetto Project

During the Cornetto project the nature of the Cornetto database structure has imposed the need of several features that were not present in the (still developing) DEB platform. The main new functionalities include:

- **entry locking** for concurrent editing. Editing of entries by distant users was already possible in DEB, however, the exclusivity in writing to the same dictionary item was not controlled by the server. The new functions offer the entry locking per user (called from the client application e.g. when entering the edit form). The list of all server locks is presented in the DEB administration interface allowing to handle the locks either manually or automatically on special events (logout, timeout, loading new entry, ...).

- **link display preview caching.** According to the database design that (correctly) handles all references with entity IDs, each operation, like structure entry preview or edit form display, runs possibly huge numbers (tens or hundreds) of extra database queries displaying text representations instead of the entity ID numbers. The drawback of this compact database model is in slowing down the query response time to seconds for one entry. To overcome this increase of the number of link queries, we have introduced the concept of *preview caching*. With this mechanism the server computes all kinds of previews in the time of saving a modified entry in special entry variables (either XML subtags or XML metadata). In the time of constructing the preview or edit form, the linked textual representations are taken from the preview caches instead of running extra queries to obtain the computed values.

- **edit form functionalities** – the lexicographic experts within the Cornetto project have suggested several new user interface functions that are profitable for other DEB-based projects like collapsing of parts of the edit form, entry merging and splitting functions or new kinds of automatic inter-dictionary queries, so called AutoLookUps.

All this added functionalities are directly applicable in any DEB application like DEBVisDic or DEBDict.

4 The New DEBVisDic Clients

Since one of the basic parts of the Cornetto database is the Dutch WordNet, we have decided to use DEBVisDic as the core for Cornetto client software. We have developed four new modules, described in more details below. All the databases are linked together and also to external resources (Princeton English WordNet and SUMO ontology), thus every possible user action had to be very carefully analyzed and described.

During the several months of active development and extensive communication between Brno and Amsterdam, a lot of new features emerged in both
The Development of a Complex-Structured Lexicon based on WordNet

Fig. 3. Cornetto Identifiers window, showing the edit form with several alternate mappings

The user interface is the same as for all the DEBVisDic modules: upper part of the window is occupied by the query input line and the query result list and the lower part contains several tabs with different views of the selected entry. Searching for entries supports several query types – a basic one is to search for a word or its part, the result list may be limited by adding an exact sense number. For more complex queries users may search for any value of any XML element or attribute, even with a value taken from other dictionaries (the latter is used mainly by the software itself for automatic lookup queries).

The tabs in the lower part of the window are defined per dictionary type, but each dictionary contains at least a preview of an entry and a display of the entry XML structure. The entry preview is generated using XSLT templates, so it is very flexible and offers plenty of possibilities for entry representation.
4.1 Cornetto Lexical Units

The Cornetto foundation is formed by Lexical Units, so let us describe their client package first. Each entry contains complex information about morphology, syntax, semantics and pragmatics, and also lots of examples with complex substructure. Thus one of the important tasks was to design a preview to display everything needed by the lexicographers without the necessity to scroll a lot. The examples were moved to separate tab and only their short resumé stayed on the main preview tab.

Lexical units also contain semantic information from RBN that cannot be published freely because of licensing issues. Thus DEBVisDic here needs to differentiate the preview content based on the actual user’s access rights.

The same ergonomic problem had to be resolved in the edit form. The whole form is divided to smaller groups of related fields (e.g. morphology) and it is possible to hide or display each group separately. By default, only the most important parts are displayed and the rest is hidden.

Another new feature developed for Cornetto is the option to split the edited entry. Basically, this function copies all content of edited entry to a new one. This way, users may easily create two lexical units that differ only in some selected details.

Because of the links between all the data collections, every change in lexical units has to be propagated to Cornetto Synsets and Identifiers. For example, when deleting a lexical unit, the corresponding synonym has to be deleted from the synset dictionary.

4.2 Cornetto Synsets

Synsets are even more complex than lexical units, because they contain lots of links to different sources – links to lexical units, relations to other synsets, equivalence links to Princeton English WordNet, and links to the ontology.

Again, designing the user-friendly preview containing all the information was very important. Even here, we had to split the preview to two tabs – the first with the synonyms, domains, ontology, definition and short representation of internal relations, and the second with full information on each relation (both internal and external to English Wordnet). Each link in the preview is clickable and displays the selected entry in the corresponding dictionary window (for example, clicking on a synonym opens a lexical unit preview in the lexical unit window).

The synset window offers also a tree view representing a hypernym/hyponym tree. Since the hypero/hyponymic hierarchy in WordNet forms not a simple tree but a directed graph, another tab provides the reversed tree displaying links in the opposite direction (this concept was introduced in the VisDic WordNet editor). The tree view also contains information about each subtree’s significance – like the number of direct hyponyms or the number of all the descendant synsets.
4.3 Cornetto Identifiers

The lexical units and synsets are linked together using the Cornetto Identifiers (CID). For each lexical unit, the automatic aligning software produced several mappings to different synsets (with different score values). At the very beginning, the most probable one was marked as the “selected” mapping.

In the course of work, users have several ways for confirming the automatic choice, choosing from other offered mapping, or creating an entirely new link. For example, a user can remove the incorrect synonym from a synset and the corresponding mapping will be marked as unselected in CID. Another option is to select one of the alternate mappings in the Cornetto Identifiers edit form. Of course, this action leads to an automatic update of synonyms.

The most convenient way to confirm or create links is to use Map current LU to current Synset function. This action can be run from any Cornetto client package, either by a keyboard shortcut or by clicking on the button. All the required changes are checked and carried out on the server, so the client software does not need to worry about the actual actions necessary to link the lexical unit and the synset.

4.4 Cornetto Ontology

The Cornetto Ontology is based on SUMO and so is the client package. The ontology is used in synsets, as can be seen in the Figure 2. The synset preview shows a list of ontology relations triplets – relation type, variable and variable or ontology term.

Clicking on the ontology term opens the term preview. A user can also browse the tree representing the ontology structure.

5 Conclusions

We have just presented the design and implementation of new tools for supporting the work on the Dutch Cornetto project developing a new complex structure lexicon. The tools are prepared on top of the DEB platform, which currently covers in six full featured dictionary writing systems (DEBDict, DEBVisDic, PRALED, DEB CPA, DEB TEDI and Cornetto). The Cornetto tools are closely related to the DEBVisDic system which, within the Cornetto project, has shown the versatility of its design as well as has been supplemented with new features reusable not only for work with other national WordNets but also for any other DEB application.
Acknowledgments

The Cornetto project is funded by the Nederlandse Taalunie and STEVIN. This work has also partly been supported by the Ministry of Education of the Czech Republic within the Center of basic research LC536 and in the Czech National Research Programme II project 2C06009.

References

1 Introduction

Chinese and Japanese are two typologically different languages sharing the same orthography since they both use Chinese characters in written text. What makes this sharing of orthography unique among languages in the world is that Chinese characters (kanji in Japanese and hanzi in Chinese) explicitly encode information of semantic classification [1,2]. This partially explains the process of Japanese adopting Chinese orthography even though the two languages are not related. The adaptation is supposed to be based on meaning and not on cognates sharing some linguistic forms. However, this meaning-based view of kanji/hanzi orthography faces a great challenge given the fact that Japanese and Chinese form-meaning pair do not have strict one-to-one mapping. There are meanings instantiated with different forms, as well as same forms representing different meanings. The character 湯 is one of most famous faux amis. It stands for ‘hot soup’ in Chinese and ‘hot spring’ in Japanese. In sum, these are two languages where their forms are supposed to be organized according to meanings, but show inconsistencies.

WordNets as lexical knowledgebases, on the other hand, assume a basic semantic taxonomy which can be universally represented regardless of the linguistic distance. In other words, they assume that the organization of words around synsets and lexical semantic relations are universal. This position is partially supported by the various languages with comprehensive WordNets.

It is important to note that WordNet and the Chinese character orthography is not as different as they appear. WordNet assumes that there are some generalizations in how concepts are clustered and lexically organized in languages and propose an explicit lexical level representation framework which can be applied to all languages in the world. Chinese character orthography intuited that there are some conceptual bases for how meaning are lexical realized and organized, hence devised a sub-lexical level representation to represent semantic clusters. Based on this observation, the study of cross-lingual homo-forms between Japanese and Chinese in the context of WordNet offers a unique window for different approaches to lexical conceptualization. Since Japanese and Chinese use the same character set with the
same semantic primitives (i.e. radicals), we can compare their conceptual system with the same atoms when there are variations in meanings of the same word-forms. When this is overlaid over WordNet, we get to compare the ontology of the two representation systems.

From a more practical point of view, unified lexical resources are necessary in advanced multilingual knowledge processing. Princeton WordNet [3] is a lexical resource commonly used. The Chinese WordNet (CWN) [4] is created already, but there is no Japanese WordNet available yet. Since the Japanese and the Chinese writing system (Hanzi) and its semantic meanings are near-related, analyzing such relation may speed up the creation of the Japanese WordNet that aligned with CWN by providing statistical information of Form-Meaning mapping of Japanese and Chinese word. In this paper, we examine and analyze the form of Hanzi and the semantic relations between the CWN and the Japanese Electronic Dictionary Research [5].

2 Literature Review

WordNet-like lexical knowledgebases for Chinese include HowNet, Chinese Concept Dictionary (CCD) [6], and Chinese WordNet [4, 7]. However, these are all constructed at the word level and do not explicitly refer to characters or character composition. Wong and Pala [8] was probably the first work linking the semantic radicals of Chinese characters to a linguistic ontology, EWN top ontology in their work. The first full-scale Lexical KnowledgeBase work based on Chinese characters and semantic radicals are two recent doctoral dissertations: Chou [9] and Hsieh [10]. Chou’s Hantology maps Chinese character radicals to SUMO ontology and build an ontology-based representation of character changes and variations. Hsieh’s HanziNet is a WordNet like knowledgebase taking Chinese characters as basic units and utilizes the semantic information form the semantic radicals. Their work have been converged and integrated in our recent proposal to utilize Chinese characters to build multilingual knowledge infrastructure [10].

For Japanese, the National Institute of Communication Technology (NICT) has recently started the first project to construct a Japanese WordNet. There is also a long tradition of working on Kanji, especially in terms of font rendition and character and word dictionaries (for instance, by the CJK Dictionary Institute, www.cjk.org). Unfortunately, we are not aware of any systematic work in Japan linking kanji with WordNet like lexical knowledgebases.

3 Resources: CWN, EDR and List of Character Variants

In order to do a character-based and sense-anchored comparison of Chinese and Japanese words, we employed three important resources: CWN, EDR, and a mapping table between Chinese and Japanese characters.
EDR

The EDR Electronic Dictionary is a machine-tractable dictionary that contains the lexical knowledge of Japanese and English constructed by the Japanese Electronic Dictionary Research Institute [5]. It contains list of 325,454 Japanese words (jwd) and their descriptions. In this study, the English translation, the English definition and the Part-of-Speech category (POS) of each jwd are used to determine their senses and semantic relations to their Chinese counterparts.

CWN

The Chinese WordNet currently contains list of 8,624 Chinese words (cwd) and their descriptions. In this experiment, the English translations, the English definition, the Part-of-Speech category (POS) and the corresponding synset of all senses of each cwd are used to determine the semantic relations. These high and mid-frequency words represent over 20,000 synsets. Since CWN is still in progress and contains only words whose senses are manually analyzed and confirmed by corpus data, we can supplement the synsets not covered by CWN with data from the translation-based Sinica BOW (Academia Sinica Bilingual Ontological WordNet http://bow.sinica.edu.tw).

List of Hanzi Variants

Modern kanji and hanzi systems are both descendent of Chinese character systems of the Tang dynasty over 1,200 years ago. However, even though frequent contacts were maintained, the long periods of developments as separate systems still result in a small set of glyph variants. For instance, a character with the basic meaning of ‘elder sister’ are represented by two graph variants in Chinese hanzi and Japanese kanji, as shown in (1).

(1) Example Character Variants in Chinese and Japanese

```xml
<c type="tw" ref="U+59CA">姊</c>
<c type="jp" ref="U+59C9">姉</c>
```

It is important to note that as these glyph variants cannot be dealt with simply as font variants. First, they are highly conventionalized can have different meaning in the context of each language. Second, they are given different codes in Unicode coding space based on the conventionality arguments. Hence, in terms of automatic searching and comparison, these pairs will be recognized as different characters unless stipulated otherwise. In our study, we use a list of 125 pairs of Japanese and Chinese character variants compiled and provided to us by Christian Wittern of Kyoto University’s Institute for Studies in Humanities.

---

1 http://www2.nict.go.jp/rr/312/EDR/index.html
4 Methodology and Procedure

4.1 EditorCharacter-based Word Mapping Between Chinese and Japanese

Each Japanese word jwd and Chinese word cwd is analyzed as a string of characters $c_1...c_n$. Each jwd is compared with all cwd’s for their character-string similarity. Each matched <cwd, jwd> pair must satisfy one of the three criteria and are classified as such. It is important to note that with the help of the list of variant characters, we are able to establish character identity regardless of their different surface glyphs.

(I) Identical Character Sequence Pairs, where the numbers of characters in the jwd and cwd are identical and all the corresponding $n$th characters in the two words are also identical. We call these pairs homographic pairs. They can be exemplified by 'head', and 'singer'.

(II) Identical Character Component Pairs, where the numbers of characters in the jwd and cwd are identical, and both contain the same set of characters in different order. We call these pairs homomorphemic pairs. They can be exemplified by Japanese 制限 vs. Chinese 限制 'to restrict', and Japanese 律法 vs. Chinese 法律 'law'.

(III) Partly Identical Pairs, where at least one Kanji in the jwd matches with a Hanzi in the cwd. For example Japanese 相合 can be paired with Chinese 相對於, 合力, 相形之下, 看相, 縫合 etc. The semantic relation between each pair, if it does exist, may be quite distant. However, including this class allows us to cover all possible mappings, as well as to study the kind of conceptual clustering represented by shared characters in either language.

Jwc-cwd word pairs in such mapping groups are searched and compared with the following algorithm: (1) A jwd and a cwd are compared. If the words are identical, then they are considered as a homographic pair. (2) For all non-homographic pairs, if the two words have the same string length, then check the characters contained in each word. If both contain the exact same set of characters, then they are a homomorphemic pair. (3) If the pair has different string length or do not contain the exact character sets, check if there is any character shared by the pair. If there is one of more shared characters, then the pair is a partly identical pair.

After the mapping procedure, if the jwd is not mapped to any of the cwd, the jwd is classified to (IV) uniquely Japanese group. If a cwd is not mapped by any of the jwd, it is classified to (V) uniquely Chinese group.

2 Note that as these forms are used in both languages but cannot be expected to have the exact meaning. Hence the free translation intends to capture only the rough conceptual equivalence in both languages.

3 Logically, all homographic pairs are also homomorphemic pairs. However, for classificatory and comparative reasons, we use homomorphemic pairs to refer only to non-homographic ones.
4.2 Establishing Semantic Relation in Word Pairs

After the character-based mapping, the senses of (I) homographic pairs and (II) homomorphemic pairs are compared in order to establish their cross-lingual semantic relations according to the following three classifications:

(I-1, II-1) Synonym pairs with identical POS:
E.g.
(1-1) 以降: afterwards (noun)
兄弟 (Japanese) and 弟兄 (Chinese): brother (noun)

(I-2, II-2) Synonym pairs with unmatched POS: words in a pair are synonym with different POS or POS of at least one of the words in the pair is missing.
E.g.
(1-2) 意味: sense (noun in EDR and verb in CWN)
(2-2) 定規 (Japanese) and 規定 (Chinese): rule (noun in EDR and no POS is indicated in CWN)

(I-3, II-3) unknown relation: the relation is not determinable by machine processing with the given information at this point.
E.g.
(1-3) 灰: ash (noun)
灰: dust (no POS indicated)
(2-3) 愛心: affection (noun)
心愛: dear, darling (no POS indicated)

In order to find the relation of J-C word pairs, the jwd and the cwd in a pair are compared according to the following information;

(2)
Jwd: English translation in EDR (jtranslation), POS
Cwd: English translations in CWN (ctranslations), POS, cwd synset (English)

The comparisons are done in the following manner; Check if the jtranslation matches with any of the ctranslations or a word in the cwd synset. If no match was found, the pair is unknown relation. If any match was found, check if the POS are identical. If the POS are identical, the pair is a synonym pair with identical POS. Otherwise the pair is a synonym pair with unmatched POS.

After the process, synonym pairs with identical POS and synonym pairs with unmatched POS are examined manually to see if they are really synonyms.

Unknown Relation Analysis
The pairs with unknown relation are divided into the following four different groups.
Only jtranslation is missing (Only comparison info. of jwd is missing)
E.g.
(1-3-A) No English translation for 兄 in EDR
(2-3-A) No English translation for 運命 in EDR
Only ctranslations and cwd synset are missing (Only comparison info. of cwd is missing)
E.g.
(1-3-B) No English translation nor synset for 有無 in CWN
(2-3-B) No English translation nor synset for 明星 in CWN
No comparison info. is missing.
E.g. (1-3-C)
Japanese  Chinese
火力: firepower (noun) 火力: power, powerfulness, potency (no POS)
(2-3-C)
Japanese  Chinese
末期: end (noun) 期末: concluding, final, last, terminal (noun)

日本 translation, ctranslations and cwd synset are missing (Both comparison info. are missing)
E.g.
(1-3-D) No English translation nor synset for 機動 in both EDR and CWN
(2-3-D) No English translation nor synset for 山中 in EDR and for 中山 in CWN

Then the group, (A), (B) and (C), are sorted into possible synonym pairs and non-
synonym pairs by using the following method.
Check if the definition of jwd contains any of the ctranslations or cwd synset. If
the definition contains any of them, then the pair is possible synonym pairs. Otherwise they are non-synonym pairs.
Check if the definition of cwd contains the jtranslation. If the definition contains the
jtranslation, then the pair is possible synonym pairs. Otherwise they are non-
synonym pairs.
Do both the methods that for (A) and (B).

5 The References Section

Hanzi Mapping

Table 1. J-C Hanzi Similarity Distribution.

<table>
<thead>
<tr>
<th></th>
<th>Number of Words</th>
<th>Number of J-C Word Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Identical Hanzi Sequence Pairs</td>
<td>2881 jwds</td>
<td>20580</td>
</tr>
<tr>
<td>Without variant mapping</td>
<td>2815 jwds</td>
<td>20199</td>
</tr>
<tr>
<td>Difference</td>
<td>+66 jwds</td>
<td>+381</td>
</tr>
<tr>
<td>(2) Different Hanzi Order Pairs</td>
<td>207 jwds</td>
<td>481</td>
</tr>
<tr>
<td>Without variant mapping</td>
<td>204 jwds</td>
<td>473</td>
</tr>
<tr>
<td>Difference</td>
<td>+3 jwds</td>
<td>+8</td>
</tr>
</tbody>
</table>
Finding Synonyms (Word Relations)

| Table 2. Identical Hanzi Sequence Pairs (20580 pairs) Synonymous Relation Distribution |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
|                               | Number of 1-to-1 Form-Meaning Pairs Found by Machine Processing (% in (1)) | Number of 1-to-1 Form-Meaning Pairs Found by Manual Analysis (% in (1)) | * Number of Many-to-Many Form-Meaning Pairs Found by Manual Analysis |
| (1-1) Synonym with the same POS pairs | 92 (0.4%) | 35 (0.2%) | 26 |
| Without variant mapping | 92 | 35 | 26 |
| Difference | 0 | 0 | 0 |
| (1-2) Synonym with unmatched POS pairs | 439 (2.1%) | 262 (1.3%) | 153 |
| Without variant mapping | 425 | 254 | 150 |
| Difference | +14 | +8 | +3 |
| (1-3) unknown relation | 20049 (97.4%) | - | - |
| Without variant mapping | 19682 | - | - |
| Difference | +367 | - | - |
Table 3. Identical Hanzi But Different Order Pairs (481 pairs) Synonymous Relation Distribution

<table>
<thead>
<tr>
<th><strong>(2-1) Synonym with the same POS pairs</strong></th>
<th>Number of 1-to-1 Form-Meaning Pairs Found by Machine Processing (% in (2))</th>
<th>Number of 1-to-1 Form-Meaning Pairs Found by Manual Analysis (% in (2))</th>
<th>* Number of Many-to-Many Form-Meaning Pairs Found by Manual Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without variant mapping</td>
<td>0 (0.0%)</td>
<td>0 (0.0%)</td>
<td>0</td>
</tr>
<tr>
<td>Difference</td>
<td>±0</td>
<td>±0</td>
<td>±0</td>
</tr>
<tr>
<td>(2-2) Synonym with unmatched POS pairs</td>
<td>14 (2.9%)</td>
<td>11 (2.3%)</td>
<td>10</td>
</tr>
<tr>
<td>Without variant mapping</td>
<td>14</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Difference</td>
<td>±0</td>
<td>±0</td>
<td>±0</td>
</tr>
<tr>
<td>(2-3) unknown relation</td>
<td>467 (97.1%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Without variant mapping</td>
<td>459</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Difference</td>
<td>+8</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* Many-to-Many Form-Meaning Pair refers to a mapping between a group of jwds, which have the same senses, and a group of cwds that corresponds with the jwds.
** No pair found in (2-1), because, all of jwds in (2-1) also has identical Hanzi sequence cwd in the given data.

Unknown Relation Analysis

Table 4. Identical Hanzi Sequence Pairs with Unknown Relation (20049 pairs) distribution

<table>
<thead>
<tr>
<th><strong>(A) Missing the Japanese translation</strong></th>
<th>Number of Pairs (% in 1-3)</th>
<th>Number of Possible Synonym Pairs (% in 1-3)</th>
<th>Number of Non-Synonym Pairs (% in 1-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without variant mapping</td>
<td>8618 (43.0%)</td>
<td>607 (3.0%)</td>
<td>8011 (40.0%)</td>
</tr>
<tr>
<td>Difference</td>
<td>8428</td>
<td>590</td>
<td>7838</td>
</tr>
<tr>
<td>*** (B) Missing Chinese translation and the synset</td>
<td>2298 (11.5%)</td>
<td>0 (0.0%)</td>
<td>2298 (11.5%)</td>
</tr>
</tbody>
</table>
### Table 5. Identical Hanzi But Different Order Pairs with Unknown Relation (467 pairs) distribution

<table>
<thead>
<tr>
<th></th>
<th>Number of Pairs (% in 2-3)</th>
<th>Number of Synonym Pairs (% in 2-3)</th>
<th>Possible Synonym Pairs</th>
<th>Non-Synonym Pairs (% in 2-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Missing the Japanese translation</td>
<td>207 (44.3%)</td>
<td>7 (1.5%)</td>
<td>200 (42.8%)</td>
<td></td>
</tr>
<tr>
<td>Without variant mapping</td>
<td>199</td>
<td>5</td>
<td>194</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>+8</td>
<td>+2</td>
<td>+6</td>
<td></td>
</tr>
<tr>
<td>*** (B) Missing Chinese translation and the synset</td>
<td>46 (9.9%)</td>
<td>0 (0.0%)</td>
<td>46 (9.9%)</td>
<td></td>
</tr>
<tr>
<td>Without variant mapping</td>
<td>46</td>
<td>0</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>+0</td>
<td>+0</td>
<td>+0</td>
<td></td>
</tr>
<tr>
<td>(C) No missing information</td>
<td>151 (32.3%)</td>
<td>10 (2.1%)</td>
<td>141 (30.2%)</td>
<td></td>
</tr>
<tr>
<td>Without variant mapping</td>
<td>151</td>
<td>10</td>
<td>141</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>+0</td>
<td>+0</td>
<td>+0</td>
<td></td>
</tr>
<tr>
<td>(D) Missing both translations and the synset</td>
<td>63 (13.5%)</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Without variant mapping</td>
<td>63</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>+0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

*** In both group (B), all of the CWN has no definition either, therefore no possible synonym pair is found.
6 Conclusion

In this paper, we present our study over Japanese and Chinese lexical semantic relation based on the Hanzi sequences and their semantic relations. We compared Electric Dictionary Research [5] with the Chinese WordNet [4] in order to examine the nature of cross-lingual lexical semantic relations.

The following tables are summarized tables showing the Japanese-Chinese form-meaning relation distribution examined from this experiment.

Table 6. Identical Hanzi Sequence Pairs (20580 pairs) Lexical Semantic Relation

<table>
<thead>
<tr>
<th></th>
<th>Pairs Found to be Synonym (% in (1))</th>
<th>Pairs Found to be Non-Synonym (% in (1))</th>
<th>Unknown Relation (% in (1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Analysis</td>
<td>1460 (7.1%)</td>
<td>15819 (76.9%)</td>
<td>3301 (16.0%)</td>
</tr>
<tr>
<td>Without variant mapping</td>
<td>1403</td>
<td>15537</td>
<td>3259</td>
</tr>
<tr>
<td>Difference</td>
<td>+57</td>
<td>+282</td>
<td>+42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Pairs Found to be Synonym (% in (2))</th>
<th>Pairs Found to be Non-Synonym (% in (2))</th>
<th>Unknown Relation (% in (2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Analysis</td>
<td>31 (6.4%)</td>
<td>387 (80.5%)</td>
<td>63 (13.1%)</td>
</tr>
<tr>
<td>Without variant mapping</td>
<td>29</td>
<td>381</td>
<td>63</td>
</tr>
<tr>
<td>Difference</td>
<td>+2</td>
<td>+6</td>
<td>+10</td>
</tr>
</tbody>
</table>

Hanzi variants were not taken into account in the previous experiment. This time, when Hanzi variants are taken into account, there found more J-C word pairs such that each of the word in the pair contains a Hanzi that they are actually variants of each other, thus the words are actually related in a sense.

As the table shows, there are more than 75% of pairs are found to be non-synonyms. However, it is not certain whether if the pairs are really non-synonyms and what their actual semantic relations are. In the further experiment, we will try to find the semantic relation (not only synonymous relation) of those pairs found to be non-synonym pairs at this point and analyze the relation of Japanese and Chinese Hanzi narrower and to get more accurate result.
References

Paranymy: Enriching Ontological Knowledge in WordNets

Chu-Ren Huang, Pei-Yi Hsiao, I-Li Su, and Xiu-Ling Ke
Institute of Linguistics, Academia Sinica
Nankang, Taipei, Taiwan 115
{churen, pyxiao, isu, vitake}@gate.sinica.edu.tw

Abstract. This paper studies and explicates the rich conceptual relations among sister terms within the WordNet framework. We follow Huang et al. [1] and define those sister terms by a lexical semantic relation, named paranymy. Using paranymy, instead of the original indirect approach of defining the sister terms as words that have the same hypernym, enables WordNets to represent and enrich ontological knowledge. The familiar ontological problem of ‘ISA overload’ [2] can be solved by identifying and classifying conceptually salient groups among those sister terms. A set of paranyms are terms which are grouped together by one conceptual principle, which are evidenced by their linguistic behaviors. We believe that treating paranymy as a bona fide lexical semantic relation allows us to explicitly represent such classificatory information and enriches the ontological layer of WordNets.

1 Introduction

One area where formal ontologies are considered to be more powerful than WordNets (some times referred to as linguistic ontologies) is the explicit mechanism for defining concepts that allows inferences [3]. It is easy to observe that the sister terms in Princeton WordNet can still contain a cluster of undifferentiated concepts. For instance, the direct hypernyms of ‘cardinal compass point’ in Princeton WordNet, are four sister terms: north, south, east, and west. These four terms are not really equal because there are two conceptual salient pairs: north-south and east-west that play important roles in other conceptual classifications. In terms of conceptual representation, the simple IS-A defining hypernymy in WordNets is inadequate to deal with the complex conceptual relations among the sister terms, known as ‘ISA overload’ [2]. There are at least two possible approaches to deal ISA overload in ontologies. Guarino [4] suggests a finer classification of upper categories, while SUMO [3] implements the two contrast pairs with antecedent axioms. In this paper, we explore the possibility of solving this ISA overload problem while maintaining the original WordNet structure and sister term relations. We present two critical parts of our treatment of paranymy as a lexical semantic relation in this paper: (i) the classificatory criteria for those elements in order to define X as a C, and (ii) the salient relation(s) among those different elements in X.
Huang et al. [1] examined sets of coordinate terms and discovered that the semantic relation, antonymy, was commonly used to explain the relation among those coordinate terms. However, antonymy and other relations, such as near-synonymy, are inadequate to account for their conceptual clustering or entailments. In order to give a more precise and richer semantic representation of lexical conceptual structure and ontology, the idea of paranymy was proposed. It was claimed that this proposal allows WordNets to incorporate representations on semantic field.

In Princeton WordNet (PWN, [5]), sister terms are defined as those coordinate words that have the same hypernym (also called “superordinate” in PWN). Such approach indeed enables the representation for some ontological knowledge. However, the hypernymy is quite general, and it is not specific concept to cover the further detailed relations for its set of hyponyms. When we reconsider the relation among those hyponyms, we realized that these coordinate terms could be reclassified into conceptually salient groups. In the earlier works on the theory of semantic field such as [6] and [7], they had provided the clear explication of how lexical concepts cluster without actually laying out a comprehensive conceptual hierarchy. Therefore, in this paper, we plan to identify the salient groups, try to improve the comprehensive conceptual hierarchy, and enrich the ontological knowledge of WordNets by means of classificatory information with a richer layer.

In what follows, section 2 discusses the different phenomena of sister terms in WordNets. The types and definitions of paranyms are explained in the section 3, with practical examples in Chinese. The conclusion of this paper is given in the section 4.

2 Sister terms (coordinate terms) in WordNets

It is easy to observe that not all coordinate terms are equal when detailed lexical analysis is done for a set of coordinate terms sharing the same hypernym. For example, when people talk about seasons, the first intuition for this concept will be four seasons—spring, summer, fall (or autumn), and winter. Other terms for seasons, such as dry season and rainy season, are not thought of intuitively as parallel as the four seasons although all of them share the same superordinate concept, “seasons in a year”. The same situation happens in the contrast between North vs. Southeast. Generally speaking, North and Southeast are both hyponyms of the concept, geographic direction. However, when we talk about the concept of geographic direction, only the four cardinal compass points, namely East/West/South/North, would come up intuitively as a set of hyponyms under this concept. Neither the North/Southeast pair nor the South/Northeast pair may be viewed as the four main directions at an equivalent level. In WordNets, the knowledge representation for those phenomena is very unclear because all those situations are simply classified by using the relation, sister terms. This is typical dilemma of ISA overload when two sets of unequal hyponyms: four seasons vs. dry season and rainy season, are grouped as sister terms. Similarly, although the four cardinal compass points and all other non-cardinal compass points such as Southeast and Northeast, are all directions, grouping them with the same ISA relation has two parallel problems: that they are not equally
privileged, and that they form contrast pairs among themselves based on the relation of opposite directions.

It is important to notice that conceptual dependencies entail linguistic collocations. For instance, the relations among the four cardinal compass points are revealed in various collocations formed by the North/South pair (or the East/West pair). Such collocations are fairly productive, while other combinations, such as South/East, would be regarded as the rare pair. In terms of conceptual structure and knowledge representation, it is essential to further specify the direction contrast pairs of North/South and East/West among the four main directions. Such conventional collocation will play an important role in our reclassification of hyponyms.

3 The definition and types of paranymy

The semantic relation paranymy is used to refer to the relation between any two lexical items belonging to the same semantic classification in [1]. A paranyms relation must conform to the following basic requirements. The first requirement is that paronyms need to be a set of coordinate terms since they share the same hypernym (also called “superordinate”). Secondly, paronyms have to share the same classificatory criteria. The second requirement is critical and has very interesting consequences because the same conceptual space/semantic field can be partitioned differently by different criteria. For example, as shown by example (1), (1a) and (1b) are both possible exhaustive enumerations of the concept “seasons in a year.” People who live in a certain area, such as Southeast Asia, may prefer to use (1b) to describe their “seasons in a year”; however, to other people in the world, the four seasons of (1a) is the default.

(1) Two sets of paronyms of the main concept- “seasons in a year”
a. chun1/xia4/qiu1/dong1
   “spring/summer/fall(autumn)/winter”

b. gan1 ji4/yu3 ji4
   “dry season/rainy season”

In addition, paranymy can capture how these concepts cluster by stipulating the same criterion they shared for conceptual classification. As shown in above (1), any element of these two different criteria, such as xia4(summer) in (1a) and gan1 ji4(dry season) in (1b), do not stand in direct contrast against each other although they are coordinate terms of the same concept “seasons in a year”. In other words, (1a) and (1b) do not belong to the same semantic field, which are defined by minimal semantic contrasts [6].

---

1 Please note that we are making a distinction between 'rainy season (i.e. monsoon season)' as a primary classification of seasons from the secondary classification of seasons, such as winter and spring are rainy seasons in Taiwan.
One important consequence of allowing classificatory criteria to determine a set of paranyms is that the same set of sister terms may receive overlapping classification, such as in (2)

(2) Directions in Chinese
   a. si4mian4 ‘four directions’
      dong1/xi1/nan2/bei3
      ‘East/West/South/North’

   b. ba1fang1 ‘eight directions’
      dong1/xi1/nan2/bei3/dong1nan2/xi1bei3/dong1bei3/xi1nan2
      ‘East/West/South/North/SouthEast/NorthWest/NorthEast/SouthWest’

The examples in (2) show that the cardinality of directions in Chinese does not have to be four. What is more important is that (2a) is a subset of (2b). In our definition of paranymy, the relation is governed by a classificatory criterion. For instance, South and SouthWest are paranyms under the classification of ba1fang1 (Eight directions), but not under the classification of si4mian4 (Four directions).

It is important to note this current study differs from [1] in terms of our treatment of what they called complementary paranymy. Huang et al. [1] proposed three types of paranymies: complementary, contrary, and overlapping. After our further study based on extensive examples from the Chinese WordNet (CWN) and data from two Austronesian languages (i.e., Formosan languages in Taiwan), Kavalan and Seediq [8, 9], we decided that the relations complementary paranymy intended to capture is better characterized by simply using the semantic relation, antonymy. The examples are given in (3) and (4). Basically, such complementary relation is a binary pair and the typical criterion of this type is “either A or B.” More specifically, under a concept, there are only two possible nodes, A or B and these two nodes are contradictory. Therefore, in this relation, either A or B will appear and this also infers that the positive of one term necessarily implies the negative of the other. Therefore, we can say that the relation between A and B is actually antonymy. This can be exemplified by (3) and (4).

(3) Data from CWN
   State of life: si3/huo2 “dead/alive”
   Amount: dan1/tu4 “singular/plural”

(4) Data from Kavalan and Seediq
   Kavalan: binus/putay ‘alive/dead’
   Seediq: muudus/muhuqin ‘alive/dead’

Hence we propose to revise the classification of [1] to include only two types of paranymy: contrary and overlapping.
3.1 Contrary Paranymy

Contrary paranymy conforms to a condition that each of a set of terms is related to all the others by the relation of incompatibility [10]. The paranyms in this type are gradable and their senses are usually contrary. Contrary paranymy allows intermediate terms, so it is possible to have something that is neither A nor B. For example, something may be warm if it is neither hot nor cold. Besides, contrary paranyms are usually relative, for instance, a thick pencil is likely to be thinner than a thin girl. Contrary paranyms are classified under the perceptual or conventional paradigms. The perceptual paradigm is based on human perception or senses, for example, the superordinate node of fast/slow is speed. Whether the speed is fast or slow, it all depends on somebody’s perception and such perception is variable from one to another.

The conventional paranyms is shown in Fig. 1a. There are various ways of addressing parents based on the register for conventionally using those terms in Chinese, so those terms shall be further classified into different groups rather than directly placing them all under the same superordinate, parent. As shown in Fig. 1b, after re-clustering the sister terms, we get three sub-classes because of the register for using the terms to address parents. Using such re-clustered classification makes the conceptual structures clearer and better.

![Figure 1a. Parents Addressing](image1.png)

![Figure 1b. Parents Addressing (Concepts Re-clustered by the register)](image2.png)

Besides, the contrary paranyms can be divided because of the collocation. For example, in Fig. 2a, a series of coordinate terms in Chinese that is all used to address the spouse in marriage under the colloquial register. Due to the collocation of those
terms, most native speakers think that the contrary paranym of “xian1 sheng1” (husband) is “tai4 tai4” (wife) rather than “qi1 zi5” (wife). Similarly, the contrary paranym of “zhang4 fu1” (husband) is “qi1 zi5” (wife) rather than “tai4 tai4” (wife).

By the relation of paranymy, we can give a more precise account for the coordinate terms or hyponyms, especially ones in the contrary type. A process of re-clustering sister terms can be formulated, as given in Fig. 3, and therefore, such conditions can be applied to augment wordnets with for descriptions of important linguistic collocations and relations.

3.2 Overlapping Paranymy

Overlapping paranymy is defined as the case containing a paradigmatic relation of inclusion and that of exclusion in linear structures. In other words, two sister terms belonging to this type have some features in common, and meanwhile, comprise other distinct features. Overlapping paranyms may include some cases illustrating the relation of incompatibility and oppositeness, in which the contrastive part is more

---

2 Please note that overlapping paranymy we call in this paper is not referred to overlapping antonyms that Cruse [12] terms good/bad and other antonyms having evaluative polarity as part of their meaning.
predominant than the overlap, and also contain near-synonyms, where the features they share are considerable and more salient than those different (e.g., [10, 11]). As [1] explicated, the type of overlapping paranymy is elaborated on the basis of conventions, which are consistently shared by a language community and conform to their experience. The contexts in which the contrast in a pair of overlapping paranymy is foregrounded or not, as well as how their semantics overlaps, depends on discoursal conventions. This is evident in the choice made between the two conventional expressions of greeting, good afternoon and good evening. Both expressions are alternative in a certain time period, say the late afternoon, which indicates the overlap between the time periods denoted by these two sister terms—afternoon, and evening. Therefore, they are regarded as overlapping paranymy.

The following examples extracted from CWN illustrate a similar case. For instance, both coordinate terms in (5) are overlapping, in that they can be alternative terms we choose to call a large stream of water, while they are differentiated from each other in some other contexts. Besides, in (6), both xiang1 zi5 and he2 zi5 can be used to refer to “box”, but when we see a container for a diamond ring, we may call it he2 zi5 rather than xiang1 zi5. Conversely, we may call a container for a TV set xiang1 zi5 rather than he2 zi5. To capture such relation between two sister terms, which the traditional semantic relations, such as antonymy and near-synonymy, cannot deal with, we appeal to overlapping paranymy.

(5) A large natural stream of water: jiang1/he2 “river”

(6) A (usually rectangular) container: xiang1 zi5/he2 zi5 “box”
4 Conclusion

Using paranymy definitely can further analyze the relations among the sister terms within the WordNets framework and enrich the ontological knowledge in WordNets. We introduce this semantic relation, paranymy, into our CWN system and the ontological knowledge of sister terms is indeed elaborated afterward. More precise and clear classifications for clustering the coordinate terms can be obtained. For example, according to the criteria of paranymy, the relationship between brother and sister can be clustered into three classifications. The first classification is based on the same gender but different birth order (older or younger), such as “ge1 ge1” (elder brother) and “di4 di4” (younger brother) / “jie3 jie3” (elder sister) and “mei4 mei4” (younger sister). The second classification is to classify the different genders but having the same birth order (older or younger), for instance, “ge1 ge1” (elder brother) and “jie3 jie3” (elder sister) / “di4 di4” (younger brother) and “mei4 mei4” (younger sister). The third type is based on the concept of collateral relatives by blood, so those four coordinate terms, “ge1 ge1”, “jie3 jie3”, “di4 di4”, and “mei4 mei4” are all grouped together under the concept of sibling. Such three distinctive relations for paranym illustrate the enrichment for the knowledge system.

It is important to note that the knowledge enrichment nature of paranymy comes not only from the introduction of this lexical semantic relation but also from the definition that requires different sets of paranymys be differentiated by their conceptual classificatory criteria. Such knowledge can be encodes explicitly, by simply listing the different subsets of paranymys. Such as {East, West, South, North} and {East, West, South, North, SouthEast, NorthWest, SouthWest, NorthEast} are listed as two different paranym relations. However, such criteria can also be explicitly represented, as in Figures 1 and 2. Our tentative proposal now is to maintain the item and relations only approach established by PWN. However, it can also be argued that the conceptual criteria for classification must be explicitly represented in order to resolve ISA overload. This architecture problem will be resolved in future studies.

References


Proposing Methods of Improving
Word Sense Disambiguation for Estonian

Kadri Kerner
University of Tartu
Institute of Estonian and General Linguistics
Liivi 2–308, Tartu, Estonia
kadri.kerner@ut.ee

Abstract. This paper proposes some methods of making word sense disambiguation (WSD) a more feasible task for Estonian language. Both automatic and manual WSD are kept in mind. Firstly, this paper gives an overview of WSD and manual annotation of Estonian language. Also the paper gives a brief overview of the Word Sense Disambiguation Corpus of Estonian (WSDCEst). Based on this corpus it is possible to examine contextual and other patterns of the target word in order to create disambiguation rules. This corpus is annotated based on Estonian WordNet (EstWN). The second part of this paper discusses the fine-grainedness of EstWN. And one way towards WSD improvement is reducing the fine-grained sense inventory of the resources. It can be done by grouping similar word senses in EstWN. In this paper some existing work with the WordNet is followed.

Keywords: Word sense disambiguation, semantic annotation, corpora, Estonian WordNet, similar sense grouping.

1 Word Sense Disambiguation of Estonian Language

Word sense disambiguation is the first step of semantic analysis of some language. WSD is needed for other natural language processing applications, such as machine translation, information retrieval etc. Since WSD is still one of the difficult problems in NLP, it is important to find methods of improving it.

Word sense disambiguation is closely connected to morphological and syntactic disambiguation.

Lexical entries (literals) in EstWN\(^1\) are presented nominal singular form for nouns and supine form for verbs. In real texts, the words are mostly in their full richness of forms. Lemmatizing and part-of-speech-tagging are made with Estmorf tagger [3]. In sense annotating we considered only nouns (_S_ com) and non-auxiliary verbs (_V_main or _V_mod).

\(^1\) for details see Orav et al in this volume
The modal verbs are explicitly marked in the output of the morphological disambiguator (_V_ mod). When a verb is marked as such, then the senses that don't correspond to the modal senses could be removed, e.g. the verb saama has 12 senses in EstWN, but only 2 of them (can or may) correspond to the modal use of the word.

The output of the morphological analyzer often contains valuable information for word sense disambiguation. In some cases the word-form used in the text can uniquely specify the sense of the word, although its lemma is ambiguous, e.g. the word palk can either mean salary or log of a tree, but its genitive form is different in each meaning (either palga or palgi). By using only the lemma we ignore this distinction that can be explicitly present in the text [5].

At the moment the input text contains no information about its syntactic structure, most importantly the verbal phrases and other multi-word units are not marked as such. Also, the syntactic structure can help to reduce the number of possible senses to choose from. For example the most frequent verb olema (English be, have) has five more frequent senses. Only one sense is present in complementary clauses; three senses appear in existential sentences and one in possessive sentences. Linguistic knowledge about the nature of the sentence can help the disambiguation process of human annotator [5].

## 2 Manual Tagging of Word Senses

During four years around 110 000 running words are looked over and all content words in texts are manually annotated according to EstWN word senses. Twelve linguists and students of linguistics tagged nouns’ and verbs’ senses in the texts; each text was disambiguated by two persons. Pre-filtering system added lexeme and number of senses for each annotating word found in EstWN. Annotators marked in brackets the sense number of EstWN which matched best with used sense of a word by their opinion. If the word was missing from the EstWN, “0” was marked as sense number, and if the word was found in EstWN, but missed appropriate sense, “+1” was marked. If inconsistencies were met, they were discussed until agreement was achieved. On about 20% of cases the disambiguators had different opinions. This fact also indicates the most problematic entries in EstWN and the need to reconsider the borders of meaning of some concepts.

## 3 Word Sense Disambiguation Corpus of Estonian²

The research group of computational linguistics of the University of Tartu has developed Word Sense Disambiguation Corpus of Estonian (WSDCEst). The source texts are mostly fiction and manually annotated.

It should be kept in mind, that not all words can be disambiguated, but only content words. Although normally nouns, verbs, adjectives and adverbs are considered as

² [http://www.cl.ut.ee/korpused/semkorpus](http://www.cl.ut.ee/korpused/semkorpus)
content words [10], in WSDCEst only nouns and verbs were subject to disambiguation.

WSDCEst consist of two parts: base corpus and sentences expressing motion (from the corpus of generated grammar of Estonian single clauses). For the base corpus 43 texts (each text file contains about 2500 tokens) are chosen for word sense disambiguation from Corpus of the Estonian Literary Language (CELL) sub-corpus of Estonian fiction from 1980s. Table 1 gives an overview of current data of WSDCEst.

The size of the Base-corpus is presently being increased mainly with newspaper texts.

Table 1. Words and senses in WSD Corpus of Estonian

<table>
<thead>
<tr>
<th></th>
<th>morphologically analyzed units</th>
<th>% of running words in text</th>
<th>% of all substantives</th>
<th>% of all verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base-corpus</td>
<td>113870</td>
<td>34,65</td>
<td>64,97</td>
<td>90,48</td>
</tr>
<tr>
<td>Sentences expressing motion</td>
<td>5738</td>
<td>66,76</td>
<td>72,50</td>
<td>91,48</td>
</tr>
</tbody>
</table>

4 Exploiting and Examining the WSDCEst

It is possible to observe the contextual and other patterns in word sense disambiguation corpus, in order to create certain rules for certain word senses. These rules are meant to improve the disambiguation task. For Estonian language around 200 rules for frequent nouns were found. Both automatic and manual WSD could benefit from using these rules. As we are currently increasing the WSDCEst, we hope to get detailed feedback and evaluation of using these rules from human annotators. The following work is to include these rules as independent module to an existing automatic WSD tool Semyhe3 to test their effectivity.

Rules are represented in disambiguation manual this way, that human annotator hopefully can easily follow them, for example:

RULE: choose *keel* sense-2 (English *language*):
If the directly preceding word is a genitive attribute;
RULE: choose *keel* sense-4 (English *tongue*):
If the word *keel* (English *tongue*) is in the form of singular allative.

3 http://uuslepo.it.da.ut.ee/~kaarel/NLP/eng_semyhe/
Sometimes the near/local context gives the right sense of a target word – for examples if there is a word marriage in the sentence, then the words man and woman are in the sense of married people (husband and wife) [4]. Another example is that a word *clock* is in the sense of clock time (not a watch), if there is a number in the near context. For determining some word senses it is important the directly preceding or following word. For example the word *language* (in sense of the speech) can be tagged only with a suitable sense number, if it is directly followed by one Estonian adposition *järgi*. Also, the genus of a word or the plural/singular form of a target word can point out the suitable sense. For example if the word *soul* (Estonian *hing*) is in the form of singular illative, it gets always only one suitable sense number. By examining the word sense disambiguation corpus, several other minor observations were made, such the target word’s position in the sentence helps to determine the correct sense (for example, words *man* and *woman* in the beginning of sentence) tend to carry only one sense; the target word’s appearance in a specific construction indicates the sense of a target word.

It was possible to find rules to these word senses, which are more frequent in WSDCEst. Yet, words with high frequency in texts are often abstract and therefore difficult to describe. For example words *thing* (Estonian *asi*), *mind* (Estonian *meel*), *thought* (Estonian *mõte*) etc. Since abstract nouns are very common, it could be useful to deal more with them, rather concentrating on concrete nouns.

### 5 WordNet as Resource for WSD

*WordNets* as lexical-semantic databases are used often for WSD, because of their multilingualism, various semantic relations and of course because of their free availability. Yet it has been argued that WordNet is too fine-grained resource for WSD, because natural language processing applications do not need this kind of high level of granularity. Of course, different NLP applications need different kind of granularity too – machine translation systems may need more distinct senses, information retrieval can operate even on homograph level. It is hard for an automatic WSD system to disambiguate between very many senses. Also it can be extremely complicated for human annotators to tag the correct sense, if there are too similar senses. For example, there are 11 senses of an Estonian noun *asi* (English *thing*) in EstWN, some of these senses are too similar that even the context or the whole text does not give the correct sense.

### 6 Grouping Similar Senses

Therefore it could be useful to group some similar word senses in order to make WSD a more feasible task to both automatic systems and human annotators. Ideas for Estonian language described here are closely related to existing work, e.g. [2], [8], [6], [7].
6.1 Grouping Similar Senses According to the Disagreement of Human Annotators

Part of our research focuses on exploiting agreement and disagreement of human annotators: are there any remarkable and important clusters of similar senses. By processing human annotator’s disagreement files we have found possible clusters of similar senses for 50 frequent Estonian verbs and 25 Estonian frequent nouns. Clusters represent the senses, which are similar in human annotators mind and can be therefore grouped. For example in Table 2 there is a high frequency co-occurrence of sense numbers 2 and 3, and this could indicate to the fact that these senses of an Estonian verb *hakkama* may be grouped into one sense only.

<table>
<thead>
<tr>
<th>Combination of sense numbers</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (English <em>approach, deal with</em>) -- 3 (English <em>become, come</em>)</td>
<td>17</td>
</tr>
<tr>
<td>2 (English <em>approach, deal with</em>) -- 5 (English <em>become, start</em>)</td>
<td>10</td>
</tr>
<tr>
<td>2 (English <em>approach, deal with</em>) -- 6 (English <em>catch, grab</em>)</td>
<td>9</td>
</tr>
<tr>
<td>3 (English <em>become, come</em>) -- 5 (English <em>become, start</em>)</td>
<td>3</td>
</tr>
</tbody>
</table>

There is little disagreement among word senses that doesn’t include autohyponymy and/or sisters (co-hyponyms). Also a very important observation is that human annotators disagree less when all the representation fields of EstWN are properly filled (hyperonym (s), synset, definition, explanation). The research referred to the fact that explanations of highly polysemous words seem to be very important for human annotators. When adding missing explanations, the difference between senses becomes more definite to human annotators. The fact that some words are not highly polysemous indicates usually (but not always) to a minor disagreement of human annotators.

Examining human annotator’s disagreement files also refers to problems in EstWN like missing examples, overlapping synsets or explanations and over-grained senses. In many cases it is impossible to determine the one and only sense. Sometimes it is even not necessary [12] and sometimes the nearby context allows different senses.

It is difficult to distinguish word senses that are detectable in EstWN but not visible in the real usage of text (or language). In some cases the disagreement between human annotators arises, because of the lack of lexicographical knowledge (or the human annotator is somewhat superficial).

Exploiting similar sense clusters can be helpful in referring to insufficiency of EstWN. For example, if all the sense numbers of a word combine with each other, it can be assumed that the distribution of senses is incomplete and needs to be improved. If there is no disagreement among human annotators, then there are no remarkable sense clusters and therefore the senses of this particular word are reasonably distributed and not too fine-grained. Also our research showed that words with abstract meanings are difficult to annotate and make up essential sense clusters.
Some researchers [11], [13] claim that words representing so-called Base Concepts are difficult to annotate semantically (apparently because of their broad meanings). This research also confirmed this fact. The boundaries and the area of the usage of a hyperonym or hyponym should be very precisely represented in EstWN. The tendency seems to be that hyponyms as narrower meanings are better to distinguish than hyperonyms. That is the reason, why top concepts tend to combine with many of the different word senses (example in Table 3).

<table>
<thead>
<tr>
<th>Combination of sense numbers</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 (English can) – 11 (English may)</td>
<td>24</td>
</tr>
<tr>
<td>10 (English can) – 9 (English attain, get to)</td>
<td>9</td>
</tr>
<tr>
<td>10 -- 2</td>
<td>8</td>
</tr>
<tr>
<td>10 -- 6</td>
<td>5</td>
</tr>
<tr>
<td>10 -- 3</td>
<td>4</td>
</tr>
<tr>
<td>10 -- 7</td>
<td>4</td>
</tr>
</tbody>
</table>

### 6.2 Grouping Similar Senses by Processing Semantic Relations

Considering the research of examining disagreement of human annotators, arises another possibility to group similar senses. That is by processing some semantic relations, for example auto-hyponymy and co-hyponymy/sisters. This idea relates to the work of Peters et al [9]. When two senses share the same hyperonym, then they are usually intuitively similar, they highlight different aspects of given word. Also the case of auto-hyponymy often shows the similarity and tight relatedness of two (or more) senses. For example in EstWN a noun *aasta* (English *year*) in following example:

ajavahemik 1, periood 1(English amount of time)

hyp=>aasta2

hyp=>aasta1(kalendriaasta1)(English twelwemonth)

hyp=>aasta4

hyp=>aasta3

In this example, aasta2 (English *year* in 2nd sense) is hyperonym to other senses of the same word year – aasta1 and aasta4. Comparing the representations fields of senses 1, 2 and 4 in EstWN, gives the possibility to group these senses.
6.3 Grouping Similar Senses Using “Semantic Mirroring” and Translational Equivalents

“Semantic mirroring” is a method, which allows to derive synsets and senses of words by using translational equivalents in parallel corpora [1]. Also, using this method, it is possible to determine new semantic relations (hyperonymy, hyponymy etc).

In this experiment, I tried to use the simplified version of Semantic mirroring method in order to determine similar word senses. EstWN and English translational equivalents via ILI-link were used. Mirroring was done manually for selected nouns only, in order to test the effectiveness of this method. For example, the noun thing (Estonian asi), which senses are represented in following example: (this noun has altogether 11 senses (represented in bold). Hyperonyms are represented before the sign =>)

- olev 2 (English something, somebody)
  - hyp=>asi1 (objekt 1) (English object)
  - hyp=>asi3
  - hyp=>asi4 (tehisasi 2, artefakt 2) (English artefact)
- tehisasi 2, artefakt 2, asi 4
  - hyp=>asi12 (värk 1, tühi-tähi 1, asjad 1, asjandus 2) (English stuff)
- töö 3 (English work)
  - hyp=>asi5 (toimetus 2, ettevõtmine) (English project, task, undertaking)
- ütlus 1, väljend 1 (English saying, expression)
  - hyp=>asi6
- juht 4, juhtum 1, sündmus 1 (English happening)
  - hyp=>asi7 (lugu2)
- asi iseeneses 1, idee 2, abstraktsioon 1 (English abstraction)
  - hyp=>asi8 (nähtus2) (English thing)
- tegu 2 (English action)
  - hyp=>asi10
- seisund 4, olukord 4, situatsioon 1 (English situation)
  - hyp=>asi11
- atribuut 1, omadus 2 (English attribute)
  - hyp=>asi9

The next Figure (1) shows every possible English translation for an Estonian word asi and vice versa, possible translations back to Estonian language (irrelevant translations are left out and not considered). In circles there are synsets of both languages that correspond to each other.

As a result, it can be observed, that some senses of an Estonian word asi are indeed similar and therefore can be grouped (senses 3, 7, 8, 9, 10, 11):

- (5) ASI, ettevõtmine, ülesanne, toimetus (English project, task, undertaking)
- (4) ASI, tehisasi, artefakt (English artefact, artifact)
- (3, 6, 7, 8, 9, 10, 11) ASI, lugu, nähtus (English thing)
- (1) ASI, objekt (English intimate object, object, physical object)
- (12) ASI, tühi-tähi, värk, asjad, asjad (English stuff, sundries, sundry, whatsis)
7 Conclusions and Future Work

Future work includes gathering feedback from human annotators – how much did they benefit from using the contextual etc rules for WSD of Estonian frequent nouns. Also it is important to test and evaluate these rules, which were created examining the WSDCEst in an Estonian automatic WSD system. A suitable formalism for describing these rules must be thought of. As the WSDCEst is increasing in size, it is possible to create new rules; at the same time it is possible to eliminate unsuitable ones.

There were some words, which senses could not be grouped with semantic mirroring method, although intuitively they seem similar and are used in same contexts in a text. It would be necessary to use parallel corpora for similar sense grouping, because of the real use of language. This task assumes in turn much large-sized parallel corpora.

Also, the future work is finding the best method for grouping similar word senses for Estonian language. It might be useful to combine different grouping methods (semantic mirroring, human annotator’s opinion, some semantic relations; an option would be using systematic polysemy patterns and/or computing semantic similarity. It
could be reasonable to group word senses considering different NLP applications, and specific domains (music, food etc).

Acknowledgements

The work described here was supported by the National Program "Language Technology Support of Estonian Language" projects No EKKTT04-5, EKKTT06-11 and EKKTT07-21 and Government Target Financing project SF0182541s03 ("Computational and language resources for Estonian: theoretical and applicational aspects").

References


Morpho-semantic Relations in WordNet – a Case Study for two Slavic Languages

Svetla Koeva¹, Cvetana Krstev², and Duško Vitas³

¹ Department of Computational Linguistics, Institute of Bulgarian, 52 Shipchenski prohod, 1113 Sofia, Bulgaria
svetla@ibl.bas.bg
² Faculty of Philology, University of Belgrade, Studentski trg 3, 11000 Belgrade, Serbia
³ Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
{cvetana, vitas}@matf.bg.ac.yu

Abstract. In this paper we present the problem of representing the morpho-syntactic relations in WordNets, especially the problems that arise when WordNets for languages that differ significantly from English are being developed on the basis of the Princeton WordNet, which is the case for Bulgarian and Serbian. We present the derivational characteristics of these two languages, how these characteristics are presently encoded in corresponding WordNets, and give some guidelines for their better coverage. Finally, we discuss the possibility to automatically generate new synsets and/or new relations on the basis of the most frequent and most regular derivational patterns.

Keywords: global WordNet, morpho-semantic relations, derivational relations

1 Introduction

The aims of this paper are to present the current stage of the encoding of morpho-semantic relations in Bulgarian and Serbian WordNets, briefly to sketch the derivational properties of Slavic languages based on the observations from Bulgarian and Serbian, to discuss the nature of morpho-semantic relations and its reflection to the WordNet structure and to analyze the positive and negative consequences of an automatic insertion of Slavic derivational relations into it.

The WordNet is a lexical-semantic network which nodes are synonymous sets (synsets) linked with the semantic or extralinguistic relations existing between them [3], [8]. The WordNet structure also includes semantic and morpho-semantic relations between literals (simple words or multiword expressions) constituting the different synsets. The representation of the WordNet is a graph. The cross-lingual nature of the global WordNet is provided by establishing the relation of equivalence between synsets that express the same meaning in different languages [15].
The global WordNet offers the extensive data for the successful implementation in different application areas such as cross-lingual information and knowledge management, cross-lingual content management and text data mining, cross-lingual information extraction and retrieval, multilingual summarization, machine translation, etc. Therefore the proper maintaining of the completeness and consistency of the global WordNet is an important prerequisite for any type of text processing to which it is intended.

The structure of the paper outlines the underlined goals. In the following section we present a short analysis of related work. In the third section, we briefly describe the properties of Slavic derivational morphology based on examples from Bulgarian and Serbian and their reflection into the WordNet structure. The forth section explains how the morpho-semantic relations are encoded in Bulgarian and Serbian WordNets respectively. We then discuss the manners to incorporate the (Slavic) derivational relations into the WordNet structure and some limitations of their automatic insertion. Finally, we raise some problematic questions connected with the presented study and propose future work to be done.

2 Related work

WordNets have been developed for the most of the Slavic languages – Bulgarian, Serbian, Czech, Russian, Polish, Slovenian, and some initial work has been done for Croatian. WordNets for three Slavic languages (Czech – started with the EuroWordNet (EWN), Bulgarian and Serbian) have been developed in the scope of the Balkanet project (BWN) [2], [11] and later on continue developing as nationally funded projects1 or on the volunteer basis.

Originally, the Princeton WordNet (PWN) is designed as a collection of synsets that represent synonymous English lexemes which are connected to one another with a few basic semantic relations, such as hyponymy, meronymy, antonymy and entailment [3], [8]. This same structure has basically been mirrored in most of the WordNets developed on the basis of PWN. The structural difference of Slavic languages which show many similar features has induced the enrichment of WordNets with new information. Added information is mostly related to the inflectional and derivational richness of a language in question. For instance, information related to inflectional properties has been added to all lexemes in Bulgarian [4] and Serbian [2] WordNets, and for Serbian some rudimentary semantic relations that can be inferred from the derivational connectedness, for instance derived-pos (for possessive adjectives) and (for gender motion) derived-gender [2] has been added too. On the other hand, the recognized importance of PWN, and global WordNet in general, for various NLP applications has initiated the major additions and modifications of PWN itself.

The existence of derivational relations that exhibit a fairly regular behavior and that connect lexemes that belong to the same or to the different categories seemed to many as a good starting point for the substantial WordNet enrichment. We will

---

1 http://dcl.bas.bg/bulNet/general_en.html
present the most interesting approaches. All these approaches rely on the fact that if there is a derivational relation between two lexemes belonging to different synsets then most probably there is a kind of semantic relation between the synsets to which the lexemes belong.

The automatic enrichment of WordNet on the basis of the derivational relations has been proposed and used for the Czech WordNet [10]. The basic and most productive derivational relations in Czech have been included in a Czech morphological analyzer and generator, and semantic labels were added to the derivational relations.

The sharing of semantic information across WordNets has been proposed by [1]. Namely, if WordNets for several languages are connected to each other (for instance via Interlingual index (ILI) [15], as has been done for WordNets developed in scope of EuroWordNet and Balkanet projects), then semantically related synsets in a source language for which the connection has been established on the basis of the derivational relatedness of some of the lexemes can be used to connect the synsets in a target language whose lexemes may not exhibit any derivational relation.

The method to improve the internal connectivity of PWN has been proposed in [9]. The existing synsets have been manually connected on the basis of the automatically produced list of pairs of lexemes that are (potentially) derivationally, and therefore also semantically, connected. In this paper we will try to show why we find the last approach the most appropriate for Bulgarian and Serbian.

3 Slavic derivation in WordNet structure

The derivation is highly expressive in all Slavic languages. Some of the most frequent and regular derivational mechanisms in Bulgarian and Serbian are given in the Table 1. The status of the derivational mechanisms listed is not the same. Some of them represent the more or less frequent models which are not applicable to every lemma that has certain syntactic or semantic property, while the other models can always be applied. For instance, the pattern Verb $\rightarrow$ Noun representing a profession is one of the numerous derivational pattern in Bulgarian (уча $\rightarrow$ учитель) and Serbian (учи $\rightarrow$ учитељ), while the pattern Verb $\rightarrow$ Verbal noun is a general rule that can be applied to all imperfective verbs in the two languages. Similarly, a possessive adjective exists for every animate noun [12]. We call this phenomenon a regular derivation since in some respect it enhances the notion of inflectional class.

Formally, regular derivation is performed by derivational operators that significantly influence the structuring of the lexicon of Slavic languages. The analysis of this phenomenon is given in [13], [14] on the examples of processing of possessive and relational adjectives, amplification and gender motion in various English-Serbian and Serbian-English dictionaries. Moreover, the derivational potential is, as a rule, connected to the specific sense of a lemma (see sections 5 and 7).
4 Current state of morpho-semantic relations in Bulgarian and Serbian WordNets

Eight semantic relations between synsets are represented (in a correspondence with the Princeton WordNet) in Bulgarian [4], [5] and Serbian WordNets [2]. These relations are: hypernymy, meronymy (three subtypes are registered among others recognized), subevent, caused, be in state, verb group, similar to and also see (also see in PWN actually encodes two different relations: between verbs and between adjectives, the former one being a kind of morpho-semantic relation between literals roughly corresponding to Slavic verb aspect while the second one is a semantic relation of similarity between synsets). Three extralinguistic relations between synsets are encoded as well: usage domain, category domain and region domain. The WordNet structure includes also semantic and morpho-semantic (derivational) relations among literals belonging to the same or to the different synsets. Semantic relations between literals are: synonymy and antonymy (in Bulgarian and Serbian WordNets antonymy links synsets); derivational are: derived, participle, derivative in Bulgarian, and derived-pos, derived-gm, and derived-vn in Serbian.

4.1 Encoded morpho-semantic relations

The morpho-semantic relations in Bulgarian and Serbian WordNets link synsets although they derivationally apply to the literals only (single word and multi-word

---

Table 1. Some of the derivational mechanisms in Bulgarian and Serbian

<table>
<thead>
<tr>
<th>Relation</th>
<th>Bulgarian</th>
<th>Serbian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspect pairs</td>
<td>уча → науча</td>
<td>учити → naučiti</td>
<td>teach – learn</td>
</tr>
<tr>
<td>Verb → noun</td>
<td>уча → учитель</td>
<td>учити → учителъ</td>
<td>teach – teacher</td>
</tr>
<tr>
<td>Verb → noun</td>
<td>уча → ученик</td>
<td>учити → ученник</td>
<td>learn – student</td>
</tr>
<tr>
<td>Verb → noun</td>
<td>уча → училище</td>
<td>учити → училище</td>
<td>learn – school</td>
</tr>
<tr>
<td>Verb → noun</td>
<td>уча → учебник</td>
<td>учити → уџбеник</td>
<td>learn – textbook</td>
</tr>
<tr>
<td>Verb → noun</td>
<td>уча → учен</td>
<td>учити → ученjak</td>
<td>learn – scientist</td>
</tr>
<tr>
<td>Verbal noun</td>
<td>уча → учение</td>
<td>учити → учење</td>
<td>learn – studies</td>
</tr>
<tr>
<td>Verbal noun</td>
<td>уча → учен</td>
<td>учити</td>
<td>learn – study</td>
</tr>
<tr>
<td>Collective noun</td>
<td>ученик → ученичество</td>
<td>ученик → ученифессо</td>
<td>student – schooldays</td>
</tr>
<tr>
<td>Verb → adjective</td>
<td>уча → учебен</td>
<td>учити → учен</td>
<td>learn – educational</td>
</tr>
<tr>
<td>Verb → adjective</td>
<td>уча → учен</td>
<td>учити → учеvan</td>
<td>learn – educated</td>
</tr>
<tr>
<td>Relative adjective</td>
<td>Учитель → учителски</td>
<td>учителъ → учителски</td>
<td>of or related to teacher</td>
</tr>
<tr>
<td>Possessive</td>
<td>Учитель → учителски</td>
<td>учителъ → учителев</td>
<td>male – female teacher</td>
</tr>
<tr>
<td>Diminutive</td>
<td>ученик – учениче</td>
<td>ученик – учениčи</td>
<td>student – little student</td>
</tr>
</tbody>
</table>

There is actually a whole list of perfective verbs that correspond to the imperfective verb yuwr: izučiti, naučiti, obučiti, preučiti (se), podučiti, poučiti, priučiti, proučiti.

There is actually a whole list of perfective verbs that correspond to the imperfective verb yuwr: izučiti, naučiti, obučiti, preučiti (se), podučiti, poučiti, priučiti, proučiti.

3 Today obsolete.
lemmas). On the other hand, morpho-semantic relations express different kinds of semantic relations which hold between synsets. Neither the derivational links between the exact literals, nor labels \[10\] for the respective semantics relations operating between synsets are encoded so far in Bulgarian and Serbian WordNets. The subsumed morpho-semantic relations are briefly presented below (some statistical data are shown in Table 2):

**Derivative** is an asymmetric inverse intransitive relation between derivationally and semantically related noun and verb. For example the Bulgarian literal водя from the synset \{насочвам:1, насоча:1, водя:4, направям:1, направя:1, направлявам:1\} (the corresponding English synset is \{steer:1, maneuver:1, maneuver:2, manoeuvre:2, direct:11, point:4, head:5, guide:1, channelize:1, channelise:1\} with a definition ‘direct the course; determine the direction of traveling’) is in derivative relation with the noun водач from the synset \{водач:3\} (the corresponding English synset is \{guide:2\} with a meaning ‘someone who shows the way by leading or advising’).

**Derived** is an asymmetric inverse intransitive relation between derivationally and semantically related adjective and noun. For example the literal меден from the Bulgarian synset \{меден:1\} (the English equivalence \{cupric:1, cuprous:1\} with a definition ‘of or containing divalent copper’) is in a derived relation with the literal мед from the synset \{мед:2, Cu:1\} (in English \(\rightarrow\) \{copper:1, Cu:1, atomic number 29:1\}). A productive derivational process rely Slavic nouns with respective relative adjectives with general meaning ‘of or related to the noun’. For example, the Bulgarian relative adjective \{стоманен:1\} defined as ‘of or related to steel’ has the Serbian equivalent \{челични:1\} with exactly the same definition. Actually in English this relation is expressed by the respective nouns used with an adjectival function (rarely at the derivational level, consider wooden wood, golden gold), thus the concepts exist in English as well and the mirror nodes should be envisaged.

**Participle** is an asymmetric inverse intransitive relation between derivationally and semantically related adjective denoting result of an action or process and the verb denoting the respective action or process. Consider играя from \{играя:7\} (the English equivalent \{play:1\} with a definition ‘participate in games or sport’) which is in a Participle relation with the literal игран from \{игран:1\} denoting ‘(of games) engaged of’ for the English counterpart \{played:1\}. All Bulgarian verbs produce participles (the number of participles varies from one to four depending on the properties of the source verb) which are considered as verb forms constituting complex tenses or passive voice. On the other hand, a big part of the Bulgarian participles acts as adjectives with separate meaning. The similar relations between a verb and its participles hold for Serbian.

It can be seen that the actual derivational relations are established between particular literals although the synsets are formally linked (the actual semantic relation between synsets which marker is the derivation itself is not labeled). The English derivative, derived, and participle relations are automatically transferred to Bulgarian WordNet. As they are language specific and obviously there is no one to one mapping between English and Bulgarian the expanded links are manually validated. A specification whether a given morpho-semantic relation exists in English only is declared in a synset note (SNote).

The relation eng_derivative has been also automatically transferred to Serbian although the corresponding derivational relation may hold in Serbian as well but need
not (see the Serbian example in section 5). The new relations derived-pos, derived-vn, and derived-gender have been introduced in Serbian WordNet to relate possessive and relative adjectives, verbal nouns and female (or male) doublets, assigned mainly to the Balkan specific or Serbian specific synsets.

Table 2. Statistical data for the encoded morpho-semantic relations in Bulgarian and Serbian WordNets.

<table>
<thead>
<tr>
<th>Number of</th>
<th>BG WN</th>
<th>SR WN</th>
<th>PWN 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synsets</td>
<td>29,136</td>
<td>13,612</td>
<td>115,424</td>
</tr>
<tr>
<td>Literals</td>
<td>56,223</td>
<td>23,139</td>
<td>203,147</td>
</tr>
<tr>
<td>Relations</td>
<td>53,144</td>
<td>18,2104</td>
<td>204,948</td>
</tr>
<tr>
<td>Derived</td>
<td>1,696</td>
<td>314</td>
<td>1,296</td>
</tr>
<tr>
<td>Derivative</td>
<td>8,920</td>
<td>833</td>
<td>36,630</td>
</tr>
<tr>
<td>Participle</td>
<td>212</td>
<td>0</td>
<td>401</td>
</tr>
</tbody>
</table>

4.2 Not-encoded morpho-semantic relations

The general observations are that not all existing derivative, derived, and especially participle links are marked in Bulgarian and Serbian WordNets. The main reason originates in the language specific character of the word-building in view of the fact that an exact correspondence with the PWN has been mostly followed in the expand WordNet model. As a result a lot of language-specific derivational relations (that can be described in terms of derivative, derived, and participle relations) remain unexpressed in Bulgarian and Serbian WordNets. For example the literals from the Bulgarian synset {метален:1, металлически:1} corresponding to the English synset {metallic:1, metal:1} with a definition: ‘containing or made of or resembling or characteristic of a metal’ are derived from the literal метал from the synset {метал:1, метален елемент} equal to the English synset {metallic element:1, metal:1} with a definition ‘any of several chemical elements that are usually shiny solids that conduct heat or electricity and can be formed into sheets etc’. Nevertheless the corresponding derived relation is not linked in the Bulgarian WordNet. Consider the following more complicated example. The literal пекар from the Bulgarian synset {пекар:1, хлебар:1, фурнаджия:1} (English equivalent {baker::2, bread maker:1} with a definition ‘someone who bakes bread or cake’) is in a derivational relation with the literal пека from the synset {пека:1, опечам:1, опека:1, изпека:1} (in English {bake:1} with a definition ‘cook and make edible by putting in a hot oven’). Moreover the second target literal хлебар is in a derivational relation with the source literal хляб from the synset {хляб:1} (in English {bread:1, breadstuff:1, staff of life:1} with a definition ‘food made from dough of flour or meal and usually raised with yeast or baking powder and then baked’), while the third one фурнаджия is in a derivational relation with the source literal фурна from {пекарница:1, фурна:2} (in PWN {bakery:1, bakeshop:1, bakehouse:1} with a definition ‘a workplace where baked goods (breads and cakes and pastries) are

4 Without extralinguistic relations: category and region, and relation eng-derived.
5 Includes relations: derived-pos, derived-vn, and derived-gender.
produced or sold'). None of the three existing derivational relations is encoded in the Bulgarian WordNet so far.

In Serbian, for instance, the adjective synset \{zamisliv:1\} (English equivalent is \{conceivable:2, imaginable:1, possible:3\} with a definition 'possible to conceive or imagine') is not linked with the verbal synset \{zamisliti:2y, koncipirati:1b\} (in English \{imagine:1, conceive of:1, ideate:1, envisage:1\} with a definition 'form a mental image of something that is not present or that is not the case'), although relation derived, or some more specific, would be appropriate.

### 4.3 Language-specific morpho-semantic relations

There are systematic morpho-semantic differences concerning derivational mechanisms between English and Slavic languages [7]. Some of the most productive derivational relations in Slavic languages are briefly presented here: namely verbal aspect pairs, gender pairs, and diminutives.

#### 4.3.1 Aspect pairs

The verb aspect is a category that occurs in all Slavic languages, its nature is very sophisticated. Generally speaking, the verb aspect in Slavic languages can be described as a relation between the action and its bound (limit) regardless of the person, speaker and speech act. The perfect aspect verbs express integrity and completeness, while the imperfect aspect verbs – lack of integrity or a process (duration, recurrence). Each Slavic verb is either perfective or imperfective; there are a number of verbs that are bi-aspectual and act as both imperfective and perfective. Most verbs form strict pairs where perfective and imperfective members form a derivational relation between two lexemes expressing generally the same meaning. The Bulgarian verbs are classified as: imperfective (perfective correspondent exists), perfective (imperfective correspondent exists), bi-aspectual, imperfective tantum (perfective correspondent does not exist), perfective tantum (imperfective correspondent does not exist). In Bulgarian WordNet the aspect pairs are introduced in one and the same synset with an LNote (literal note) describing the respective aspect. For example \{съчинявам:2 LNOTE: imperfective, съчиня:2 LNOTE: perfective, пиша:4 LNOTE: imperfective, написвам:2 LNOTE: imperfective, напиша:2 LNOTE: perfective\} (an equivalent of the English synset \{write:1, compose:3, pen:1, indite:1\} with a definition 'produce a literary work'). Similarly, in Serbian WordNet the aspect pairs are introduced in a same synset. For instance in a synset \{zamišljati:2x, zamisliti:2x, dočaravati:2x, dočarati:2x, predočavati:1, predočiti:1\} (in English \{visualize:1, visualise:3, envision:1, project:9, fancy:1, see:4, figure:3, picture:1, image:1\} with a definition 'imagine; conceive of; see in one's mind'), LNOTE element corresponding to each literal describes inflectional and derivational properties of each verb, e.g. LNOTE content for the imperfective verb \textit{zamišljati} is V1+Imperf+Tr+Iref+Ref, while LNOTE content the perfective correspondent \textit{zamisliti} is V162+Perf+Tr+Iref+Ref [6]. In most cases, however, perfective verbs derived from the imperfective by prefixation express different meaning and are not in the same synset, for example the perfective verb \textit{uraditi} ‘do, perform’ and its imperfective correspondent \textit{raditi} are not in the same synset.
4.3.2 Gender pairs

The gender pairing is a systematic phenomenon in Slavic languages that display binary morpho-semantic opposition: male $\rightarrow$ female, and as a general rule there is no corresponding concept lexicalized in English. The derivation is applied mainly to nouns expressing professional occupations, but also to female (or male) correspondents of nouns denoting representatives of animal species. For example, Bulgarian synset {преподавател:2, учител:1, инструктор:1} and Serbian synset {предавач:1} that correspond to the English {teacher:1, instructor:1} with a definition: ‘a person whose occupation is teaching’ have their female gender counterparts {преподавателка, учителька, инструкторка} and {предавачица} with a feasible definition ‘a female person whose occupation is teaching’.

There are some exceptions where like in English one and the same word is used both for masculine and feminine in Bulgarian and Serbian, for example {президент:1} which corresponds to the English synset {president:3} with a definition: ‘the chief executive of a republic’, and as a tendency the masculine noun can be used referring to females. Following the PWN practice the female counterparts are encoded in Bulgarian and Serbian WordNets as hyponyms of the corresponding synset with the male counterpart. For example {актриса:1} (English equivalent {actress:1} with a definition ‘a female actor’) is a hyponym of {актьор:1, артист:} (corresponding to the English synset {actor:1, histrion:1, player:3, thespian:1, role player:2} expressing the meaning ‘a theatrical performer’). It might be foreseen of introducing a new relation describing the female – male opposition of nouns in Slavic languages as has already been done for Serbian.

4.3.3 Diminutives

Diminutives are standard derivational class for expressing concepts that relate to small things. The diminutives display a sort of morpho-semantic opposition: big $\rightarrow$ small, however sometimes they may express an emotional attitude too. Thus the following cases can be found with diminutives: standard relation big $\rightarrow$ small thing, consider {стол:1} corresponding to English {chair:1} with a meaning ‘a seat for one person, with a support for the back’ and {столче:1} with an feasible meaning ‘a little seat for one person, with a support for the back’; small thing to which an emotional attitude is expressed. Also, Serbian synset {lutka:1} that corresponds to the English {doll:1, dolly:3} with a meaning ‘with a replica of a person, used as a toy’ is related to {lutica} which has both diminutive and hypocoristic meaning. There might be some occasional cases when this kind of concept is lexicalized in English, {foal:1} with a definition: ‘a young horse’, {filly:1} with a definition: ‘a young female horse under the age of four’, but in general these concepts are expressed in English by phrases.

For the moment the diminutives are included in Bulgarian and Serbian WordNets only in the rare case when the English equivalent is lexicalized. On the other hand, almost from every concrete noun a diminutive (in some cases more than one lexeme) can be derived. Consequently a place for the diminutives in the WordNet structure has to be provided.
5 The nature of morpho-semantic (derivational) relations

One of the most important features of the morpho-semantic relations is that being derivational relations between literals (i.e. assistant is a person that assists; participant is the person that participates etc.) they express also regular semantic oppositions holding between synsets [9]. The derivational relation linking *assist* and *assistant* from the respective synsets {help:1, *assist*:1, aid:1} ‘give help or assistance: be of service’ and {*assistant*:1, helper:1, help:4, supporter:3} ‘a person who contributes to the fulfillment of a need or furtherance of an effort or purpose’ implies a kind of semantic relation over synsets formulated in [10] as an agitative relation existing between an action and its agent.

Given morpho-semantic relation may be realized by different derivation mechanisms. Consider the literals from the Bulgarian synset {певец:2, *вокалист*:1} (in English {singer:1, vocalist:1, vocalizer:2, vocaliser:2} with a definition ‘a person who sings’), the former one певец is derived with the suffix –ец from the literal пея constituting the synset {пея:1} (the English equivalent {sing:2} with a definition ‘produce tones with the voice’), while the second one вокалист is derived with the suffix –ист from the literal вокализирам belonging to the synset {вокализирам:1} (in English {vocalize:2, vocalise:1}” with a definition ‘sing with one vowel’).

On the other hand, different derivational mechanisms might correspond to different semantic relations. For example in Bulgarian, as well as in English the verb че́та from the synset {че́та:3; прочита́м:2; прочита:2} corresponding to the English synset {read:1} with a definition ‘interpret something that is written or printed’ has the following derivates among others:

– the noun чете́нє from the synset {чете́нє:1} ↔ {reading:1}, with a definition: ‘the cognitive process of understanding a written linguistic message’. The derivation transforms the verb into a verbal noun. The respective relation between synsets is formulated as an action relation in [10].

– the noun чита́тел from the synset {чита́тел:1} ↔ {reader:1}, with a definition: ‘a person who enjoys reading’. The derivational relation links the source verb with a noun build by an affixation. The respective relation between synsets expresses a property over the underlying action.

In some cases when the source literal has more than one meaning the exact correspondences with the derivates can be traced. Consider the verb че́та from the synset {че́та:1, прочита́м:1, прочита:1} equivalent with the English synset {read:3} with a definition ‘look at, interpret, and say out loud something that is written or printed’. Its verbal noun derivative чете́нє from the synset {чете́нє:1; поетическо чете́нє:1; решита́л:1} (in English {recitation:2, recital:3, reading:7}) expresses a meaning which is related with the meaning of the source ‘a public instance of reciting or repeating (from memory) something prepared in advance’. As the source derives counterpart in two different synsets (equivalent to {read:1} and {read:3}), this presupposes the corresponding difference in the meanings of the resulting derivatives. Thus the same derivational mechanism might indicate for different semantic oppositions if it targets graphically equivalent literals expressing different meaning (the observed difference in the semantic oppositions remains undistinguished). It is natural that the synsets {read:1} and {read:3} are related with a *verb group* relation.
The semantic part of the morpho-semantic relations is not language specific, language specific are the derivational mechanisms of lexicalization. There are several English derivatives of the literal paint from \{paint:3\} with a definition ‘make a painting of’:

En 1. \{paint:1\} – ‘a substance used as a coating to protect or decorate a surface (especially a mixture of pigment suspended in a liquid); dries to form a hard coating’

En 2. \{painter:1\} – ‘an artist who paints’

En 3. \{painting:1, picture:2\} – ‘graphic art consisting of an artistic composition made by applying paints to a surface’

En 4. \{painting:2\} – ‘creating a picture with paints’

Neither of the corresponding Bulgarian equivalents:

Bg 1. \{боя:2\}

Bg 2. \{живописец:1, художник:1\}

Bg 3. \{картина:3\}

Bg 4. \{живопис:1\}

are derivatives of the Bulgarian synset equivalent to \{paint:3\} – \{рисувам:2; нарисувам:2\}. Nevertheless the same semantic oppositions exist in Bulgarian although they are not marked with any semantic or morpho-semantic relations.

In Serbian some of the related synsets to \{naslikati:1 LNOTE: V101+Perf+Tr+Iref\} (equal to \{paint:3\}) include derivatives, while the other do not (e.g. \{боja:2x, farba:1x\}). The derivative relation is transferred from English to Serbian WordNet, but the name of the relation has not been changed in order to indicate that the origin of the relation is English, and that it may hold for Serbian but need not, as shown by the same example.

Sr 1. \{боja:2x, farba:1x\}

Sr 2. \{сликар:1\}

Sr 3. \{сliка:1\}

Sr 4. \{сликарство:1\}

This means that the derivational relations in a particular language might be successfully used not only for the detecting of a given semantic opposition. Moreover they can be exploited for the identification of the corresponding semantic relations in other languages where lexicalization is expressed by different mechanisms. Thus we have to make clear distinction between the derivation as a literal relation (asymmetric, inverse, and intransitive) and the semantic oppositions between synsets for which the derivation itself might be a formal pointer.

6 Approaches to cover Slavic specific derivations in WordNet

There are several possible approaches for covering different lexicalizations resulting from derivation in different languages [7], [11]:

– to treat them as denoting specific concepts and to define appropriate synsets (gender pairs in Bulgarian and Serbian; relative adjectives in Bulgarian and Serbian);

– to include them in the synset with the word they were derived from (verb aspect in Bulgarian and in most of the cases in Serbian);

– to omit their explicit mentioning (diminutives in Bulgarian);
to provide source literals with flexion-derivation description encompass these phenomena as well.

Treating morpho-semantic relations such as verb aspect, relative adjectives, gender pairs and diminutives among others in Slavic languages as relations that involve language specific concepts requires an ILI addition for the languages where the concepts are presented (respectively lexical gaps in the rest). This solution takes grounds from the following observations:

– Verb aspect pairs, relative adjectives, feminine gender pairs and diminutives denote an unique concept;
– Verb aspect pairs, relative adjectives, feminine gender pairs and diminutives are lexicalized with a separate word in Bulgarian, Serbian, Czech and other Slavonic languages;
– Relative adjectives, feminine gender pairs and diminutives in most of the cases belong to different category or different inflectional class comparing to the word from which they are derived (there are some exceptions in the difference of the category, like diminutives that are derived from neuter nouns in Bulgarian).

Although the new WordNets do not compare yet with PWN’s coverage, the former are continuously extended and improved so that a balanced global multilingual WordNet is foreseen. For that reason the task of proper encoding of different levels of lexicalization in different languages is becoming more and more important in the view of the various Natural Language Processing tasks. The Slavic languages possess rich derivational morphology which has to be involved into the strict one-to-one mapping with the ILI.

7 Automatic building of derivational relations in Bulgarian and Serbian

The derivational relations for literals that already exist in WordNet can be interpreted in terms of derivational morphology, e.g., the noun teacher is derived from the verb teach and so on. WordNet already contains a lot of words that are produced by the derivational morphology rules: verbal nouns are linked with verbs, etc. In order to make explicit the morpho-semantic relations that exist already it would be necessary to include more links. On the other hand, a special attention has to be paid on the language specific derivational relations (some of them valid for big language families as Slavic languages). Several problems can be formulated following the observations and analyses presented in this study:

It is necessary to distinguish the pure derivation form the semantic relations which meaning is presupposed by the derivation itself. Concerning Bulgarian and Serbian WordNets this will be reflected particularly in the proper encoding of the derivational links between exact literals as it has been done in PWN; in the identification of derivational relations between literals already encoded in WordNets (comparing with PWN or exploiting language-specific derivational models), and in the introducing of language specific derivations in their appropriate place in the WordNet structure providing the exact correspondence with other languages.
In more general plan a theoretical investigation is needed to describe the nature of the semantic relations to which derivations are formal pointers. Ones a consistent classification is provided the respective semantic relations might be identified in the WordNets on the basis of the derivational ones in a particular language.

Several tasks may be done semi-automatically: to link literals instead of synsets with derivational relations; and to identify synsets where the potentially derivationally related literals appear. Bellow we provide some observations why the complete automation is not appropriate; although the derivational regularities are in most of the cases well established.

Although derivation is in many cases regular in the sense that it yields predictable results, it cannot be freely used for generation since it can lead to over-generation; namely, one could generate something which exists in a language system but does not exist in language usage. For instance, in Bulgarian and Serbian an abstract noun can be regularly derived (with a suffix –ост; –ost) from a descriptive adjective X meaning ‘the quality of something that has the characteristic X’, and a prefix (–не, –ne; –bez, –bez, etc.) can be used to produce both the adjective and a noun with the opposite meaning. One such example in Serbian is osećajan ‘be able to respond to affective changes’ → osećajnost ‘the ability to respond to affective changes’ → bezosećajan ‘not being able to respond to affective changes’ → bezosećajnost ‘the inability to respond to affective changes’. However, if the same pattern is applied to the adjective sličan ‘marked by correspondence or resemblance’ → sličnost ‘the quality of being similar’ → ?nesličan ‘not similar’ → ?nesličnost ‘the quality of being dissimilar’, two last lexemes in a sequence though easily understood are not lexicalized.

In a context of a WordNet production it is not sufficient to produce new synsets, for instance by applying the regular derivational mechanisms. It is equally important to place the generated synsets in the already existent network consisting of various relations. For instance, in Serbian the nouns sposobnost, vidljivost and popustljivost are regularly generated from the adjectives sposoban ‘having the necessary means or skill to do something’, vidljiv ‘having the characteristics that make it visible’ and popustljiv ‘easily managed or controlled’. However, the produced nouns have three different hypernyms: osnovna karakteristika {quality:1}, svojstvo {property:3}, and osobina {trait:1}. The correct placement of newly generated synsets in an existent network is not straightforward.

It has been noted (in section 5) that many senses of some words are distinguished by their different derivational capabilities. For instance, Serbian verb polaziti has five different meanings according to the Serbian explanatory dictionary, and one submeaning of the second presented meaning is ‘to go somewhere regularly and often to perform some duty’. That meaning is the only one from which the noun polaznik ‘someone who attends a school or a course’ can be derived by the agentive relation (realized by a suffix –ik).

It has already been stated in [8] that even derivation that seems very predictable can show very unpredictable behavior. Some derivational mechanisms in Bulgarian and Serbian are very predictable, like production of possessive adjectives that are produced from (mostly) animate nouns. As a consequence possessive adjectives are not listed in traditional Serbian dictionaries. The production of verbal nouns from imperfective verbs is also regular and produces a predictive meaning, the act of doing something. The verbal nouns are however, listed as a separate entries in Bulgarian and
Serbian dictionaries. Besides the predicted meaning they often acquire the additional meaning. For instance, the verbal nouns учение in Bulgarian, уčenje in Serbian and pečenje in Serbian are derived from imperfective verbs уча, уčiti ‘to study’ and пеци ‘to roast’. Besides the predicated meanings ‘the act of studying’ and ‘the act of roasting’ they have acquired in Serbian the additional meaning, ‘doctrine’ and ‘roast meat’, respectively. In the case of other derivational mechanism it can be more difficult to establish the meaning of the derived word. For instance, adjectives приличив and читлив in Serbian are derived respectively from the verbs прича ‘to talk’ and чита ‘to read’ using the same suffix –ив. Both verbs are imperfective and can be used both as transitive and intransitive: Марко прича ‘Marko tells the story’, Марко пuno прича ‘Marko speaks a lot’, Пuno лjudи чита knjigu ‘A lot of people read the book’, Марко пuno чита ‘Marko reads a lot’. The meaning of the adjective приличив is derived from the intransitive usage of a verb (namely, Марко пuno прича implies Марко je приличив ‘Marko is talkative’), while the adjective читлив is derived from the transitive usage (here Пuno лjudи чита knjigu implies Кnjига je читлива ‘The book is easy to read’).

The complexity of the issue of automation is best illustrated by the derivation of gender pairs in Serbian, since they exhibit all the previously mentioned problems. If we consider the derivation of female counterparts for the nouns of professions we encounter the following situations:
- The female counterpart morphologically does not exist: for instance, судија ‘judge’ is therefore used for both men and women;
- The female counterpart morphologically exists but is never used, војник ‘soldier’ vs. *војница and женa војник ‘(woman) soldier’;
- The female counterpart exists and is exclusively used for women performing that profession or function: келнер ‘waiter’ and келнерica ‘waitress’;
- The female counterpart exists but the male noun is also sometimes used for women: професор ‘(man or woman) professor’ and професорка ‘(woman) professor’;
- The female counterpart exists but it does not mean quite the same as a noun it was derived from: секретар ‘secretary’ is treated as someone performing an highly responsible function, as opposed to секретарica ‘(woman) secretary’ who is performing the low-level tasks in an organization;
- The female counterpart exists but it has acquired a different meaning, so it is not used to denote a woman performing certain function: саобраћај ‘traffic cop’ vs. саобраћајка ‘car accident’.

8 Conclusions and future work

We have briefly presented the current stage of the encoding of morpho-semantic relations in Bulgarian and Serbian WordNets. Grounding on the derivational properties of Slavic languages we provided some observations over the sophisticated nature of morpho-semantic relations and presented some examples proving the negative consequences from a purely automatic insertion of Slavic derivational relations into the WordNet structure. We believed we added additional evidences supporting the approach presented in [9] namely the utilization of a semi-automatic
identification or insertion of morpho-semantic relations. Such an approach would significantly facilitate the WordNet development although a manual connection on the basis of the automatically produced lists of suggested pairs has to be provided.

Further development of both Bulgarian and Serbian WordNets is narrowly connected with an investigation towards the theoretical grounds of the nature of morpho-semantic relations. At first stage the encoding of derivational relations between exact literals instead of synsets is foreseen. Another important task is the introducing of Slavic language specific derivations in a uniform way providing at the same time ILI correspondences. The accomplishment of these tasks will also reflect in the successful implementations of approaches based on cross-lingual information extraction, retrieval, and data mining, multilingual summarization, machine translation, etc.

References

Language Independent and Language Dependent Innovations in the Hungarian WordNet

Judit Kuti, Károly Varasdi, Ágnes Gyarmati, and Péter Vajda

Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest
1068 Budapest, Benczúr u. 33.
{kutij,varasdi,aagnes,vajda}@nytud.hu

Abstract. In this paper we present innovations that proved to be useful during the development of the Hungarian WordNet (HuWN). Part of these are language independent, but part of speech related expansions of the structure, which hopefully serve the accuracy of representation, as in the case of new relation types in the adjectival WordNet. Others seemed necessary because of the restricted applicability of the expand method to building the WordNet of a typologically different language than that of the model-WordNet, English. The Hungarian system of preverbs called for an expansion of the verbal structure in a way that it fits the characteristics of Hungarian verbs expressing aspect and Aktionsart. Treating verbs as eventualities and using the notion of nucleus introduced by Moens & Steedman we classified some of the Hungarian verbal synsets according to aspectual types and introduced new relations to the WordNet. This enabled us to represent both linguistically and psycholinguistically relevant pieces of information in the network.

Keywords: WordNet, verb, event structure, event ontology, aspect, Aktionsart

1 Introduction – the Hungarian WordNet

In the present paper we examine some of the problems encountered during the development of the Hungarian WordNet (HuWN), which, due to their language- or part of speech category specific nature called for different, alternative solutions than the ones offered by the WordNets serving as models for HuWN. The development of HuWN started along the lines of the general principle of the expand model.

However, this methodology presupposes the adaptability of a “source-WordNet” structure to another database of this kind, and with this, the basic compatibility of the internal semantic net of the lexicalised concepts in the two languages. It is, accordingly inevitable when dealing with typologically different languages that some

---

1 The basic model for the HuWN was the Princeton WordNet, but we have used some ideas of the GermaNet database, as well.
2 When talking about a specific WordNet of a given language, we refer to it with the widespread, trademark-like spelling, using capital ‘W’ and ‘N’, while when referring to the database type as to a common noun we use lower case letters.
3 The term was introduced by Vossen, [10] p.53.
complementary methodological steps be added or certain modifications be carried out in order to do justice to the linguistic characteristics of the “target” language. One of the most relevant characteristics of Hungarian, as opposed to English, from the point of view of building a WordNet, is that, through preverbs, the verb contains much information related to aspect and Aktionsart. When building a verbal semantic network, it is, thus, essential to examine the event structure of verbs. Accordingly, the first part of our paper shows what ways of storing and representing certain semantic relations stemming from the event structure of verbs we have worked out – within the framework facilitated by WordNet as a genre. In the second part of the paper we present new relation types introduced to the Hungarian WordNet on the basis of more general – not language-dependent – considerations. They concern the adjectival WordNet, and are meant to be alternative suggestions for the representation of some atypical adjective-clusters.

2 Language-dependent factors in structuring HuWN

As WordNets were originally designed to describe the hierarchical structure of nouns, and it is nouns that constitute a preponderant part of existing WordNets, one has to pay special attention to representing verbal relations in the given framework as accurately, as possible. For example the choice of two distinct names for relation types that can be considered equivalent – troponymy for verbs vs hyponymy for nouns – in the two respective parts of speech in PWN indicates already that a meaning representation framework for verbs cannot be solely designed on the basis of the existing grounds for a nominal hierarchy, not even in the case of a language like English, in which verbs as lexical units bear little or no information related to aspect or Aktionsart. With this in mind, in this section we first present some fundamental statements on event structure and aspectuality of verbs, and an elementary event-structure called nucleus introduced by Moens&Steedman. Subsequently, we hope to show that by using the notion of nucleus we acquire a means that enables us to

1. incorporate lexicalised meanings into WordNet more easily than was possible previously
2. represent psycholinguistically relevant pieces of information that were so far missing from the Hungarian WordNet
3. store information that proves to be useful for computational linguistic applications of the HuWN.
2.1 Eventualities and their aspectual properties

2.1.1 Logical implication between verbal meanings

It is necessary to examine in what way the relation of logical implication holds between verbs since this is what both the relations *troponymy* and *hyponymy* are based on. The propositions implied by a sentence are highly dependent on its aspect, illustrated by the following examples:

1. *Mari éppen ment át az utca túloldalára, amikor megpillantotta Jánost.*
   'Mary was crossing the street when she saw John.'
   'Mary crossed the street when she saw John.'

While sentence (1) does not imply that Mary actually crossed the street – she might have turned back to greet John, sentence (2) does imply that Mary did finish crossing the street (moreover, the pragmatical implicature suggesting that Mary crossed the street because she had noticed John, is also present).

The difference between the two main clauses in Hungarian is merely aspectual: the first one is in progressive, while the second one is in perfective aspect, each possessing a different logical potential.\(^4\) It is, thus, obvious that the question concerning what implications the preverb and verb as a whole can take part in is not separable from its aspectual value in the sentence. Although in Hungarian the actual aspect of a sentence is of course determined by many factors in the sentence besides the verb, its aspectual potential – as well as the sentences it can imply – is largely determined by the event structure of the verb.

In Hungarian some preverbs can bear information related to both aspect and Aktionsart. This alone might make Hungarian seem to be similar to Slavic languages. However, on the one hand, Hungarian does not express aspect in as a predictable manner, as e.g. Russian, whose WordNet we could have used as a basis for the Hungarian one, if the two languages had had enough similarities. On the other hand, aspect and Aktionsart in Hungarian are interwoven in a way that is unique among the languages that so far have been developed a WordNet for. The perfective aspect for example goes almost always hand in hand with one of the Aktionsart-types that are present in Hungarian (see [4], p.45.).

Furthermore, Hungarian has an extremely rich system of preverbs which can modify the meaning of the verb, making it inevitable when dealing with Hungarian, to consider aspectual characteristics as much as possible within the given framework. As already mentioned, the basic verbal relation, hypo- and hypernymy as well as troponymy, were elaborated based on the pattern of nominal relations, meaning that the WordNet-methodology requires that semantic relations between morphemes hold

\(^4\) When talking about aspectual properties of verbs, we should, in fact, be talking about verbal phrases, since verbs on their own are underspecified with respect to this kind of information, see [9].

\(^5\) The above phenomenon is known as the imperfective paradox, see [2].
through logical implications. While in the case of nouns one can show that N1 is a hyponym of N2 – by checking whether the pattern “it is true for each X that if X is an N1, then X is an N2” holds –, this is not possible for verbs, since one can only establish logical relations between propositions or the sentences expressing them, but the logical structure of sentences is determined by verbs together with their modifiers and complements. However, the verb-complement relation is highly asymmetrical: the logical potential of the sentence is determined by the verb; complements are only more or less passive participants. As the PWN, which has served as a basis for HuWN, does not contain aspectual information due to the lack of morphological marking of aspect in English, another way had to be found for representing typically occurring phenomena related to aspect in Hungarian.

A first framework of approaching aspect in general is provided by Zeno Vendler [8], who developed a type of event ontology in a way that would become useful for linguistic theory. This system was later elaborated by Emmon Bach [1], and worked out for computational linguistics by Marc Moens and Mark Steedman [6]. Drawing on Moens & Steedman's work we would like to suggest a way to structure aspectually related verb meanings in WordNet.

### 2.1.2 Aspectual classes according to Vendler and Bach

Vendler's classification of eventualities\(^7\) distinguishes between four aspectual classes according to the internal temporal structure of the event expressed by the verb. Together with their arguments and with context the four event types according to Vendler may take different aspects: activities (e.g. *swim*) typically take the progressive aspect, accomplishments (*go out of the room*) take both the progressive and perfective aspect, and achievements (*blow up*) take the perfective aspect. States take neither the progressive nor the perfective aspect. The classification as further developed and extended by Bach represents aspectual categories in a binary system, highlighting the existence of point expressions that are different from achievements (e.g. *click*). In Bach's terminology Vendler's accomplishments are called *protracted events*, achievements are called *culminations*, while *point* expressions are called *happenings*.

Vendler's four aspectual classes are also characterised by whether the interval of the event is divisible or not – i.e. whether the eventuality denoted by the verb holds for most of the subintervals, as well. Accordingly, of the four aspectual classes *activities* and *states* may be considered homogeneous eventualities, since they are expressed by predicates any sub-intervals of which may be described by the very same predicates.

---

\(^6\) In the case of verbs with direct object complements it is also the direct object that takes part in determining the aspect of the sentence. However, the impact of the direct object on the aspect can be relatively well predicted from the event structure of the verb and the properties of the object, so we do not have to specifically deal with this in the framework of the WordNet.

\(^7\) We are using the term *eventuality* after Bach, see [1].
Accomplishments and achievements on the other hand are in this respect coherent units of different kinds of heterogeneous event-components. Point expressions are also taken to be non-complex eventualities. From the point of view of constructing the Hungarian verbal WordNet it is the representation of complex eventualities – achievements and accomplishments – in a way that does justice to their aspecual properties that is of the greatest interest to us. One may interpret their complexity with the help of the so-called nucleus-structure introduced by Moens & Steedman.

2.1.3 The event-nucleus of Moens & Steedman

Moens & Steedman introduce a classification of eventualities relying on but further refining Vendler’s aspectual classes. Their central notion is that of an event-nucleus, which might be called a tripartite structure or triad, as well. The reason for the latter name is that an idealised eventuality consists of potentially three components belonging together: preparatory phase, telos/culmination and consequent state.

- preparatory phase
- telos
- consequent state

One may also represent the triad as an ordered triple \(< a; b; c >\) where \(a = \text{preparatory phase}, b = \text{telos}\) and \(c = \text{consequent state}\). Moens & Steedman place this idealised event-unit beyond the level of linguistically manifested lexicalised meanings. The components of the event-nucleus are thus filled with meta-linguistic and not with lexicalised linguistic elements.\(^8\) Treating the three nucleus-components\(^9\) as a unit may be justified as follows. When testing a lexicalised expression with linguistic tests sensitive to aspecual properties (in Hungarian the tests of the progressive and the perfective) the co-occurrence of no more than the three

---

\(^8\) Since we may only refer to these with linguistic elements, we will use small capitals so that they can be distinguished from italicised, linguistic elements

\(^9\) Here we are dealing with the event-components irrespective of whether they are lexicalised or not.
components outlined above may be shown. Since we are examining eventualities from an aspectual point of view, this fact must be considered relevant. We may, thus, acquire information about the aspectual properties of a verb expressing a certain eventuality by looking at which of the three event-components described above are conceptually present. On the example of the eventuality lexicalised with the verbal phrase *go out of the room*: The existence of the first component can be tested by looking at whether the expression can be put into the progressive. The existence of the third component, which practically goes hand in hand with the presence of the second one, can be tested by examining whether the expression can be put into the perfective (see [6]). Due to certain characteristics of the Hungarian language the easiest way we can test whether the second and third components of the triad are conceptualised is by translating the Hungarian sentence into English and putting the translated equivalent into Present Perfect / Progressive.10

3. János éppen ment ki az épületből, amikor találkoztam vele.
   ‘John was going out of the building when I met him.’

   ‘By the time Sue arrived, John has gone out of the building.’

As a result of the two tests we can see that the phrase *go out of the building* conceptualises all the three components of the triad:

<GOES TOWARDS THE GATE, PASSES THE THRESHOLD, IS OUTSIDE OF THE BUILDING>

Moens&Steedman elaborate on the categories named by Vendler/Bach by adding the factors of the existence or lack of the triad-components. In order to see how the classification according to the triad-components relates to the classification of Vendler/Bach, let us look at Table 1. This shows the classification of eventualities according to the factors taken into consideration by Moens&Steedman (+/- atomic and +/- consequent state), explicitly referring to the equivalents in Vendler's and Bach's system, where possible (in cases where the new terminology differs from the former one, we have indicated the latter in brackets).

---

10 Since this methodology may be surprising at first, some explanation is in order. In Hungarian – as opposed to English – there are no clear-cut and simple tests that are sensitive enough to the aspectual properties of a sentence (or verb phrase). Realizing that the impossibility of providing a usable test battery for Hungarian, we chose a detour, as it were, through a proxy in English. Benefitting from the situation that everybody in the WordNet developers’ team spoke English on an advanced level and had learnt to be sensitive to certain aspectual features in English, we decided to rely on our tacit knowledge of the aspectual features we wanted to test. When translating a Hungarian sentence into the English Present Perfect or Progressive, one had to judge its aspectual acceptability irrespective of whether the translation was correct in any other respect. Obviously, this methodological shortcut should be backed by further research in second language acquisition to be of sound theoretical value, but we believe that used with sufficient care it provides a reliable tool when the tests in the object language prove too complicated for practical usage.
Table 1. Eventualities based on the system of Moens&Steedman

<table>
<thead>
<tr>
<th></th>
<th>non-states</th>
<th>states</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>atomic</td>
<td>extended</td>
</tr>
</tbody>
</table>
| +conseq| culmination (=ACHIEVEMENT)        | culminated process (=ACCOMPLISHMENT) | resemble\,
|        | recognition,                      | build a house,       | understand\,
|        |                                   | eat a sandwich       | love\,
| -conseq| point                             | process              | know\,
|        | hiccup,                            | run, swim, walk      |                       |
|        | tap, wink                          | play the piano       |                       |

Theoretically $2^3$ different potential aspectual types may be distinguished according to the conceptual presence of the nucleus-components, listed as follows.\(^\text{11}\)

\[
<\emptyset, \emptyset, \emptyset> \quad <\emptyset, b, c> \quad <a, \emptyset, \emptyset> \\
<a, b, c> \quad <a, \emptyset, c> \quad <\emptyset, b, \emptyset> \\
<a, b, \emptyset> \quad <\emptyset, \emptyset, c> 
\]

The coherence of the nucleus components is more than mere temporal sequentiality, it is what Moens&Steedman call contingency "a term related, but not identical to a notion like causality" [6].

The mutual dependency among the three components of the nucleus means that none of them can be seen as preparatory phase, culmination or consequent state per se. An eventuality that, based on the above tests, seems to possess a preparatory phase, but lacks both culmination and consequent state (could be marked as \(<a;\emptyset;\emptyset>\)) cannot be seen as a preparatory process as it does not precede anything. By analogy an eventuality that, based on the above tests, seems to possess a consequent state but lacks a culmination (could be marked as \(<\emptyset,\emptyset,c>\)) cannot be seen as a consequent state, just like an eventuality with what seems to be a point of culmination, but lacking both preparatory phase and consequent state (could be marked as \(<\emptyset,b,\emptyset>\)) cannot be interpreted as a telos. In other words, a triad having a consequent state implies that the triad also has a culmination point. However, the three respective components seemingly appearing on their own may easily be interpreted as corresponding to the notion process and state as used by Vendler and to the Bachian point expression.

Although the three non-complex eventualities (process, point, state) are not discussed in depth further by Moens&Steedman, we found it important to deal with them in HuWN, and follow the above convention of showing the aspectual information in an ordered triple. Accordingly, the above listed possible combinations of the nucleus-components, each standing for one possible aspectual verb-subtype, are illustrated with examples, as follows:

\(^{11}\) The sign \(\emptyset\) refers to non-conceptualised components of the triad.
Three of the possible combinations are excluded based on epistemologic grounds:
(i) A nucleus having no components at all cannot be discussed neither conceptually nor linguistically. An eventuality (ii) having a preparatory phase and a culmination point, as well as one (iii) having a preparatory phase and a consequent state cannot be lexicalised due to the coherence of the telos and the consequent state.

Besides the remaining five lexicalised possibilities of nucleus-component combinations we have, however, seen the need for marking a sixth possible aspectual type in HuWN. As mentioned above, in many cases linguistic tests in Hungarian are unreliable in the sense that they provide ambiguous results even for native speakers. For the sake of usability in Hungarian language technology applications we considered it necessary to explicitly mark those cases in HuWN where the Hungarian test for the progressive did not result in a clearly grammatical sentence, but the English equivalent did. One such example can be seen in (5):

5. János éppen gyógyult meg, amikor huzatot kapott a füle és újra belázasodott.
    John was getting better when his ear caught cold and he got fever again.

In cases like the above mentioned we decided to mark the first component of the nucleus "unmarked", designating this with an x: <x,b,c>

2.2 The notion of the nucleus in HuWN

As we have seen, the conceptual presence or absence of meta-language elements beyond the lexicalised expressions can be tested with the help of Moens&Steedman's nucleus structure. The number of components a verb conceptualizes compared to an idealized complex event unit provides information on the telicity or atelicity of a given eventuality. If the third component of a nucleus denoted by a given verb is expressed\(^{12}\), the eventuality is telic, if the component is not present, the eventuality is atelic.

2.2.1 Representing telicity in HuWN

From the six mentioned possible patterns whose lexicalisation the presence of the respective nucleus-components enables it is only complex eventualities that can be

\(^{12}\) As mentioned in the previous section, the presence of the consequent state as third component entails the presence of the culmination point as second component.
telic. If we would like to get an overview of these complex eventualities from an aspectual point of view, the representation in ordered triples as introduced in 2.3. seems appropriate, as it can be seen in Table 2:

Table 2. Telicity of complex eventualities illustrated by the tripartite event structure of Moens & Steedman

<table>
<thead>
<tr>
<th>Compositions of the triad</th>
<th>The metalinguistic name for the conceptualised components of the phrase lexicalising the triad</th>
<th>Telicity of the VP</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;a, b, c&gt;</td>
<td>to exit: &lt;TO GO TOWARDS THE DOOR, TO STEP THROUGH THE DOOR, TO BE OUTSIDE&gt;</td>
<td>+consequent state telic</td>
</tr>
<tr>
<td>&lt;Ø, b, c&gt;</td>
<td>blow up: &lt;Ø, THE MOMENT OF BLOW-UP, BLOWN-UP STATE&gt;</td>
<td>+consequent state telic</td>
</tr>
</tbody>
</table>

Of the simple eventualities, processes and states are usually considered atelic while point expressions (on their own, without context) are underspecified for this kind of information. When constructing HuWN, the question arises whether and how to represent meanings that should be synonyms according to the notion of synonymy in WordNet and yet differ aspectually. The notion of the nucleus help us answer: aspectual differences can and should be represented in HuWN. If a meaning represented as a synset in the WordNet is transformed into a minimal proposition, one can determine whether the consequent state of the appropriate nucleus is present. By encoding whether a meaning has a consequent state (and hence a telos), through assigning to it one of the six conceptualization patterns of the triad components, the telicity of the eventuality expressed by the verb will be made explicit. This information is stored in HuWN in a similar way as in the case of the information on verb frames: we indicate which of the three triad components is conceptualised in Hungarian on the level of the literals.

As already introduced, for the sake of uniformity and transparency we follow the convention of showing the aspectual information in an ordered triple even in the case of simple eventualities mentioned in 2.2 and 2.3. Accordingly, the ordered triple of the verb fut ’run’ is (<a, Ø, Ø>). This triple shows on the one hand that the eventuality expressed by the verb fut is atelic, and on the other that it is a Vendlerian process, indicated by the preparatory phase being solely present.

2.2.2 Complex eventualities in HuWN

Besides the possibility of storing a minimal amount of aspectual information concerning the given literal in a verb synset, the relational structure of the WordNet and the nucleus taken as a single unit allow us to propose another extension to the
verb synset structure. In the case of complex eventualities whose certain triad components are not only conceptually present, but are *lexicalised*, as well, the unity of these components can be represented. Although the structure of PWN is based on a hierarchical system, an alternative structure has already been accepted for adjectives in PWN. By analogy it should also be possible to organise the verb synsets in a slightly modified way than nouns. The tripartite structure described above may be mapped onto the system of WordNet in the form of relations. The meta-language level described by Moens&Steedman's nucleus structure can be mapped onto the level of lexicalised elements, represented by WordNet synsets. The connection of the two levels is shown in Figure 3.

**Fig. 3.** Applying the event-nucleus of Moens&Steedman to the synsets of WordNet

Artificial nodes introduced in HuWN [5] are suitable for naming meta-language nuclei, e.g. the complex eventuality denoting the change of state from wet to dry, in the above example\(^\text{14}\). The relational structure of the WordNet allows introducing three new relations according to the respective triad-components being related to the meta-language nucleus-unit, represented by an artificial node. These new relations point to the appropriate artificial node and they are called *is_preparatory_phase_of*, *is_telos_of* and *is_consequent_state_of*, respectively, based on the names of the different nucleus components.

Meanings that are lexicalised by a single verb in English but not in Hungarian can thus be distinguished: the same meaning might be present in Hungarian both as a verb with a preverb providing more aspectual information and as a verb without a preverb, more underspecified for aspectual information. In the above example, the Hungarian

\(^{14}\) Artificial nodes are written with capital letters to distinguish them from natural language synsets.
szárad and megszárad synsets are both equivalent to the English {dry:2}\textsuperscript{15}. Without integrating the nucleus system into the WordNet, the synset megszárad could be placed into HuWN only as a hyponym of szárad, considering the originally available relations. However this kind of representation would not distinguish the different implicational relation between the above mentioned two meanings, but would merge them into a hyponym–hypernym relation.\textsuperscript{16} After having integrated the nucleus system into the HuWN, there is no need for an additional explicit relation between the components of a nucleus: they are already connected through the artificial node. Following the path of the relations \textit{is_preparatory_phase_of} and \textit{is_telos_of}, it is easy to determine that the synset szárad represents the preparatory phase of the nucleus whose another lexicalized component is megszárad, hence megszárad implies szárad while the implication does not hold in the other direction.\textsuperscript{17}

As we have seen, verbs belonging to the same triad (often with and without a preverb respectively) can be placed more accurately in HuWN with the help of the new relations. Furthermore, the relation \textit{is_consequent_state of} is not restricted to verbs, the third component of the triad mentioned above is the adjective synset száraz ({dry:1}). This psycholinguistically relevant piece of information is present in HuWN but would be lost if we had strictly held onto the structure of PWN without the tools for representing triads.

### 2.3 Possible applications

Besides the fact that one of the main tasks of a WordNet is to provide a uniform representation for the idiosyncratic properties of the lexical items, the extension of HuWN in the proposed way may bring practical benefits, as well. As we have seen, it can be easily deduced from the triad whether a given verb is telic or atelic, perfect or progressive, respectively. A Hungarian-English MT system may be improved by using this information provided in HuWN, e.g. in the area of matching the verb tenses in the source and the target language more appropriately. Since there are only two morphologically marked tenses in Hungarian (present and past), a rule-based MT system would select the same two tenses in the target language, simple present and simple past, respectively. Inaccurate translations would emerge inevitably. However, the above outlined information integrated into HuWN would improve the system. In Hungarian, for example, morphologically present tense forms of a telic verb have a future reference. The English equivalent of the Hungarian sentence \textit{Felhívom Pétert} is not \textit{I call Peter}, but \textit{I will call Peter}. Similarly, progressive past tense verb forms should be matched with the past continuous form of the appropriate verb, instead of selecting the simple past form: the Hungarian \textit{Péter az udvaron játszott} should be...

\textsuperscript{15} szárad 'is drying' (v), megszárad 'get dry' (v), száraz 'dry' (a)

\textsuperscript{16} By analogy to the nominal hypernymy relation, one way of conceiving of this relation between verbs would be basing it on selectional restrictions. E.g. the synsets \{hervad, fonnyad\} ('fade', 'wither') and \{rohad\} ('rot') would have such an ideal hypernymy relation, since the former selects plants as subject, while there is no such restriction on the subject of the latter one.

\textsuperscript{17} See Section 2.1.3 for a discussion on the connection (called contingency by Moens&Steedman) between the components of a triad.
matched to the English *Peter was playing in the yard*, instead of the expected *Peter played in the yard*. Aspectual information may be used in generating sentences, as well, let it be translation from English to Hungarian, or some other tasks requiring generation.

These properties of verbs may be helpful in text understanding, as well. The knowledge of these idiosyncratic properties of verbs is an important component of the inner representation of a computer. Without this information, just by considering the temporal adverbials (possibly) present in the sentence, it is not possible to accurately represent or reconstruct the temporal structure of a narrative.

3 Language independent new relation types in the HuWN

When constructing the HuWN, we tried to remain as faithful as possible to the structure of PWN 2.0 on which HuWN is based on. This was not always possible, as English and Hungarian show differences in their organising the lexicon and in some word-association tests. In what follows, we will describe two problematic cases we have faced, and for which we provide alternative solutions to the one suggested in PWN.

Some descriptive adjectives do not fit into the typical bipolar cluster structure of PWN. They occur in clusters having more focal synsets than the usual number, ie. more than two adjectives are meant to express opposing values of an attribute, see Figure 4.

![Figure 4. Atypical adjective clusters](image)

The focal synsets of these domains form a „triangle” along the *near_antonym* relations running between each pair among them. Considering this representation, it might be deduced that these attributes are not bipolar but are of 3 dimensions, having three marked "poles”. In the present section we argue for an alternative kind of representation, which, with the help of two new relations, enables adherence to the original bipolar structure of adjective clusters.

Descriptive adjectives are organised in clusters along semantic similarity and antonymy between words (instead of concepts), reflecting psychological principles [3]. Consider the example in Figure 5/b. The adjective pair *pozitív* ‘positive’-*negatív*
'negative' are the opposing poles of their domain. The situation of the word *semleges* 'neutral' is odd. Its English equivalent occurs as a third focal synset in the same domain as *positive* and *negative* in PWN. Relying on word association tests for Hungarian, we did not follow the solution of PWN when inserting *semleges* ('neutral') into HuWN. While the words *pozitív* and *negatív* do evoke each other in word association tests, the relation between *pozitív* and *semleges*, and *negatív* and *semleges*, respectively is not as straightforward. Although the word *semleges* does evoke *pozitív*, the antonym pair of *pozitív* is the adjective *negatív*. Loosening the scope of the usage of the relation *near_antonym* in order to enable antonym triangles to fit into a WordNet might cause anomalies in regular bipolar clusters as well (cf. direct and indirect antonyms). Therefore we have defined a new relation as an alternative to dealing with the case of triangles described above.

The adjectives *pozitív* and *negatív* determine a bipolar domain. This domain differs from the typical domains in the number and structure of its members. Apart from the two focal synsets, there is another adjective whose role is marked, but, as we have already shown, it is no real antonym of neither *pozitív* nor *negatív*. Furthermore, this special adjective expresses a value lying exactly in the middle of the domain. Therefore, the new relation we are proposing here, and we have already used in HuWN is called *scalar middle*, and points to both focals of the given domain (Fig. 5.).

Fig. 5. The middle relation

It should be noted that the newly introduced relation *scalar middle* can be used in any bipolar domain where the exact value (either being actually or considered conceptually as a discrete point) is lexically marked, e.g. in the domain determined by the adjectives *alsó-felső-középső* ('lower-upper-middle'). Although we have defined *scalar middle* in relation to HuWN, it may be used in other WordNets, as well, since the above described case is not limited to the Hungarian language alone.

At first sight the scalar *middle* relation could be used in the example shown in (5.a). The two opposing poles of the domain are {működő, aktív} 'active' and {kialudt} 'extinct, inactive', while the midpoint is denoted by {alvó, inaktív} 'dormant, inactive'. In this domain, however, the middle value of the attribute cannot be considered as discrete. Furthermore, the synset {alvó, inaktív} might be considered to be in *similar_to* relation with {működő, aktív}, as the adjective *alvó* 'dormant' refers to a "presently not functioning volcano", thus having a closer meaning to {működő, aktív}, just as *langyos* 'lukewarm' is in *similar_to* relation with *meleg* 'warm'.

The domain specified by these three synsets differs from the aforementioned domains not only because of the similarities and contrasts between its members.
These adjectives also constrain their scope: they can only refer to *volcanos*, and the WordNet has to account for this semantic relation. PWN and BalkaNet relate these adjectives through the use of the antonymy relation, and do not even indicate the relation with the noun exclusively modified by these adjectives from one point of view.

The synset-triple concerning volcanos is not the only triangle of this kind present in the semantic lexicon. For another simple example, we refer to the adjectives *egyényári-kényári-évelő* ‘annual, biennial, perennial’. Had we only the *near_antonym* relation at our disposal, the information that the respective adjectives can only refer to plants would have to be omitted, and the fact that these three adjectives belong together could indeed only be present in a triangle form among them.

When taking a closer look, one can see that the adjectives mentioned above *partition* the extension of the particular noun, i.e. they divide the set of nouns, e.g. all the plants in the last example, into disjoint subsets. This motivates the name of the suggested new relation: *partitions*, which is represented as a pointer pointing from the adjectives to the noun synset they partition (see Fig. 6.).

![diagram](image)

**Fig. 6.** The *partitions* relation

With the introduction of this new relation the explicit designation of the opposition between the adjective synsets becomes redundant, since due to the nature of the partitioning relation they may only be mutually exclusive. Although the *partitions* relation is similar to the *category_domain* relation of the WordNet, the two relations should not be confused. *Category_domain* relates the given adjectival meaning and the domain it can be used in, e.g.: {egyegyértékű, monovalens} ‘monovalent’ – {kémia, vegyészet} ‘chemistry’, but does not specify the noun(s) it can modify, even if it can modify a certain noun exclusively.

## 4 Conclusion

In the present paper we have tried to show on the example of the Hungarian WordNet in what ways the WordNet-structure as conceived of in PWN may be exploited and extended in order to represent some language-specific phenomena of typologically different languages than English. Although specifically implemented for solving a
linguistic situation in the Hungarian language, the implementation of the nucleus-structure in the WordNet in the form of relations might prove to be useful for other languages with a rich morphology showing aspectual distinctions, as well. Similarly, the use of the adjectival relations as suggested in the present paper enables us to represent semantic relations partly in a form that they remain faithful to some of the original ideas behind the WordNet-structure, while hopefully allowing for an equally accurate representation of semantic relation as with the originally offered alternative.

References

Introducing the African Languages WordNet

Juri le Roux1, Koliswa Moropa1, Sonja Bosch1, and Christiane Fellbaum2

1Department of African Languages, University of South Africa,
PO Box 392, 0003 UNISA, South Africa
{lrouxjc, moropck, boschse}@unisa.ac.za
2Department of Psychology, Princeton University, USA
fellbaum@clarity.princeton.edu

Abstract. This paper introduces the African Languages WordNet project which was launched in Pretoria during March 2007. The construction of a WordNet for African Languages is discussed. Adding African languages to the WordNet web will enable many natural language processing (NLP) applications such as crosslinguistic information retrieval and question answering, and significantly aid machine translation. Some of the accomplishments of the WordNet workshop as well as the main challenges facing the development of Wordnets for African languages are examined. Finally we look at future challenges.

Keywords: WordNet, African languages, Bantu language family, agglutinating languages, noun class system

1 Introduction

We discuss the construction of a WordNet for African Languages. Our main focus is on the challenges posed by languages that are typologically distinct from those for which the original WordNet design was conceived.

1.1 African Languages WordNet: Laying the Foundations

African Languages WordNets present an exciting addition to the WordNets of the world. NLP applications will be enabled not only for each of the African languages in isolation, but powerful cross-linguistic applications such as machine translation will be made possible by linking the African languages WordNets to one another and to the many global WordNets. Moreover, our initial investigations into the lexicons of different African languages suggest interesting similarities and differences among the

---

1 The African Languages WordNet effort, aiming to create an infrastructure for WordNet development for African languages, began with a week-long workshop funded by the Meraka Institute (CSIR) at the University of Pretoria in March, 2007. Christiane Fellbaum (Princeton), Piek Vossen (Amsterdam) and Karel Pala (Brno) facilitated. Linguists and computer scientists representing 9 official South African languages were introduced to WordNet lexicography and familiarized with the lexicographic editing tools DebVisDic (http://nlp.fi.muni.cz/trac/deb2/wiki/DebVisDicManual).
African languages. We work with the nine official African languages of South Africa, viz. isiZulu, isiXhosa, isiSwati, isiNdebele, Tshivenda, Xitsonga, Sesotho (Southern Sotho), Sesotho sa Leboa (Northern Sotho) and Setswana (Western Sotho). These languages all belong to the Bantu language family and are grammatically closely related. The Nguni languages, i.e. isiZulu, isiXhosa, isiSwati and isiNdebele form one group. The Sotho languages, viz. Sesotho, Sesotho sa Leboa and Setswana (Western Sotho) form another group with Tshivenda and Xitsonga being more or less on their own.

An important reason for distinguishing between these groups of languages lies in their distinct orthographies. In the case of the Nguni languages, a conjunctive system of writing is adhered to with a one-to-one correlation between orthographic words and linguistic words. For example, the IsiZulu orthographic word siyakuthanda (si-ya-ku-thand-a) ‘we like it’ is also a linguistic word. The Sotho languages as well as Tshivenda and Xitsonga on the other hand, are disjunctively written, and the above mentioned single IsiZulu orthographic word is written as four orthographic words in Sesotho sa Leboa, namely re a go rata ‘we like it’. These four orthographic entities constitute one linguistic word.

1.2 Language resources

A problem facing the African languages WordNet is the limited availability of electronic language resources such as large corpora, parallel corpora, electronic dictionaries and machine-readable lexicons, particularly in comparison to those available for other languages. For the moment we need to rely on both monolingual and bilingual dictionaries available in some African languages for semantic information. African languages are still lagging behind with regard to corpus compilation. General corpora are available for all nine African languages at the University of Pretoria, but with access restrictions which involve on site computer processing of the corpora and downloading only the results of the queries. The sizes of the various corpora range from 1 million tokens for isiNdebele to 5.8 million tokens for Sesotho sa Leboa [1].

2 Challenges

2.1 Morphological complexities

For the African languages we identify various specific challenges that have not been addressed by other WordNets. The first and foremost is that these languages are primarily agglutinative languages:

a) based on a noun class system according to which nouns are categorised by prefixal morphemes; and

b) having roots/stems as the constant core element from which words or word forms are constructed.
These morphological complexities make the notion of "words" and their lexicographic treatment particularly challenging when one tries to form synsets. Traditional lexicography does not reflect the linguistic intuitions of native speakers and does not withstand modern linguistic analysis as illustrated in the following examples representing the disjunctively as well as the conjunctively written languages. Entries in some traditional Setswana (Western-Sotho) dictionaries follow the alphabetical order of stems rather than the prefix that precedes the stem (a complete noun is formed by a prefix and a stem, a complete verb is syntactically formed by prefixes and/or suffixes around a root). The stem with the prefixes following it, where applicable, are presented in bold. The following is an example of an entry found under a in a Setswana dictionary:

ádimí, mo- ba- dev < adima, borrower, lender

This implies that "words" are listed under the first letter of their stems or roots except in those cases where prefixes have coalesced with the stems or roots, e.g.

mmútla pl mebútla, hare

Other Setswana dictionaries do, however, take the noun as it is as an entry, for example,

moadimi N. cl 1 mo-, SING. OF baadimi, DER. F. adima, a lender; a borrower.

Verbs are entered under the verb stem, e.g.

ádimā, borrow, lend
ádingwā pass < adima, be lent or borrowed

Adjectives are entered under the specific adjectival root but with all the noun class prefixes that it may take e.g.

ntlé. 1. adj dí-, bó- má-, gó-, lé- má-, ló- dí-, mó- bá-, mó- mé- or sé- dí-, beautiful, pretty, handsome

Adverbs do not present major problems since only a few adverbs are primitive, and do not show derivation from any other part of speech. The great majority are derivative, being formed mainly from nouns and pronouns by prefixal and suffixal inflexion. There are a considerable number of nouns and pronouns which may be used as adverbs without undergoing any change of form at all. Take the following entry for example,

ntlé, (ntlé), 1. adv in (i) fa ntlé, here outside; (ii) kwa ntlé, outside; 2. conj in kwa ntlé ga, besides or with the exception of; 3. n le-, outside; le- ma-, faeces (human)
The Nguni languages (isiXhosa, isiZulu, siSwati and isiNdebele) are particularly challenging, as they are written conjunctively; for example, basic noun forms are written together with concord morphemes. Entries in traditional dictionaries for these languages follow the alphabetical order of stems rather than the prefix that precedes the stem (a complete noun is formed by a prefix and a stem). The stem is presented in bold to distinguish it from the prefix. Below are example entries found under k in isiXhosa dictionaries:

- i-khaka (shield)
- isi-khumba (skin)
- u-khokho (ancestor -usually great-grandparent)

The three nouns begin with different prefixes i.e. i-, isi-, u- which denote different noun classes, but the common feature is the first letter of the stem, k. Verbs are entered with the infinitive prefix uku-, followed by the verb stem. WordNets for these languages will follow this pattern for representing synset members. As WordNets are organized entirely by meaning and not alphabetically, any look-up difficulties that this format poses for conventional dictionaries will not arise.

2.2 Roots, words, and word class membership

As with the original WordNet, African Languages WordNets will include only "open-class words", viz. nouns, verbs, adjectives and adverbials. For example, all Setswana words can feature as one of these categories because of the fact that most "closed-class words", such as pronouns, are in any case derived from these word classes. This standpoint may seem to create some problems but we will pursue it. Take the example of the root -ntlê [ntl], i.e. -ntlê and -ntlé, depending on the tone, it generally means 'beauty' and 'outside', therefore noun, adjective and also adverbial. One entry can be bontlé (beauty), a noun as:

- bontlé (beauty)
  Leina bontlé le na le bokao ba (Noun bontlé has the sense):
  1. bontlé - boleng bo bo kgatlhang (quality that pleases)
  Malatodi: maswê, mmê, mpe (Antonyms: dirt, ugly)

However, if we now try to enter this 'word' as an adjective, we start to encounter problems. If we enter it as a word, i.e. as a complete linguistic unit, we need to take into account that different prefixes can be taken by the root -ntlê, in fact, all the noun classes generate a prefix for this root and we must now decide what our entry should be. If we take the root as our entry, we need to indicate all the possible prefixes that it may take. Our entry can then be something like:

- ntlê (-ntlê)
  Modi wa letlhaodi -ntlê o na le bokao jwa (Adjective stem -ntlê has the sense):
1. yô montlê, ba bantlê, ô montlê, ê mentlê, lê lentlê, a mantlê, sê sentlê, tsê dintlê, ê ntlê, lô lontlê, bô bontlê, gô gontlê
   - boleng bo bo kgatlang (quality that pleases)

Malatodi: maswê, mmê, mpe
(Antonyms: dirt, ugly)

The user will now only get the word, for example yo montle, as part of the synset -ntlê and not as a word on its own. On the other hand, if we take every possible word, i.e. the combination of the root with the prefixes as our entries we will need to make 12 different entries for -ntlê (beautiful) as an adjective. Since the user encounters the complete form in speech/writing the latter option seems to be the most practical. However, because we are working with 'lexemes' per se, we need to find a way to make one entry to present -ntlê (beautiful) as an adjective and as an adverb. It therefore seems that we need to opt for the entry as -ntlê. We can then accommodate -ntlê as a noun and also as an adverbial. This can be done with an entry for bontlê, i.e. as a noun and also as an adverb sentlê. The entry for the lexeme ntlê can then be:

-ntlê (-ntlê)
  Modi wa letlhaodi -ntlê o na le bokao jwa (Adjective stem -ntlê has the sense):
  1. yô montlê, ba bantlê, ô montlê, ê mentlê, lê lentlê, a mantlê, sê sentlê, tsê dintlê, ê ntlê, lô lontlê, bô bontlê, gô gontlê
     - boleng bo bo kgatlang (quality that pleases)
  Malatodi: maswê, mmê, mpe
  (Antonyms: dirt, ugly)

Modi wa letlhalosi -ntlê o na le bokao jwa (Adverbial stem -ntlê has the sense):
  1. sentlê - ka mokgwa o o kgatlang e bile o siame (with quality that pleases and is right)
  Letswa: -ntlê
  (Derivative: -ntlê)

Leina bontlê le na le bokao jwa (Noun bontlê has the sense):
  1. bontlê - boleng bo bo kgatlang (quality that pleases)
  Malatodi: maswê, mmê, mpe
  (Antonyms: dirt, ugly)
  Letswa: -ntlê
  (Derivative: -ntlê)

For the adverbial it is now necessary to call it an adverbial root (modi wa letlhalosi), but since it is still the 'lexeme' which expresses the notion of 'beautiful', 'pretty', 'nice' and 'well' it can feature under the same entry.

The root -ntlê i.e. -ntlê, as a lexeme, expresses the meaning 'outside'. It is first and foremost adverbial but can take noun prefixes and be part of a noun. It can also be a conjunctive according to traditional dictionary entries. We will therefore form a synset for -ntlê (-ntlê), first as an adverb with the prefixes it takes in which we incorporate it also as a conjunct (adverbial), but then also as a noun with the prefixes it then takes. The entry can then be:
-ntlê (-ntlé)

Modi wa letlhalosi -ntlê ó na le bokao jwa (Adverbial stem -ntlê has the sense):
1. fa ntlê - gaufi le bokañntlê ga sengwe kgotsa lefelô (close to the outside of something or place)
   Malatodi: mo ñêng
   (Antonym: inside)
2. kwa ntlê - bokañntlê ga sengwe kgotsa lefelô (outside something or place)
   Malatodi: ka mo ñêng
   (Antonym: on the inside)
3. kwa ntlê ga - go tlogêla kgotsa go lesa mo go tse di leng teng (separate or leave from those present)
   Malatodi: le, e bile, gape
   (Antonyms: and, also, more)
Leina lentlê le na le bokao jwa (Noun bontlê has the sense):
1. lentlê - masalêla a dijô a a ntshetswang kwa ntlê (body waste matter)

For verbs the verb stem, i.e. the root plus the ending of the infinitive form, is taken as the lexeme. A typical entry for a verb can be:

-búa (speak)

Lediri go búa le na le bokao jwa (Verb go búa has the sense):
1. go búa - go dumisa mafoko ka maikaêlêlô a go itsese yo mongwe sengwe (utter words in order to let someone else to know something)
   Malatodi: go didimala, go tuulala
   (Antonyms: keep quite, to silence)

The infinitive form of the verb can also be used as a noun. An extra entry should therefore be made under -búa to tender for this possibility, viz.

Leina go búa le na le bokao jwa (Noun go búa has the sense):
1. go búa - ntlha ya go dumisa mafoko ka maikaêlêlô a go itsese yo mongwe sengwe (the uttering of words in order to let someone else know something)
   Malatodi: go didimala, go tuulala
   (Antonyms: keeping quite, silencing)

2.3 Deverbatives

As the formation of deverbatives (nouns from verb stems) is very common in the languages under discussion, we also need to include these elements under the verb stem since it forms part of the same lexeme. Since verb synsets are connected by a variety of lexical entailment pointers [2], the above entry can be extended by adding the deverbative forms as well, as illustrated in the following Setswana examples:
Introducing the African Languages WordNet 275

-búa
Lediri go búa le na le bokao jwa (Verb go búa has the sense):
1. go búa - go dumisa mafoko ka maikaëlêlô a go itsese yo mongwe sengwe (utter words in order to let someone else know something)
   Malatodi: go didimala, go tuulala
   (Antonyms: keep quite, to silence)

Leina go búa le na le bokao jwa (Noun go búa has the sense):
1. go búa - ntlna ya go dumisa mafoko ka maikaëlêlô a go itsese yo mongwe sengwe (the uttering of words in order to let someone else know something)
   Malatodi: go didimala, go tuulala
   (Antonyms: keeping quite, silencing)

Leina mmuí le na le bokao jwa (Noun mmuí has the sense):
1. mmuí - yo o dumisang mafoko ka maikaëlêlô a go itsese yo mongwe sengwe (person who utters words in order to let someone else know something)
   Letswa: -búisa
   (Derivative: -búisa)

Leina mmúisi le na le bokao jwa (Noun mmúisi has the sense):
1. mmúisi - yo o balang sengwe (reader of something)
   Letswa: -búisa
   (Derivative: -búiséwa)

Leina mmúisiwa le na le bokao jwa (Noun mmúisiwa has the sense):
1. mmúisiwa - yo go dumiswag mafoko ka maikaëlêlô a go itsese go ene (one to whom words are pronounced)
   Letswa: -búiséwa
   (Derivative: -búiséwa)

Leina mmúêlêdi le na le bokao jwa (Noun mmúêlêdi has the sense):
1. mmúêlêdi - yo o dumisang mafoko ka ntlna ya go buêlêla (one who utters words in order to speak for)
   Letswa: -búêlêla
   (Derivative: -búêlêla)

2.4 Heteronyms

It should also be noted that we need to make a separate entry for heteronyms, i.e. partial homonyms, which are totally different lexemes. In this case it is a difference in tone. The entry for the heteronym is then:

-bua
Lediri go bua le na le bokao jwa (Verb go bua has the sense):
1. go bua - go tlosa letlalo ka go dirisa thipa (to take of skin by using a knife)

Leina go bua le na le bokao jwa (Noun go bua has the sense):
1. go bua - nthla ya go tlosa letlalo ka go dirisa thipa (process of taking of skin by using a knife)

From the above it becomes clear that the Setswana WordNet will follow a pattern of using different morphological elements as entries, i.e. linguistic words for nouns, stems for verbs and roots for adjectives and adverbs. As WordNets are organized entirely by meaning and not alphabetically, this will not cause any look-up difficulties.

As stated above all Setswana words can be accommodated within WordNet's four prescribed word classes, i.e. nouns, adjectives, verbs and adverbs. All, so-called 'qualificatives', i.e. noun qualifiers, can be accommodated under nouns and adjectives. All verbal qualifiers can feature under verbs and adverbs. Some problems may be encountered, but none seems to be unsolvable.

3 Prior work

Traditionally, word categories in the Bantu languages are grouped together according to their function in the sentence and their grammatical relationship to one another. Cole [3] and Doke [4] distinguish six major categories for Setswana and IsiZulu respectively, viz. Substantives (Subjects and Objects), Qualificatives (Nominal modifiers), Predicatives (Verbs and Copulatives), Descriptives (Adverbs and Ideophones), Conjunctives (Connectives), and Interjectives (Exclamations). This classification is based on grammatical features and falls away the moment we consider language units in terms of lexemes.

Substantives and Qualificatives are all nominal and can feature under nouns and adjectives with the derived forms featuring under verbs and/or adverbs. The categories for verbs and adverbs will logically accommodate verbs and adverbs. Copulatives are non-verbal structures and need not be accommodated since their existence is based on structure.

Doke [5] proposed the term "ideophone" for a part of speech which describes a predicate, qualitative or adverb in respect to manner, colour, sound, smell, action, state or intensity. In contrast to the linguistic word in the Bantu languages, which is characterised by a number of morphemes such as prefixes and suffixes, as well as a root or stem, the ideophone consists only of a root which simultaneously functions as a stem and a fully-fledged word. This is illustrated in the following IsiZulu examples:

*Bathula bathi du*  "They kept completely quiet"

*Ingilazi iwe yathi phahla phansi*  "The glass fell smashing on the floor"

*Amanzi abomvu klubhu!*  "The water is as red as blood"
As lexemes, ideophones are descriptive of sound, colour, smell, manner, appearance, state, action or intensity. The majority of ideophones are therefore adverbial and will feature under the specific adverbial root. Where they indicate colour, for instance, they will feature under adjectives.

Most conjunctives are derived forms and will therefore feature under the category from which they are derived. Since we see most so-called conjunctives as adverbials anyway [6] only a few remain as true conjunctions in Setswana. These elements can however also be accommodated as adverbials since they generate some modifying meaning. By far the majority of interjectives are only nouns or pronouns being used vocatively or verbs being used imperatively. The rest are expressive of some aspect and can be accommodate under the specific lexeme expressing that particular aspect, e.g. cold.

4 Examples of African Languages WordNets

Examples of noun and verb synsets in isiXhosa and isiZulu are given below. Table 1 shows how the single concept expressed by English vehicle corresponds to two concepts (synsets) with distinct word forms in isiXhosa; the difference is revealed in the definitions, "wheeled vehicle" vs. "vehicle used for transporting people and goods." For each concept, hyponyms are grouped according to the mode of movement (vehicles that travel on air, road, water, rail, etc.). The second definition refers to a vehicle without wheels, with one example provided. Table 1 also shows meronyms (engine, tyres and wheels) linked to vehicle by the part-whole relation familiar from other WordNets.

Importantly, the isiXhosa and isiZulu synsets are connected to one another, so that corresponding words and synsets are given, allowing a direct comparison of corresponding and distinct concepts and lexicalizations.

Verb synsets are connected by a variety of lexical entailment pointers [2]. Tables 2 and 3 show the verbs corresponding to English walk and put along with their troponyms (manner subordinates). For example, -gxanyaza refers to walking in a certain manner (fast with fairly long strides). It is interesting to note that English also has many verbs expressing manners of walking, but they do not always match those in the African languages.

5 Software

African WordNets will use the editing tool DebVisDic, freeware multilingual software, designed for the development, maintenance and efficient exploitation of the aligned WordNets [7]. Initial experiments showed it to be well suited and adaptable to the construction of African Languages WordNets.
<table>
<thead>
<tr>
<th>ISIXHOSA</th>
<th>ISIZULU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ingwelo (vehicle)</strong></td>
<td><strong>isithuthi (vehicle)</strong></td>
</tr>
<tr>
<td></td>
<td>Def. 1. isithuthi esinamavili sokuthutha abantu nemphila (a vehicle with wheels for transporting people and goods)</td>
</tr>
<tr>
<td>Ezomoya (air)</td>
<td></td>
</tr>
<tr>
<td>ingvelomoya (aeroplane)</td>
<td>Isithuthi saseemoyeni (air)</td>
</tr>
<tr>
<td>ingvelontaka (helicopter)</td>
<td>indiza (aeroplane)</td>
</tr>
<tr>
<td>ijethi (jet)</td>
<td>ibhanoyi (aeroplane)</td>
</tr>
<tr>
<td>Ezendlela (road)</td>
<td></td>
</tr>
<tr>
<td>imoto / ikari (car)</td>
<td>Isithuthi saseemgaqweni (road)</td>
</tr>
<tr>
<td>ibhashi / udula-dula (bus)</td>
<td>isithuthuthu (motorbike)</td>
</tr>
<tr>
<td>itrakhi (truck)</td>
<td>imoto (car)</td>
</tr>
<tr>
<td>ilori (lorry)</td>
<td></td>
</tr>
<tr>
<td>iveni (van/bakkie)</td>
<td></td>
</tr>
<tr>
<td>itekisi (taxi)</td>
<td></td>
</tr>
<tr>
<td>ibhaysekile (bicycle)</td>
<td></td>
</tr>
<tr>
<td>isithuthuthu (motorbike)</td>
<td></td>
</tr>
<tr>
<td>Ezesiporo (rail)</td>
<td></td>
</tr>
<tr>
<td>ulolwe / uujuju (train)</td>
<td>Isithuthi sikajantshi (rail)</td>
</tr>
<tr>
<td>igutsi (goodstrain)</td>
<td>ingolovane (cocopan)</td>
</tr>
<tr>
<td>utramu (tram)</td>
<td>isitimela (train)</td>
</tr>
<tr>
<td>Ezasemanzini (water)</td>
<td></td>
</tr>
<tr>
<td>inganawa (ship)</td>
<td>Isithuthi saseemanzini (water)</td>
</tr>
<tr>
<td>inkwiwi (submarine)</td>
<td>umkhumbi (boat)</td>
</tr>
<tr>
<td>isikhephe (boat)</td>
<td>umkhumbingwenya (submarine)</td>
</tr>
<tr>
<td>Meronym (part-whole relation)</td>
<td></td>
</tr>
<tr>
<td>injini (engine), ivili lokuhuba (steering wheel), amavili (wheels),</td>
<td></td>
</tr>
<tr>
<td>Def. 2. isithuthi esingenamavili sokuthutha abantu nemphila (a vehicle without wheels for transporting people and goods)</td>
<td></td>
</tr>
<tr>
<td>isileyi (sledge)</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. IsiXhosa verb –*hamba*

<table>
<thead>
<tr>
<th>ISIXHOSA</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>-hamba</em> (walk)</td>
</tr>
<tr>
<td><em>-gxanyaza</em> (walk fast with fairly long strides)</td>
</tr>
<tr>
<td><em>-cotha</em> (walk slowly)</td>
</tr>
<tr>
<td><em>-khusa</em> (crawl)</td>
</tr>
<tr>
<td><em>-vantaza</em> (walk aimlessly)</td>
</tr>
<tr>
<td><em>-thwakuza</em> (walk aimlessly)</td>
</tr>
<tr>
<td><em>-nyalasa</em> (walk boldly)</td>
</tr>
<tr>
<td><em>-ndolosa</em> (walk proudly)</td>
</tr>
<tr>
<td><em>-khawuleza</em> (walk fast)</td>
</tr>
<tr>
<td><em>-qhiwalela</em> (limp)</td>
</tr>
</tbody>
</table>

Table 3. IsiXhosa and IsiZulu verb –*beka*

<table>
<thead>
<tr>
<th>ISIXHOSA</th>
<th>ISIZULU</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>-beka</em> (put)</td>
<td><em>-beka/ faka</em> (put)</td>
</tr>
<tr>
<td><em>-gcina</em> (keep)</td>
<td><em>-fulela</em> (put a roof on)</td>
</tr>
<tr>
<td><em>-londoloza</em> (save, e.g save money)</td>
<td><em>-ggoka</em> (wear)</td>
</tr>
<tr>
<td><em>-thwala</em> (put on)</td>
<td><em>-emboza</em> (cover)</td>
</tr>
<tr>
<td><em>-ngcwaba</em> (bury someone)</td>
<td><em>-thwala</em> (put on)</td>
</tr>
<tr>
<td></td>
<td><em>-endlala</em> (make)</td>
</tr>
</tbody>
</table>

6 Future work and conclusion

It should be clear from the above discussion that the intention with the African Languages WordNet is not to just produce another wordlist or a conventional dictionary based on grammatical analysis. With the lexeme and not the word as our point of departure we hope to produce a unique WordNet for African languages, driven by 'meaning' and not by copying existing WordNets. It is our belief that 'meaning' can only come from within the language and cannot be interpreted in terms of another language.

The long term aim of this project is the development of aligned WordNets for African languages spoken in South Africa (i.e. languages belonging to the Bantu...
language family) as multilingual knowledge resources which could be extended to include a wide variety of related languages from other parts of Africa. Such research and development would depend on the commitment of researchers to continue the work begun with great enthusiasm, the co-operation of numerous language institutions, the availability of a variety of language resources as well as further financial support following the seed research funding.

References

1. University of Pretoria Department of African Languages: http://www.up.ac.za/academic/humanities/eng/eng/afrilan/eng/initiative.htm
Towards an Integrated OWL Model for Domain-Specific and General Language WordNets

Harald Lüngen¹, Claudia Kunze², Lothar Lemnitzer², and Angelika Storrer³

¹ Justus-Liebig-Universität Gießen, ² University of Tübingen, ³ University of Dortmund
luengen@uni-giessen.de, kunze@sfs.uni-tuebingen.de, angelica.storrer@uni-dortmund.de, lothar@sfs.uni-tuebingen.de

Abstract. This paper presents an approach to integrate the general language WordNet GermaNet with TermNet, a German domain-specific ontology. Both resources are represented in the Web Ontology Language OWL. For GermaNet, we adopted the OWL model suggested by van Assem et al. [3] for the Princeton WordNet, for TermNet we developed a slightly different model better suited to terminologies. We will show how both resources can be inter-related using the idea of plug-in relations (as proposed by Magnini and Speranza 2002). In contrast to earlier plug-in approaches, our method of connecting general language WordNets with domain-specific terminology does not impose changes on the structure of these two types of lexical representations. We therefore consider our proposal to be a step towards the interoperability of lexical-semantic resources.

Keywords. WordNets; GermaNet; OWL; terminology

1 Introduction

WordNets (like the Princeton WordNet, cf. [1]) have been used in various applications of text processing, information retrieval, and information extraction. When these applications process documents dealing with a specific domain, one needs to combine knowledge about the domain-specific vocabulary represented in domain ontologies with lexical repositories representing general language vocabulary. In this context, it is useful to represent and inter-relate the entities and relations in both types of resources using a common representation language. In this paper we discuss an integrated representation model for domain-specific and general language resources using the Web Ontology Language OWL. The model was tested by relating entities of the German WordNet GermaNet to corresponding entities of the German domain ontology TermNet [2].

In Section 3, the main characteristics of these two resources are described. We built on the W3C approach to convert Princeton WordNet in RDF/OWL [3] and adapted them to GermaNet. For the domain ontology TermNet a different model was developed. The main classes and properties of both models are discussed in Section 4. The focus of this paper is on the question how the entities of the two OWL models — the model of the general language WordNet GermaNet and the model of the domain ontology TermNet — can be linked in a principled fashion. For this purpose, we defined
OWL-properties that relate entities of the two lexical resources, following the basic idea of the so-called plug-in approach by [4] for linking general language with domain-specific WordNets. Section 5 discusses the plug-in approach and our adaptation of it with reference to appropriate examples from GermaNet and TermNet. With our work, we aim at contributing to the emergent issue of interoperability between language resources.

2 Related Work

The work presented in this paper is inspired by the plug-in approach, which was developed in the context of ItalWordNet [5] and was originally proposed by [4]. However, rather than focusing on the processing aspects of the original method, in the present study we propose a declarative model of interlinking general language with domain-specific WordNets from the perspective of explicitly defined plug-in relations, which differ slightly from the ones proposed by Magnini and Speranza [4]. These relations allow for connecting specific terms with appropriate concepts, but do not modify the original resources and concepts.

Subsequent applications of the plug-in approach, like ArchiWordNet [6] or JurWordNet [7], implement plug-in relations for extending generic resources with domain terms from a processing perspective. The procedures lead to merged concepts and additional features being integrated into or added to the original databases.

De Luca and Nürnberger [8] describe an approach that relates an OWL representation of EuroWordNet to an OWL representation of domain terms. In their approach, terms are directly mapped onto synsets without any reference to intermediate relations. By defining distinct OWL plug-in properties our model aims to capture, in addition, different types of semantic correspondence between general language and domain-specific concepts.

In Vossen’s [9] approach, WordNet is adapted to the field and the needs of a specific organisation by extending it to include domain-specific vocabulary and removing concepts (and thus word senses) that are irrelevant for the organisation. In contrast, both the plug-in approach and the approach introduced in this paper are neutral with respect to the question whether a global ontology is extended by a specialised ontology or the other way around. Furthermore, the plug-in approach and the present approach do not address the question of how to automatically derive a domain ontology from a text collection; they are applicable to both automatically derived ontologies and hand-crafted ones. Moreover, Vossen’s [9] approach is procedural, meaning that its focus is on the specification of an extraction and integration algorithm, whereas the aim of the present paper is to declaratively model and specify the relational structure of the interface between a general and a domain-specific ontology in a formal language, i.e. the Semantic Web Ontology Language OWL.

3 Lexical and Terminological Resources

Our approach was developed and tested using an OWL model for a representative subset of the German WordNet GermaNet and an OWL model for the German termi-
nological WordNet TermNet. In this section we outline the main characteristics of GermaNet and TermNet.

3.1 Characteristics of GermaNet

GermaNet is a lexical-semantic WordNet for German which has been developed along the lines of the Princeton WordNet [1], covering the most important and frequent general language concepts and lexical-semantic relations holding between the concepts and lexical units represented, like hyponymy, meronymy and antonymy [10]. As is typical of WordNets, the central unit of representation is the synset, which comprises all synonyms or lexical units of a given concept. GermaNet presently covers more than 53 000 synsets with some 76 000 lexical units, among them nouns, verbs and adjectives. A basic subset of GermaNet (15 000 concepts) has been integrated into the polylingual EuroWordNet database [11]. The following features distinguish GermaNet from the data model of the Princeton WordNet, version 2.0:

1. The use of so-called artificial, non-lexicalised concepts, in order to achieve well-formed taxonomic hierarchies. For example, the artificial concept Schultyplehrer (‘school type teacher’) has been introduced to act as a hyper(o)nym of the lexicalised concepts Grundschullehrer (‘primary school teacher’), Realschullehrer (‘secondary school teacher’), Berufsschullehrer (‘vocational school teacher’) etc.;
2. Named entities are explicitly marked. Proper names in GermaNet primarily occur in the geographic domain;¹
3. In GermaNet, the taxonomic approach is also applied to the representation of adjectives, as opposed to WordNet’s satellite approach (based upon the notion of similarity with regard to different adjective clusters);
4. Meronymy is deemed a generic relation in GermaNet;
5. GermaNet verbs are provided with an exhaustive list of sub-categorisation frames and example phrases.

The data model of GermaNet is depicted in Fig. 1 as an entity-relationship graph. This model guided the conversion process of GermaNet objects and relations into XML elements and attributes.

¹ In version 2.1 of WordNet, however, over 7,600 synsets were manually classified as instances and tagged as such (cf. [12]).
3.2 Characteristics of TermNet

TermNet is a lexical resource that was developed in a project on automated text-to-hypertext conversion (cf. [13]). TermNet represents more than 400 German technical terms occurring in a corpus with documents in the domains “text-technology” and “hypertext research.” Most terms are noun terms, including multiword terms composed of a noun and an adjective modifier such as bidirektionaler Link (engl. ‘bidirectional link’). The entities and relations introduced for the Princeton WordNet [1] are fundamental for the structure of TermNet. The two basic entities of the TermNet model are terms (the analogue to word/lexical unit in the WordNet model) and termsets (the analogue to synsets in the WordNet model). Terms in TermNet are lexical units for which the technical meaning is explicitly defined in the documents of our corpus. Termsets contain technical terms that denote the same or a quite similar topic in different approaches to a given domain (cf. [14]). Terms are related by lexical relations, e.g. isAbbreviationOf, and termsets are related by conceptual relations, e.g. isHyponymOf, isMeronymOf. The data model of TermNet is illustrated by the ER-diagram in Fig. 2.

For automated hyperlinking, and probably for other applications, it is useful to know that term A occurring in document X denotes a category similar to the one denoted by term B occurring in document Y. Unlike other standards and proposals for representing thesauri (e.g. [15, 16, 17]), TermNet focuses on the representation of semantic correspondences between terms defined in different taxonomies or in competing scientific schools. Since competing taxonomies or schools may all have their benefits, we do not want to decide which terminology is to be preferred. Thus, the current TermNet model deliberately does not label terms as “preferred term.”

Since the entity type TermSet is crucial for the purpose of representing semantic correspondences between technical terms defined in competing schools, we want to explain the idea behind it using an example from German hypertext terminology: Kuhlen [18] and Tochtermann [19] both introduced a terminology for hypertext concepts that influenced the usage of technical terms in German publications on hypertext.
research. Both authors provide definitions for the concept hyperlink and specify a taxonomy of subclasses (external link, bidirectional link etc.). But Kuhlen uses the term Verknüpfung in his taxonomy (extratextuelle Verknüpfung, bidirektionale Verknüpfung) while in Tochtermann’s taxonomy the term Verweis is used (with subclasses like externer Verweis, bidirektionaler Verweis). The definitions of the concepts and subconcepts given by these authors are slightly different, and the two taxonomies are not isomorphic. As a consequence, in a scientific document on the subject domain, a term from the Kuhlen taxonomy cannot be replaced by the corresponding term from the Tochtermann taxonomy. After all, the purpose of defining terms is exactly to bind their word forms to the semantics specified in the definition. The usage of technical terms in documents may then serve to indicate the theoretical framework or scientific school to which the paper belongs. In our OWL model of TermNet, on the one hand we represent relations between terms of the same taxonomy, on the other hand we capture categorial correspondences between terms of competing taxonomies by assigning similar terms to the same termset.

4 OWL Models of GermaNet and TermNet

The Web Ontology Language OWL was created by the W3C Ontology Working Group as a standard for the specification of ontologies in the context of the Semantic Web. OWL comprises the three sublanguages OWL Light, OWL DL, and OWL Full, which differ in their expressivity. An ontology in the sublanguage OWL DL can be
interpreted according to description logics (DL), and DL-based reasoning software (e.g. RacerPro\(^2\) or Pellet\(^3\)) can be applied to check its consistency or to draw inferences from it. To take advantage of this, our OWL models of GermaNet, TermNet and the plug-in structure all remain within the OWL DL dialect.

Several approaches to convert PWN to OWL and to make it available for Semantic Web applications exist (e.g. [20, 21, 3]). In all these, the individual synsets and lexical units are rendered as instances of the OWL ontology. Although alternative modelling options have been discussed (cf. [22]), in the present project we adhere to an instance model as proposed by [3].

### 4.1 GermaNet OWL Model

In our OWL model, sets of GN concepts are represented as classes (<owl:Class>), while the properties of and relations between concepts are represented as OWL properties (<owl:ObjectProperty> or <owl:DatatypeProperty>) of these classes. For the two basic objects in the E-R-model of GN (Fig. 1), the classes Synset and LexicalUnit are introduced. Following the W3C model for PWN [3], we introduce NounSynset, VerbSynset, AdjectiveSynset, and AdverbSynset as immediate subclasses of Synset, as well as NounUnit, VerbUnit, AdjectiveUnit, and AdverbUnit as immediate subclasses of LexicalUnit.

#### Table 1. Features of OWL object properties for GermaNet

<table>
<thead>
<tr>
<th>Property</th>
<th>Domain</th>
<th>Range</th>
<th>Characteristics</th>
<th>Inverse Property</th>
<th>Local Restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>hasMember</td>
<td>Synset</td>
<td>LexicalUnit</td>
<td>inverse-functional</td>
<td>memberOf</td>
<td>POS-based</td>
</tr>
<tr>
<td>memberOf</td>
<td>LexicalUnit</td>
<td>Synset</td>
<td>functional</td>
<td>hasMember</td>
<td>POS-based</td>
</tr>
<tr>
<td>hasExample</td>
<td>Synset</td>
<td>Example</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conceptual Relations (CR)**

- `isHyperonymOf` Synset Synset transitive `isHyperonymOf` POS-based
- `isHyperonymOf` Synset Synset transitive `isHyponymOf` POS-based
- `isHolonymOf` NounSynset NounSynset - - -
- `isMeronymOf` NounSynset NounSynset - - -
- `isAssociatedWith` Synset Synset - - -
- `entails` VerbSynset VerbSynset - - -
- `causes` VerbSynset VerbSynset - - -

**Lexical-semantic relations (LSR)**

- `hasAntonym` LexicalUnit LexicalUnit symmetric `hasAntonym` POS-based
- `hasPertainym` LexicalUnit LexicalUnit - - -
- `isParticipleOf` VerbUnit VerbUnit - - -

---

\(^2\) cf. http://www.racer-systems.com

\(^3\) cf. http://pellet.owldl.com
Towards an Integrated OWL Model for Domain-Specific and…

For modelling the lexicalisation relation between synsets and lexical units, an OWL Object Property called `hasMember` with domain `Synset` and range `LexicalUnit` is introduced. For each POS-based subclass of `Synset` (e.g. `NounSynset`), a restriction of the range of `hasMember` to the corresponding subclass of `LexicalUnit` (e.g. `NounUnit`) is encoded using `<owl:allValuesFrom>`.

OWL is particularly well-suited to model the two basic relation types CR and LSR. Both types hold between internally defined classes and thus correspond to object properties in OWL. Like classes, properties can be arranged in a hierarchy in OWL using the `<rdfs:subPropertyOf>` construct. Our model thus contains two top-level object properties, `conceptualRelation` (with domain and range = `Synset`) and `lexicalSemanticRelation` (with domain and range = `LexicalUnit`). The OWL characteristics of their respective subproperties are shown in Table 1. Hypernymy, for example, is encoded as an `<owl:TransitiveProperty>` called `isHyperonymOf` with domain and range = `Synset`, as an immediate subproperty of `conceptualRelation`, and as the inverse property of hyponymy, cf. Listing 1.

Similar to `hasMember`, for each POS-based subclass of `Synset`, the range of `isHyperonymOf` is restricted to synsets of the same subclass. Relations that do not hold between internally defined classes, but in which a range in the form of an XML Schema data type like string or boolean is assigned to an internal class, are modelled as OWL datatype properties. In the case of GN, they are obviously the ones that are represented as ellipses in the E-R model of GN (Fig. 1). Table 2 contains a survey of datatype properties in the OWL model of GN with their respective domain, range and function status.

---

**Listing 1**: OWL Code for the introduction of hypernymy

```
<owl:TransitiveProperty rdf:about="#isHyperonymOf">
    <rdfs:subPropertyOf rdf:resource="#conceptualRelation"/>
    <rdfs:domain rdf:resource="#Synset"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <owl:inverseOf>
        <owl:TransitiveProperty rdf:about="#isHyponymOf"/>
    </owl:inverseOf>
</owl:TransitiveProperty>
```

---

**Table 2. Features of OWL datatype properties for GermaNet**

<table>
<thead>
<tr>
<th>Property</th>
<th>Domain</th>
<th>Range</th>
<th>Functional</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>Synset</td>
<td>“N”</td>
<td>”V”</td>
</tr>
<tr>
<td>hasParaphrase</td>
<td>Synset</td>
<td>xs:string</td>
<td>no</td>
</tr>
<tr>
<td>isArtificial</td>
<td>Synset ∪ LexicalUnit</td>
<td>xs:boolean</td>
<td>(yes)</td>
</tr>
<tr>
<td>isProperName</td>
<td>NounSynset ∪ NounUnit</td>
<td>xs:boolean</td>
<td>(yes)</td>
</tr>
<tr>
<td>hasOrthographicForm</td>
<td>LexicalUnit</td>
<td>xs:string</td>
<td>yes</td>
</tr>
<tr>
<td>hasSenseInt</td>
<td>LexicalUnit</td>
<td>xs:positiveInteger</td>
<td>yes</td>
</tr>
<tr>
<td>isStylisticallyMarked</td>
<td>LexicalUnit</td>
<td>xs:boolean</td>
<td>(yes)</td>
</tr>
<tr>
<td>hasFrame</td>
<td>VerbUnit ∪ Example</td>
<td>xs:string</td>
<td>no</td>
</tr>
<tr>
<td>hasText</td>
<td>Example</td>
<td>xs:string</td>
<td>yes</td>
</tr>
</tbody>
</table>
A subset of GermaNet (54 synset and 104 lexical unit instances including all conceptual and lexical-semantic relations holding between them) has been encoded in OWL according to the model presented above, using the Protégé ontology editor4. The GermaNet subset contains most of the candidate synsets for plugging in TermNet terms. Furthermore, this exemplary subset contains at least one instance of each conceptual and each lexical-semantic relation. We employed the reasoning software RacerPro to ensure its consistency within OWL DL. An automatic conversion of the complete GermaNet.5.0 is under way.

4.2 TermNet OWL Model

The complete TermNet in its OWL representation contains 425 technical terms and 206 termsets. In the OWL model we define all terms as classes, the instances of which are those objects in the real world that are denoted by the respective terms (e.g., an instance of the term externer Verweis is a concrete hyperlink in a hyperdocument compliant with Tochtermann’s definition of this term). Since we only account for nominal terms, all terms are subclasses of the superclass NounTerm. We use the <rdf:subClassOf> property to relate narrower terms to broader terms within the same taxonomy (e.g., we define Kuhlen’s term extratextuelle Verknüpfung as a subclass of his broader term Verknüpfung). By modelling terms as classes we benefit from the mechanism of feature inheritance related to the predefined <rdf:subClassOf> property. In addition, we are able to represent disjointness between classes using the OWL <owl:disjointWith> construct. By defining that the sets of instances denoted by the terms externer Verweis and interner Verweis are disjoint, we make sure that a link object in a document can only be assigned to one of these classes. In other words, a link object can either be an instance of the class externer Verweis or an instance of the class interner Verweis (although it may quite well be an instance of both externer Verweis and bidirektionaler Verweis). Terms of competing taxonomies that represent similar categories (like externer Verweis and extratextuelle Verknüpfung from the example in Sect. 3.2) are assigned to the same termset. For this purpose termsets are defined as subclasses of the superclass NounTermSet. Terms are assigned to termsets us-

Listing 2: OWL code for the assignment of terms to termset

```xml
<owl:Class rdf:ID="Verweis">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:ID="isMemberOf"/>
      </owl:onProperty>
      <owl:allValuesFrom>
        <owl:Class rdf:ID="TermSet_Link"/>
      </owl:allValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

4 http://protege.stanford.edu
ing the object property \texttt{tn:isMemberOf} (with \textit{NounTerm} as domain and \textit{NounTermSet} as range). The inverse property is \texttt{tn:hasMember}. Since termsets and terms are modelled as classes, we cannot simply adopt the definition of the \textit{gn:MemberOf} object property specified in the GermaNet OWL model (cf. Sect. 4.1.). Instead, we had to use the \texttt{<owl:allValuesFrom>} restriction to assign all instances of a term class to the respective termset class. Listing 2 illustrates how the term \textit{Verweis} is assigned to the termset \textit{Termset\_Link} (which comprises other terms like \textit{Verknüpfung}, \textit{Link}, \textit{Hyperlink}, \textit{Kante} etc.).

In addition to the taxonomic relations specified between terms of the same taxonomy by means of the \texttt{<rdf:subClassOf>} property, we also represent hierarchical relations between termsets, e.g. we want to account for the fact that all terms assigned to the termset \textit{TermSet\_Link} have a broader meaning than the terms assigned to the termset \textit{TermSet\_Monodirektionaler\_Link}. For this purpose, we defined the \texttt{tn:isHypernymOf}-Property, which relates termsets containing broader terms to termsets containing more specific terms. Its inverse property is \textit{isHyponymOf}. Listing 3 demonstrates how the more specific termset \textit{TermSet\_Monodirektionaler\_Link} is defined to be a hyponym of the broader termset \textit{TermSet\_Link} by means of the property \textit{isHyponymOf} and the \texttt{<owl:allValuesFrom>} restriction.

<table>
<thead>
<tr>
<th>Property</th>
<th>Domain</th>
<th>Range</th>
<th>Characteristics</th>
<th>Inverse Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{hasMember}</td>
<td>\textit{NounTermSet}</td>
<td>\textit{NounTerm}</td>
<td>inverse-functional</td>
<td>\texttt{isMemberOf}</td>
</tr>
<tr>
<td>\texttt{isMemberOf}</td>
<td>\textit{NounTerm}</td>
<td>\textit{NounTermSet}</td>
<td>functional</td>
<td>\texttt{hasMember}</td>
</tr>
</tbody>
</table>

Relations between termsets

<table>
<thead>
<tr>
<th>Property</th>
<th>Domain</th>
<th>Range</th>
<th>Characteristics</th>
<th>Inverse Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{isHyponymOf}</td>
<td>\textit{NounTermSet}</td>
<td>\textit{NounTermSet}</td>
<td>transitive</td>
<td>\texttt{isHyponymOf}</td>
</tr>
<tr>
<td>\texttt{isHyponymOf}</td>
<td>\textit{NounTermSet}</td>
<td>\textit{NounTermSet}</td>
<td>transitive</td>
<td>\texttt{isHyponymOf}</td>
</tr>
<tr>
<td>\texttt{isHolonymOf}</td>
<td>\textit{NounTermSet}</td>
<td>\textit{NounTermSet}</td>
<td>transitive</td>
<td>\texttt{isHolonymOf}</td>
</tr>
<tr>
<td>\texttt{isMeronymOf}</td>
<td>\textit{NounTermSet}</td>
<td>\textit{NounTermSet}</td>
<td>transitive</td>
<td>\texttt{isMeronymOf}</td>
</tr>
</tbody>
</table>

Relations between terms

<table>
<thead>
<tr>
<th>Property</th>
<th>Domain</th>
<th>Range</th>
<th>Characteristics</th>
<th>Inverse Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{IsAbbreviationOf}</td>
<td>\textit{NounTerm}</td>
<td>\textit{NounTerm}</td>
<td>isExpansionOf</td>
<td>\texttt{IsExpansionOf}</td>
</tr>
<tr>
<td>\texttt{IsExpansionOf}</td>
<td>\textit{NounTerm}</td>
<td>\textit{NounTerm}</td>
<td>isAbbreviationOf</td>
<td>\texttt{IsAbbreviationOf}</td>
</tr>
</tbody>
</table>

### Listing 3: OWL code for a hyponym relation between Termsets

```xml
<owl:Class rdf:ID="tn:TermSet\_Monodirektionaler\_Link">
  <rdfs:subClassOf rdf:resource="#tn:NounTermSet"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#tn:isHyponymOf"/>
      <owl:allValuesFrom rdf:resource="#tn:TermSet\_Link"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

Table 3. Features of OWL object properties for TermNet

The object properties \textit{isMeronymOf} and \textit{isHolonymOf} were introduced to account for part-whole-relations between objects denoted by the terms of two termsets. The property \textit{isAbbreviationOf} relates short terms to their expanded forms within the same taxonomy. Table 3 provides an overview of the properties defined in the OWL Term-
Net model. The `<rdf:subClassOf>` property between terms of the same taxonomy is not included in this overview because its semantics is predefined.

5 Representing Plug-in Relations in OWL

Providing domain-specific extensions for general language resources in order to capture and exploit the respective advantages of both resource types in natural language processing and semantic web applications has been discussed in the approaches by [9] and [4].

Vossen [9] describes a procedure to extract a hierarchy of terms (called “topics”) from a document collection, e.g. the set of all documents used in a specific organisation, and to subsequently combine it with WordNet. This is achieved by merging topics from the extracted hierarchy with matching WN concepts. The kind of matching criterion used is not specified; from the examples given one can assume that simple string matching is applied. Similar to the plug-in approach, one of the features of the resulting hierarchy is that the lower levels of the WN hierarchy and the possible upper levels of the terminological hierarchy are discarded. Vossen’s procedure only identifies plug-in synonymy and plug-in near synonymy, which are not differentiated in the new hierarchy.

The resulting hierarchy is subsequently trimmed by automatically removing those concepts that are irrelevant in the domain of the document collection, i.e. removing unwanted sense ambiguities that were introduced by the merger of the two resources. Finally, a procedure to fuse the compositional hierarchy with a so-called “private” or “personal” ontology, which apparently is a more domain-specific upper level ontology designed for the organisation and its document collection, is presented. For the fusing procedure, an “interface level” with matching concepts or topics from the source and target hierarchies seems to be externally defined, i.e. criteria other than string matching could potentially be applied. In this step, subtrees of the combined hierarchy are placed under the interface nodes of the private ontology; thus, it can be regarded as another instance of merging a global with a specialised ontology.

Whereas Vossen first builds ad-hoc terminologies from large document collections using information retrieval and term extraction methods and then links the resulting terms to WordNet synsets, Magnini and Speranza have proposed the plug-in approach which serves to link two (independently) existing resources of different types, namely the general-language ItalWordNet (IW)N and the specialised ontology ECOWN from the economic domain. Plug-in is a special instance of ontology merging, which is normally concerned with aligning resources of the same type.

Various kinds of plug-in relations serve to combine the relevant synsets of both resources. The plug-in approach yields a common hierarchy in which the top concepts of the specialised ontology are “eclipsed” while the subordinate concepts, the terms, are imported into the general language ontology. A relatively small number of instances of plug-in relations (269) suffices to integrate 4662 ECOWN concepts into ItalWordNet (cf. [4]).

ECOWN synsets are linked to a small domain ontology; 100 basic terms dominating relevant subhierarchies have been selected by experts due to relevance and frequency of use. The following scenarios of correspondences between IWN synsets and ECOWN terms are discussed:
Towards an Integrated OWL Model for Domain-Specific and… 291

1. Overlapping concepts: generic terms from the economic domain which also play a role in general language;
2. Overdifferentiation: a given ECOWN synset corresponds to more than one IWN concept, or an IWN synset corresponds to more than one ECOWN concept—these phenomena can be traced to different sense distinctions made by lexicographers vs. terminologists;
3. Gaps: for terms which have no general language counterpart, a suitable hypernym in the generic resource is selected.

The first scenario is captured by plug-in synonymy for overlapping synsets in IWN and ECOWN. A new plug-synset is created which replaces the corresponding IWN and ECOWN synsets in the integrated resource. This plug-synset takes its synonyms and hyponyms from the terminological resource and its hypernym from the generic resource. As a consequence, the terminological hypernym and the general language hyponyms are eclipsed.

The case of overdifferentiation is dealt with by plug-in near-synonymy. A new plug-synset is being created which also takes its hypernym from IWN and its synonyms and hyponyms from ECOWN.

In order to bridge the gap between IWN synset and ECOWN synset in the third scenario, plug-in hyponomy is applied. Two new plug-synsets are derived: one for the superordinate IWN synset (Plug-IWN) and one for the subordinate ECOWN synset (Plug-ECOWN). Plug-IWN takes its synonyms and hyponyms from IWN, and its hyponyms also include the Plug-ECOWN node. Plug-ECOWN relates to synonyms and hyponyms from ECOWN. Plug-ECOWN is assigned a new hypernym: Plug-IWN replaces the former hypernym from ECOWN.

The integration process is realised in four steps that centre around the plug-in relations. Thus, plug-in can be seen as a dynamic device with regard to merging two resources. The procedure yields new concepts, the plug-in concepts. The status of these merged plug-in concepts remains unclear—whether they constitute new lexical items, new terms or artificial concepts.

The plug-in approach has also been used and enhanced for Jur-WordNet [7] and ArchiWordNet [6], two domain-specific WordNet extensions. Jur-WordNet addresses theoretical considerations regarding common language versus expert language, and emphasises the citizens’ perspective on law terms, applying more or less the original plug-in relations. For ArchiWordNet, several plug-in procedures (substitutive, integrative, hyponymic and inverse plug-ins) are developed to replace or rearrange MultiWordNet hierarchies and integrate them with ArchiWordNet hierarchies. Furthermore, synsets may be enriched with terminological features, synonyms may be added or deleted from synsets, and relations may be added or deleted for specific synsets. Within this merging process, a lot of manual work specific to the resources in question had to be done which might possibly not be representative for any other pair of resources.

The plug-in approach offers an attractive model for linking TermNet to GermaNet, as both resources are also WordNet-based and of different coverage and specificity with a significant number of overlapping concepts. We primarily focus on modelling the relationships between general language and domain-specific concepts, and we use the plug-in metaphor for the relational model, less for the integration process. Thus, from our linking procedure, no new plug-in concepts evolve as the outcome of merging general language synsets with terms. The original databases, GermaNet and
TermNet, remain unchanged, but are supplemented with the relational structure provided by the established plug-in links.

As described in Section 4, in our OWL models, TermNet terms are modelled as classes and GermaNet synsets as individuals. Within OWL DL, a meta-class of term classes cannot be built, i.e. OWL classes cannot be declared to be OWL individuals without resorting to OWL Full. Thus, within OWL DL, the alignment can only be realised by restricting the range of a plug-in property to the individual that represents the corresponding GN synset. We distinguish three different linking scenarios between TermNet terms and GermaNet synsets:

1. Correspondence between a given TermNet term and a GermaNet synset, for example between the term `tn:term_Link` and the GermaNet noun synset `gn:Link`. The corresponding object property `plg:attachedToNearSynonym` has `tn:NounTerm` as domain and `gn:Synset` as range. By using an `<owl:Restriction>` over `plg:attachedToNearSynonym`, every individual of the class `tn:Term_Link` is assigned the individual `gn:Link` (see Listing 4). Since we do not assume pure synonymy for a corresponding term-synset pair, no synonymy link is established for plugging general language with domain language; the closest sense-relation being near-synonymy.

2. A TermNet term cannot be assigned a corresponding GermaNet synset but is a subclass of another TermNet term which in turn is linked to a GermaNet synset by `plg:attachedToNearSynonym`. For instance, the term `tn:Term_MonodirektionalerLink` stands in a subclass relation with the term `tn:Link`, which itself is linked to the GermaNet synset `gn:Link` by `plg:attachedToNearSynonym`. The property `plg:attachedToGeneralConcept` relates a term class like `tn:Term_MonodirektionalerLink` with a GermaNet synset which stands in a `plg:attachedToNearSynonym` relation with a superordinate term. Thus, a relation between indirectly linked concepts is made explicit and also serves to reduce the path length between semantically similar concepts for applications in which semantic distance measures are calculated. In this respect, we go beyond the scope of the plug-in approach which does not account for indirect links.

3. A TermNet term cannot be assigned a corresponding GermaNet synset, and, furthermore, no suitable hypernym for linking the term is available in the GermaNet data. But the term can be linked to a holonym concept in GermaNet, via the plug-in relation `plg:attachedToHolonym`. For example, the TermNet term `tn:Term_Anger

Listing 4: OWL code for a relation instance of attachedToNearSynonym

```xml
<owl:Class rdf:ID="tn:Term_Link">
  <rdfs:subClassOf rdf:resource="#tn:NounTerm"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:hasValue rdf:resource="#gn:Link"/>
      <owl:onProperty>
        <owl:ObjectProperty rdf:ID="plg:attachedToNearSynonym"/>
      </owl:onProperty>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```
(meaning ‘anchor’, i.e. a part of a link in the domain of hypertext research) has no semantic counterpart in GermaNet, but can be linked to the superordinate holonym in GermaNet, the synset gn:Link, by a plg:attachedToHolonym relation. This plug-in relation is unique to our approach and has not been derived from the original model.

Using the Protégé ontology editor and the reasoners RacerPro and Pellet, we encoded 150 OWL restrictions representing plug-in relations for plugging terms into the Synsets of the representative subset of GermaNet, 27 of which are plg:attachedToNearSynonym, 103 plg:attachedToGeneralConcept, and 20 plg:attachedToHolonym. In the actual integration of resources, the OWL construct <owl:imports> is applied to import both GermaNet and TermNet into the OWL file containing the plug-in specifications, while the original GermaNet and TermNet OWL ontologies remain unchanged and reside in their separate files. The integrated ontology is within OWL DL, and the reasoning software confirms its consistency.

For identifying the necessary plug-in relation instances, we adapted the basic concepts identification and alignment steps specified for the integration procedure in [4], using a correspondence list of GermaNet synset and TermNet term pairs which was derived on the basis of string matching. The remaining 377 TermNet terms will be linked when the complete GermaNet is available in OWL.

Applying this approach to integrating the residual TermNet terms or even further terminologies, we might possibly encounter terms without any corresponding hypernymic or holonymic concept in GermaNet. A complete alignment of both resources will yield the relevant number of instances regarding different plug-in relations and the number of concepts that cannot be linked by one of the three relations. The outcome will show whether the introduction of further types of plug-in relations is required.

Since we decided to model TermNet terms as OWL classes and GermaNet synsets as OWL individuals, the inverse relations of the plug-in relations cannot be defined within OWL DL, i.e. with Synset as their domain and the meta-class of all terms as range. This would however be a desirable feature of the model, even if drawing inferences is possible without it.

6 Conclusions and Future Work

Recently, the discussion about interoperability of language resources, including lexical resources of all kinds, has gained momentum. Interoperability issues are, for example, the focus of the newly-launched EU-project CLARIN (Common Language Resources and Technology Infrastructure, cf. www.clarin.eu). Interoperability issues include the development of standards for various kinds of resources. For WordNets and similar resources, the Lexical Markup Framework (LMF, [23]) is of utmost importance. True interoperability, however, is more than imposing format standards on resources. It should pave the way to merging and combining resources in the context of an application, even if they do not adhere to a common format standard, a requirement which often cannot be met. The plug-in approach, as we present it here, shows how lexical resources can be merged by a set of relations, while the resources themselves are left untouched. We will demonstrate that our approach can be applied to other terminological resources and WordNets.
The next steps in our research are the automatic conversion of the complete GermaNet into the OWL model presented above, and a completion of the definition of plug-in relation instances needed to connect TermNet to it. We will also implement and process a test suite of queries to the integrated ontology that are typical of text-technological applications such as thematic chaining and discourse parsing (cf. [22, 23], i.e. determining (transitive) hypernyms and calculating path lengths and semantic distances between synsets or units. In our approach, the merging of plug-in configurations and the pruning of the upper level of the specialised ontology as well as the lower level of the general ontology are deliberately shunned. Thus, if the effect of eclipse as described in [4] is desired, it will have to be produced by the query resolution procedure. However, we believe that this is the right place for it to go.

Another aspect of our work is worth mentioning. The aforementioned conversions of Princeton WordNet into an OWL format [20, 3, 21] convert synsets into OWL individuals. This is surprising both from a lexicographical and a terminological point of view. Synsets are assumed to represent concepts that are lexicalised by the lexical units which a synset contains. The conversion of a synset into an OWL class seems therefore more natural. For instance, the concept dog represents a class of animals, of which e.g. Fido is an instance. Arguably a conversion of synsets into instances is due to restrictions of the OWL-DL formalism and in particular of the tools which process OWL-encoded data. Sanfilipo et al. [25], for instance, deem the modelling of a larger amount of synsets as classes “impractical for a real-world application.” Elsewhere we have reported about experiments with an alternative modelling of GermaNet, in which synsets as well as lexical units have been modelled as OWL classes (cf. [22] for details).

We will therefore investigate how and with which consequences OWL classes such as the Synset class can be modelled as meta-classes, with the individual synsets being instances of this meta-class. Schreiber [26] pointed out the growing need for such an extension in the Semantic Web. Thus far, the definition of meta-classes was only possible within the dialect of OWL Full. Pan et al. [27] introduce a variant of OWL, called OWL FA, which provides a well-defined meta-modelling extension to OWL DL, preserving decidability. Still, the success of such an extension of OWL DL hinges on the availability of processing tools for this dialect of OWL. From our point of view, though, such an extension will facilitate linguistically more adequate representations of lexical-semantic and terminological resources. We will continue to investigate and to tap the potential of upcoming modelling standards.

References


The Possible Effects of Persian Light Verb Constructions on Persian WordNet

Niloofer Mansoory and Mahmood Bijankhan
University of Tehran, Linguistics Department
nmansoory@gmail.com, mbjkhan@ut.ac.ir

Abstract. This paper deals with a special class of Persian verbs, called complex verbs (CVs). The paper can be divided into two parts. The first part discusses the Persian verbal system, and is mainly concerned with the syntactic and semantic properties of Persian complex verbs or light verb constructions (LVCs). In the second part of the paper we have discussed the possible effect of these syntactic and semantic properties on Persian verb hierarchy and Persian WordNet.

Keywords: complex verbs, simple verbs, verb hierarchy, Persian WordNet

1 Introduction

Persian complex verbs have been discussed by a number of authors with respect to their syntactic and semantic properties. Some scholars have studied the phenomena within a syntax-based approach and have considered Persian CVs as elements, the syntactic and semantic properties of which are determined post-syntactically rather than in the lexicon. Among these scholars are Karimi [1], Dabir-Moghadam [2], Folli [3], Karimi-Doostan [4], among others. Some authors like Vahedi-Langrudi [5] have taken an in between approach. For example, Vahedi-Langrudi provides evidence for Persian CVs as \(V\) (lexical units) and \(V^{\text{max}}\) (verb phrases). With reference to recent approaches in the Minimalist Program, he suggests that Persian CVs be placed both in the lexicon and the syntax. In addition to theoretical linguists, researches in NLP systems have also been interested in Persian light verb constructions (LVCs). As Multiword expressions, Persian LVCs pose many problems in processing and generating Persian language since they have both lexical and phrasal properties. Megerdoomian [6], [7] discusses the issue from computational and NLP perspective and states that we will ignore the productive nature of CVs if we simply list them in the lexicon.

This paper is concerned with the lexical and computational aspect of the issue and proposes the possible effects of the compositional and idiomatic aspects of Persian CVs on Persian WordNet. The paper is organized as follows: Section 2 explains the

---

1 complex verbs, light verb constructions, compound verbs, and complex predicates are synonymously used in the literature to refer to Persian complex verbs.
Persian verbal system. Section 3 deals with the semantic connections between the light verb (LV) and the nonverbal (NV) elements of Persian LVCs. In section 4 we have elaborated the nature of the semantic connections illustrating the semantic regularities in LV uses of *zaadæn* (to hit). Finally in section 5 the possible effects of Persian LVCs on Persian WordNet are discussed.

### 2 The Persian verbal system

One of the significant characteristics of the Persian verbal system is its small number of simple verbs. Verbal concepts in Persian are mostly expressed by the combination of a nonverbal element (NV) and a light verb, the result of which is traditionally called “compound verb”. This process is very productive and as a result, the number of simple verbs is less than 200, which is very small in comparison to languages like English. Some of the Persian simple verbs can also function as LVs in LVCs (complex verbs). Examples include *zaadæn* (to hit), *kesændæn* (to do), *Xordæn* (to eat), *dadæn* (to give), *bordæn* (to take), and others. The preverbal (or nonverbal) elements in such constructions range over a number of phrasal categories which are usually nouns and sometimes adjectives, adverbs, and prepositional phrases (Karimi [1]; Karimi-Doostan [4]. Some examples are as follows:

\[
\begin{align*}
N + V \\
A) & \text{Pa zaadæn} & \text{Leg hit} & \text{‘To pedal’} \\
B) & \text{Email zaadæn} & \text{Email hit} & \text{‘Send an email’} \\
adj + V \\
C) & \text{lus kærdæn} & \text{louse doing} & \text{‘make louse’} \\
D) & \text{agah kærdæn} & \text{informed doing} & \text{‘to inform’} \\
adv + V \\
E) & \text{birun ændaxtæn} & \text{out throwing} & \text{‘to fire’} \\
F) & \text{bala keshidæn} & \text{up pulling} & \text{‘to advance’} \\
pp+V \\
G) & \text{æz dæst dadæn} & \text{of hand giving} & \text{‘to lose’} \\
H) & \text{be xatar aværdæn} & \text{to memory bringing} & \text{‘to remember’}
\end{align*}
\]

The most conflicting characteristic of Persian CVs is that they have both word-like properties and phrasal properties. As Karimi [1] and Goldberg [8] suggest, Persian CVs have one single word stress like simple words. Meanwhile, they undergo derivational processes that are typically restricted to zero level categories. Nevertheless, the majority of them can not be considered as frozen lexical elements since their NV and LV can be separated by some elements such as the feature auxiliary, negational and inflectional affixes and emphatic elements (Karimi, [1]; Müller [9] ; Folli, [3]).

---

2 As Karimi [1] mentions, the nonverbal element in Persian LVCs is not restricted to native Persian elements, and includes borrowed words.
Considering these and other syntactic features of Persian CVs, most of the authors suggest that LV and NV elements in Persian LVCs are separately generated and combined in syntax and become semantically fused at a different, later level (Karimi, [1]; Dabir-Moghaddam, [2]. From the semantic point of view, LVCs are traditionally listed as separate entries in Persian dictionaries since their semantic properties are the same as simple verbs. But the problem is that simply listing Persian CVs in the lexicon and adopting a lexical approach cannot explain their phrase-like properties and the productivity of the process in this language. As Folli [3] suggests Persian LVCs pose a more serious problem for lexicalist accounts in that it would essentially need to claim that Persian CVs are instances of idioms receiving a separate entry along with their syntactic structure. On the other hand, a nonlexicalist approach seems to have more capabilities in accounting for the syntactic freedom of NV elements and LVs in these constructions.

In the existing literature, among the authors who have explained the phenomenon from the syntactic perspective, Dabir-Moghaddam [2] suggest that some Persian CVs are the result of syntactic incorporation. On the contrary, Karimi [1] maintains that Persian CVs cannot be the result of syntactic incorporation suggesting that for the better description of such constructions, it would be more helpful to consider both their semantic and syntactic specifications. In the next section, we will discuss some semantic connections between the NV elements and LVs in Persian LVCs.

3 Some Semantic Properties of Persian CVs

The Persian CVs do not yield a uniform interpretation. They may receive either an idiomatic or compositional interpretation. From the LVCs which are compositional we cannot extract a clear pattern showing that they are fully compositional. In other words, the meaning of the whole is not directly derived from the meaning of the parts. Consider the following examples:

1) ræng zædæn paint hit ‘to paint’
   roghæn zædæn oil hit ‘to oil’
2) hærf zædæn words hit ‘to talk’
   færyad zædæn shout hit ‘to shout’
3) dæst zædæn hand hit ‘to clap’
   pa zædæn leg hit ‘to pedal’
4) zæng zædæn ring hit ‘to call’
   telefon zædæn phone hit ‘to call’
   email zædæn email hit ‘to send an e mail’
   fax zædæn fax hit ‘to send a fax’
5) lægæd zædæn foot hit ‘to kick’
   sili zædæn hand’s flat hit ‘to strike with the flat of the hand’
   chækosh zædæn hammer hit ‘to hammer’
All the examples above are a sample of Persian CVs containing a noun and the LV 
zađæn. The meaning of the verb zædæn as a simple verb in Persian is “to hit”. As the 
examples suggest, this verb, apart from its full verbal usage, can occur in LVCs. In 
LV uses of zædæn, where it cooccurs with different PV elements, it conveys new 
meanings that are not directly related to its simple verbal meaning. In (1) it means 
coating; in (2) it denotes doing, in (3) movement; in (4) sending, and in (5) the 
meaning of the LV zædæn is exactly similar to its full verbal meaning. The most 
conflicting examples are those in (6), where it seems impossible to postulate a 
meaning for zædæn.

The polysemious nature of the LV zædæn in the above examples is that some 
constructions are semantically opaque and idiomatic (6), some are compositional (5) 
and some are semi-compositional (1–4). Classifying Persian CVs, Karimi [1] suggests 
that most Persian compositional CVs can be considered as idiomatically combining 
expressions whose idiomatic meaning is composed on the basis of the meaning of 
their parts [1]. In this sense, we may consider most of the above semi-compositional 
examples as idiomatically combining expressions. However, the problem is that it 
would be counter-intuitive to analyze each of the constructions as a separate lexical 
entry in the lexicon, since it is possible to extract certain patterns from them. It is 
certain that giving an elaborate and detailed picture of such patterns requires an 
examination of a large set of data and adoption of an approach that can explain their 
compositional and productive properties. In this paper we have attempted to show the 
existence of these productive patterns.

4 The semantic patterns in LV uses of zædæn

In (1) zædæn is combined with two nominal elements, namely ræng (paint) and 
roghæn (oil). The two nouns belong to the same semantic class; that is, coating (in 
WordNet 2.0, paint and oil are under the same synset {coat, coating}). There are other 
examples in which zædæn combines with nominal elements with the same semantic 
feature as ræng and roghæn3:

7) kerem zædæn cream hit ‘to cream’
    shampoo zædæn shampoo hit ‘to shampoo’
    sabun zædæn soap hit ‘to soap’

In the examples above, the meaning of LV zædæn is {coat, cover} or {put on, 
apply}. This meaning is drastically far from its full verb meaning (hit).

---

3 It is noteworthy that in WordNet 2.0 coating is a kind of artifact. However, cream is not a 
coating but as an instrumentation under the concept artifact. Shampoo and soap are under 
none of these concepts, are kinds of substance.
Now consider examples (2). Both *hærf* (words or a brief statement) and *færyad* (out cry) are a kind of communication (as WordNet 2.0, also classifies their English equivalents as a kind of communication). In this semantic space, the common properties of the nominal elements seem to activate a different meaning of *zædæn*; that is, *doing*. Other nouns with the same semantic attributes may, also trigger the same meaning of *zædæn*. Examples are *naäre zædæn* (to rear) and *jigh zædæn* (to scream). The nominal PV elements in these constructions are also a kind of *auditory communication* (similar to their English counterparts in WordNet 2.0). Examples (3) show another type of semi-compositional constructions in which the nominal element seems to trigger a new meaning of the LV *zædæn*; that is, *movement*. Both *dæst* (hand) and *pa* (leg) are *external body parts*. The meaning of the whole constructions in such cases indicates a repeated movement of the organ involved. A similar example is *pelk zædæn* (to move the eye lids). The nominal elements in (4), except for *zæng* (in its literal sense as *ringing*), are a kind of *instrumentation* used for communication. Here, the meaning implied by the LV is *communicating via*, which is totally different from its full verb meaning. Among the nouns involved, *zæng* in the CV *zæng zædæn* (to call) can be interpreted as *phone* in its connotational sense (similar to the English verb *ring* meaning *call*). In this sense, it poses a problem for our analysis since it does not fall within the class of *instrumentation* as the other nominal elements in (4). Examples (5) illustrate a different pattern. Here, the light verb is synonymous to its full verb meaning (*hit*) and the preverbal nominal elements do not belong to the same semantic class. The meaning involved here is compositional in the sense that the meaning of the LVCs is semantically transparent and can be given by the sum of its parts.

The examples in (6) illustrate a completely different class of LVCs in which the meaning is idiomatic. Here, the nominal elements, like the examples in (5) do not belong to a specific semantic domain as they have no common semantic attribute. The most important property of these LVCs is that the meaning that the LV implies is not predictable so that the meaning of the whole construction can not be interpreted compositionally. Here, no semantic pattern emerges and as such the nominal elements involved can not be classified under a specific semantic group (unlike those in 1-4). So it seems plausible to consider them as totally frozen expressions and suppose them to be stored in the lexicon as individual lexical entries.

In this section, we illustrated the variation involved in the interpretation of Persian LVCs and CVs by studying some constructions resulting from the combination of the LV *zædæn* and a number of nominal PV elements. We revealed the existence of some semantic regularities in these constructions. Our analysis implies that the group of nominal elements with the same semantic property, when combined with the same LV produce a group of LVCs in which the meaning of the whole can be interpreted compositionally (1-5). In these constructions, the meaning that the verbal element or LV implies is predictable with respect to the semantic regularities among PV elements.

When there is no regularity among nominal elements or nominal elements do not have common semantic attributes (5 and 6), two different groups of LVCs result: (1)
LVs in which the meaning of the LV does not deviate from its full verb meaning and the interpretation of the CV takes place compositionally; (2) CVs in which the LV is semantically different from its full verb. This group of LVs are idioms and we can not find any semantic pattern in them.

In the next section we will discuss the possible effects of these findings on the structure of Persian WordNet.

5 The Possible effects of Persian LVCs on Persian WordNet

So far no serious attempt has been made by Iran's governmental centers or universities to build Persian WordNet. Some preliminary steps have been taken outside Iran (Keyvan, [10]) there is also a research done inside addressing the adjective hierarchy for Persian WordNet (Famiyan, [11]).

The present paper provides a theoretical basis for adopting an efficient strategy to build Persian WordNet. In order to decide between the Merge or Expand approach (or the combination of the two) which are applied in constructing WordNets for dozens of languages around the world, we found it reasonable to concentrate on the language specific properties of Persian in the first step. One of the most conflicting issues in the Persian verbal system, as mentioned in section 2, is the specific characteristics of its LVs. In order to construct the Persian WordNet, it was very important to answer a crucial question: should we consider all Persian CVs as frozen lexical elements and simply place them as individual lexical entries in Persian verb hierarchy? If not, It would be necessary to find a way to illustrate the productive nature of some semantic pattern in Persian LVs.

Studying a group of Persian LVs, a sample of which is presented in (3), we have classified the constructions under the three categories of compositional, semi-compositional and idiomatic. According to this classification, we propose that in building Persian verb hierarchy, in order to show the productive nature of the first two classes, we should connect each Persian LV with the specific class of nouns which trigger an embedded meaning in the LV, when combined with it. To do so, we can list every LV after its full verb meaning as an individual synset and define the meaning it implies in relation to the class of nouns triggering this meaning.

For example, remind the properties of the verb *zædæn* in Persian WordNet, after the synset which illustrates the full verb meaning of this verb, we can list other synsets defining other meanings for this verb, as it is used as a LV. One meaning is to put on or apply (on a surface), when it joins with the nouns under the concept coating. Then, under the synset we can list the CVs constructed by this pattern or link the LV to the nouns involved. The other meaning would be doing, when it joins with the nouns under the concept communication. In this case, too, we can list the available CVs after the synset or link the synset to the nouns involved. A similar procedure can be followed for other meanings mentioned before.

In the case of idiomatic LVs, we have no option but to list them as individual lexical entries in Persian WordNet.

This approach is comprehensive from both theoretical and practical perspectives. First, it attempts to illustrates the existing semantic regularities that contribute to the
productivity of Persian LVCs and predict the possible LVCs in this language. Second, because in the process of building Persian WordNet we have to work on a large amount of data, a comprehensive classification of Persian CVs will be done. Moreover, From the practical point of view, Persian WordNet will not be a mere copy of PWN and will present the features and properties specific to Persian. In this way, apart from the verb hierarchy, it is possible to have some language specific classifications of nominal concepts which can be categorized into different groups with respect to their relations with the verbal concepts in the process of constructing LVCs.

6 Conclusion

Building a WordNet for Persian requires a comprehensive study of this language. In this paper we discussed one of the properties of this language, namely LVCs. We introduced a classification of LVCs and proposed a new method in listing Persian CVs in an electronic database serving as Persian WordNet.

References

Towards a Morphodynamic WordNet of the Lexical Meaning

Nazaire Mbame

LRL, UBP Clermont 2, France
mbame@LRL.univ-bpclermont.fr

Abstract. We aim at conceiving a new form of semantic organisation that could help to build in the future what we entitled: Morphodynamic WordNet of a language like English. This new form of semantic organisation is consociationist and gestalist in contrast with the associationist one that continues to structure nowadays dictionaries. To illustrate, we are going to take the example of the lexical root “trench-” that the lexical items to trench, trenching, trencher, trenchspade, trenchcoat, trenchweapon, trenchknife, etc. repeat in their morphology. After the study of its semantic morphogenesis, we will draw up the corresponding schematic organisation that needs to be computed and implemented in the frame of the Morphodynamic WordNet project.

Keywords. Gestalt, Morphogenesis, categorization, Morphodynamic WordNet, semantic forms.

1 Introduction

By Morphodynamic WordNet, we intend a schematic representation in which lexical meanings and items derive one from another according to their natural causality. By Semantic morphogenesis (or Morphosemantic genesis) we intend, the generation process of lexical meanings by particularisation, differentiation, and categorial transposition. We are going to present in this paper the general aspects of our Morphodynamic WordNet project. We will first begin by presenting our theoretical framework, and then, will offer a descriptive example of application.

2 Theoretical Framework

Our representational model, we derive it from diverse sources. From the phenomenologist and gestalist theory of the concept as developed by Husserl, Gurwitsch, Merleau-Ponty, etc. According to this theory, the concept or noema comes from the sensible reality (objects and events) such as it is perceived and experienced through different points of view and facets. The same way this reality is perceived and experienced, the same way, the corresponding noema or concept is organised. So,
there is a kind of structural isomorphism between the concept and the corresponding experienced reality.

The other source of our inspiration is the *Theory of catastrophes* of R. Thom [1, 2] who, while talking about concept and signification, stipulated that the concept (and thus the signification) presents a “nucleus” around which we find a gestalt of its own deriving points of views, facets, forms, etc., which are its “presuppositions”. Between the “nucleus” and its “presuppositions”, we find structural genetic relationships. In the *Theory of Semantic forms* of Cadiot and Visetti [3] – in which the philosophical assumptions of phenomenology and gestalt are exploited for application in semantics, the same kind of functional organization of the concept reappears through the notions “motive” and “profile” the former denoting a kind of undifferentiated “nucleus”, and the latter, the different ways or forms by which this semantic nucleus appears to us, or can be perceived and intuitionized.

Without being subject to semiotisation by acoustic and graphic signifyings, concepts would be useless in linguistics. We find the basis of this semiotisation in the materialist and sensualist philosophy of language, in the way Cassirer [4] presents it in the *Philosophy of Symbolic forms, 1 language*.

3 Morphogenesis of the lexical meaning: example of the lexical root “trench-”

R. Thom [1, 2] defined Morphogenesis: generation (or destruction) of forms. We are going to see how this concept works in the semantic domain, especially in the generation of what we call semantic forms (uses, senses, etc.) of lexical items, which dictionaries just collect in the associationist way. A semantic form [3] is qualitative, individual and original by itself. Moreover, it is not isolated: it integrates a kind of system structured by genetic relationships. Morphogenesis of semantic forms can thus be defined as the generation process of lexical senses, meanings, uses, etc. It starts from a semantic nucleus semiotized by a lexical root, and yield semantic forms relating to this nucleus by proper reduction, particularization, differentiation and categorial transposition. The emerging semantic forms are often notified by lexical items repeating, in their morphology, the lexical root of the nucleus. For example, to trench, trenching, trencher, trenchspade, trenchcoat, trenchweapon, trenchknife, etc. are repeating the lexical root “trench-”. This is the material proof of the genetic relationships these lexical items share prior at the conceptual level. All these words are modifying the lexical root “trench-” and the corresponding nucleus in the frame of a dynamic process of categorization.

The semantic nucleus is phenomenological, descriptive and generic. Its nature should first be determined, before trying to see how its semantic morphogenesis (or morphosemantic genesis) works. As we said, the semantic forms (meanings, uses, senses...) generated by means of this morphogenesis are not isolated: they integrate a kind of gestalt, a system where they derive one from another in accordance with their natural causality.
3.1 The semantic morphogenesis of the lexical root “trench-” and its schematic representation

Let us take the example of the lexical notion “trench”. The free online dictionary lists for it at least 20 lexical inputs and uses that are, for example: to trench, trenching, the trench, trencher, trenchspade, trenchcoat, trenchweapon, trenchknife, etc. Each of these lexical items is individually introduced in its definition(s), without showing the semantic structural relationships they all share around the lexical root “trench-”. In our perspective, this lexical root gives access to a proto semantic nucleus that is continuously categorized by different ways.

So, what is this nucleus? A serious semiotic study could be done on that in the way Cassirer [4] suggests it in its Philosophy of symbolic forms, *The Language*, with the help of Indo-European roots [5]. Nevertheless, pragmatically, in comparison with the verb to trench, this proto nucleus seems to express a virtuous and undifferentiated idea of cutting the matter and taking it out to free a longitudinal opening with plane vertical sides... Husserl [6] introduced the method of the Phenomenological reduction, which consists in investigating, in a descriptive way, the perceptive noema (concept or the linguistic signified) in order to determine its actual and potential forms, structures, facets and points of view of (re)presentation. This method is going to help us in determining the semantic morphogenesis of the nucleus “trench-”. By means of morphogenesis (or conceptual categorization), semantic forms relating to this nucleus are generated. After, we will draw up the schematic representation of this morphogenesis that has to be computed and implemented for the needs of the Morphodynamic WordNet project.

The question that comes to us immediately is about the difference between the nucleus expressed by the lexical root “trench-” and the semantic form expressed by the verb to trench. In fact, objectively, the verb to trench considers the nucleus of the lexical root “trench-” from the point of view of its temporal anchorage. Besides, we find the lexical item trenching (act of trenching) that binds this nucleus into spatial praxis. Space and time are thus primitively particularizing the lexical root “trench-”.

In practice, the act of trenching - that is the act of cutting in the matter and taking it out to free a longitudinal opening... could be assimilated to the act of digging. This explains why, in some cases, to trench is defined as to dig nearby its other definition to cut. Digging implies cutting and these actions are “sequential parts” of the act of trenching. Digging and cutting are thus semantic features of this global action that, in some pragmatic contexts, the verb to trench can promote and express.

The action expressed by to trench sketches a pragmatic scenario where we find participants interacting. The actantial reduction of this pragmatic scenario yields individual lexical concepts that are: trencher (man who trench), trenching spade, trencher, etc. (instruments used in trenching). As we can notice it, the lexical item trencher is yet stabilising two meanings: the meaning of the intentional person who trench, and the meaning of the machine used in trenching (cutting, digging). These are semantic poles of its polysemy.

In praxis, the act of trenching needs to be direct, exact, calculated and strong. In the linguistic usage, the lexical expression to be trenchant transposes these qualities to human being in order to qualify him by their points of view. To be trenchant means for man, to be direct, exact, strong in judging. Qualities relating prior to the praxis of
trenching are then transferred to human being in order to point out some of its spiritual and intellectual abilities.

This categorial transposition of to be trenchant is viewed as a kind of particularization of the nucleus “trench-”. Here, transposition means: transferring properties of an object of a category A to an object of a category B by virtue of their similarities, analogy, etc. This gives justice to Cassirer [4] who claimed that, even the most abstract concepts are rooted in praxis, in sensible experience.

From the expression to be trenchant, derives conceptually and morphologically the adverb trenchantly (in the trenchant way), and the nominal trenchancy (the faculty of being trenchant, for a man). The related conceptual categorization goes along with the corresponding grammatical class changing. The principal outcome of the action of trenching is to free a longitudinal opening that the lexical item the trench denotes particularly. At its proper level, this notion specifies itself in natural trench, artificial trench; the former denoting a trench created by natural catastrophes, and the later, a trench created by human beings or animals. Natural/artificial are thus categorizing the concept of trench, as defined above. Furthermore, at their levels, natural trench / artificial trench can be subject to conceptual particularization. For example, we will find natural trenches located in land, and natural trenches located in sea, the fact of being located in land or in sea being the matter of the corresponding semantic refinement. The particularization process does not stop at these levels: it continues because we find different kinds of natural land trench (Attacama trench, for example) and different kinds of natural sea trench (Japan trench, Bourgainville trench). Along with that, we have Trenchtown that is a quarter of Kingston in Jamaica. This quarter wears the name trench because a natural land trench parts it.

We started from the generic notion trench (in the sense of a longitudinal hole), and we observed how the process of its dynamic categorization works relatively to its branching node natural trench. Relatively to its other branching node artificial trench, we have, for example, the categories draining trench (an artificial trench helping to drain water), military trench (an artificial trench dug for war), etc. Specially, we are going to focus on military trench, because its semantic morphogenesis is very wide and rich.

3.2 Military trench and its semantic morphogenesis.

The adjective military (of military trench) differentiates the generic notion artificial trench by binding it to the military context. This is the matter of its categorization. In English, many lexical items formed with the lexical root “trench-” are semantically motivated in this context. For example, slit trench categorizes military trench over the detail of its practical function of allowing the evacuation of soldiers. Slit trench is therefore a kind of military trench with specific properties and function.

During their stay in military trenches, soldiers were exposed to climate disturbances (snowing, raining, coldness...), and they needed adequate furniture for their protection. Sometimes, they could also contact diseases relating to their environment and life condition. About that, we find expressions like trench cap, trenchcoat originally signifying the cap or coat that soldiers used to wear in trenches to protect themselves from raining, coldness, etc. Nowadays, trenchcoat is totally free
from the proto military context, because it generally signifies a kind of overcoat that common people can wear to protect themselves from raining and coldness. The original properties and function of this equipment are kept. We also find trench fever, trench foot denoting specific illnesses that soldiers were able to contact during their stay in military trenches. We thus consider trench cap, trench coat, trench fever, trench foot as semantic forms relating to the generic notion military trench. They put into evidence and semiotise some of its potential points of view and facets. They are its immediate deriving semantics forms. Films where actors wear trenchcoats (or overcoats) are called trenchcoat films. Either, we find trenchcoat mafia referring to a gang of youngsters responsible of the Columbine killing in USA. This gang deserves the qualification trenchcoat because its members used to wear black overcoats. So trenchcoat films, trenchcoat mafia are particularizing the generic notion trenchcoat according to contexts of its application. They are its deriving categorial semantic forms.

As we said, the notion military trench gives access to a military context full with local particularizing points of view. For example, when a soldier died, he was buried in the trench in a way called: trench interment. In trenches, soldiers were using codes named: trench codes. Furthermore, the notion of trench warfare actualizes in praxis the military context that is just latent in military trench. Conceptually, trench interment, trench codes, trench warfare, are either deriving from the generic notion military trench. They are its categorial semantic forms.

When there is war, soldiers use arms and fighting strategies. The notions trench fire, trench raiding, trench weapons are apprehending the generic notion trench warfare from the points of view of arms used in it, of its fighting strategies, etc. They are consequently its deriving semantic forms. At its proper level, trench weapon categorizes itself in trench mortar, trench knife, trench gun, etc., which are its brands. Dictionaries also mention trench war that is a video game reproducing the trench warfare. Here, the concept of trench warfare is considered from the point of view of its possible reproduction into a video game. In relation with trench war, we find the expression trench trophy designating the trophy that the winner of a trench war video game obtains.

In the context of actual trench warfare, it could occur that, after an offensive that initially pulled soldiers out of their trenches, they move backward to these trenches for self-protection. This potential strategic retreat of soldiers is what the verb to entrench puts into evidence in the context of trench warfare. From this verb derives conceptually the nominal entrenchment. And by generalization of its meaning, we find some uses of to entrench in which it only means the act of withdrawing from an offensive position and hiding behind a protection that is no more just a trench, but that could also be a wall, for example.

In dictionaries, we find other usages of to entrench in which he promotes various ideas of digging, occupying a trench, fortifying, securing, etc. All these ideas are particularizing the verb to entrench according to points of view relating to its contextual applications and categorial transpositions. They are its deriving categorial semantic forms. For the rest, dictionaries also list expressions like: trench Schotty Barrier, trencher friend, trench effect, trencher cap, Trench mouth, etc. These notions are categorizing the nucleus “trench-” directly, or indirectly through some of its morphosemantic categories. Additional studies will be made to determine the points
of view they promote. After the complete study of the morphosemantic genesis of "trench-", we will then be able to reproduce its complete schematic representation as the following figure sketches it:
4 Conclusion

As shown above, the lexical root « trench- » is the genetic patrimony of all the lexical items which repeat it and modify at the conceptual and morphological levels. As stipulated by R. Thom [1, 2], *Semantic is thus a kind of genetics. The above systematic representation should be completed, computed and implemented in the frame of the project of creating a Morphodynamic WordNet of English. The aim of this Morphodynamic WordNet is to reproduce schematically the morphosemantic derivations of lexical meanings, uses, etc. in their natural way of processing. It should also precise the semantic links and catastrophes, which cause and support theses derivations. Moreover, the stable morphosemantic poles of this WordNet should be illustrated by specific linguistic examples concerning the individual meanings they fix. In this semantic domain, Morphogenesis generates *semantic forms by differential variations* [7] as from an undifferentiated semantic nucleus situated in the centre, and that lexical roots usually notify. This semantic morphogenesis is extendable and limitless.

References

Abstract. This paper presents a complete outline of the results of the Hungarian WordNet (HuWN) project: the construction process of the general vocabulary Hungarian WordNet ontology, its validation and evaluation, the construction of a domain ontology of financial terms built on top of the general ontology, and two practical applications demonstrating the utilization of the ontology.

1 Introduction

This paper presents a complete outline of the results of the Hungarian WordNet (HuWN) project: the construction process of the general vocabulary Hungarian WordNet ontology, its validation and evaluation, the construction of a domain ontology of financial terms built on top of the general ontology, and two practical applications demonstrating the utilization of the ontology.

The quantifiable results of the project may be summarized as follows. The Hungarian WordNet comprises of over 40,000 synsets, out of which 2,000 synsets form part of a business domain specific ontology. The proportion of the different parts-of-speech in the general ontology follows that observed in the Hungarian National Corpus and includes approximately 19,400 noun, 3,400 verb, 4,100 adjective and 1,100 adverb synsets.

In the following section, we describe our construction methodology in detail for the various parts of speech. In section 3., we present our validation and evaluation methodology, and in the last section we present the information extraction and the
word sense disambiguation corpus building applications that make use of the ontology.

## 2 Ontology construction

The development of the HuWN followed the methodology that was called expand model by [8]. Although this general principle seemed applicable in the case of the nominal, adjectival and adverbial parts of our WordNet, naturally, some minor adjustments to the language-specific needs were allowed as well. In the case of verbs, however, some major modifications were necessary. Due to the typological difference between English and Hungarian some of the linguistic information that Hungarian verbs express through preverbs − related to aspect and aktionsart − called for an additional different representation method of Hungarian verbs than the one in PWN. This new representation, together with some other innovations in the adjectival part of the Hungarian WordNet are described in detail in [2] and a separate paper submitted to the Conference.

A second principle we decided to comply with was the so-called conceptual density, as defined by [6]. This means that if a nominal or verbal synset was selected for inclusion in the Hungarian ontology, all its ancestors were also added to the ontology. This way the resulting ontology is dense, in the sense that it does not contain contextual gaps. This fact has the advantage that later extensions of the HuWN can be performed by further extending the important parts of the hierarchies, without the need of constant validation and search for gaps in the upper levels.

During the construction of the HuWN there were several work steps when the usage of monolingual resources was necessary: the Concise Hungarian Explanatory Dictionary (Magyar értelmező kéziszótár, EKSZ) ([4]), a monolingual explanatory dictionary, the Hungarian National Corpus ([7]) and a subcategorisation frame table of the most frequent verbs in Hungarian, developed by the Research Institute for Linguistics of the Hungarian Academy of Sciences.

The relation types that have been retained from the Princeton WordNet are hypernymy, antonymy, meronymy (substance, member and part), attribute (be_in_state), pertainym, similar (similar_to), entailment (subevent), cause (causes), also see (in the case of adjectives). Since the verbal relation indicating super- and subordination, called troponymy in PWN is called hypernymy in the version imported into the VisDic WordNet-building tool we have used, we have adopted the latter name. Some new relation types were also introduced, partly because of language specific phenomena – relations within the nucleus-structure have to be mentioned here – and partly due to other, language-independent reasons – two new relations introduced in the adjectival HuWN, scalar middle and partitions, represent the latter type of new relations. These are described in detail in a separate paper submitted to the Conference.

A primary concern when starting the ontology building was to provide a large overlap between the vocabulary covered by the Hungarian WordNet and other WordNets developed over the recent years. Accordingly, we have decided that we will take the BalkaNet Concept Set ([6]) (altogether 8,516 synsets) as a basis for the
expand model, and find a Hungarian equivalent for all its synsets, or state if the given meaning is non-lexicalised in Hungarian.

2.1 Nouns

2.1.1 Translation of the BCS and adding the LBC
We first implemented the nominal part of the BalkaNet Concept Set (BCS sets 1, 2 and 3 together), consisting of 5,896 Princeton WordNet 2.0 noun synsets.

First, we applied several machine-translation heuristics, developed earlier ([3]) in order to get rough translations for as many literals as possible. This comprised about 50% of all BCS synsets. These were then manually examined, corrected and extended with further synonyms using the VisDic editor. We also allowed for many-to-one and one-to-many mappings between the ILI and HuWN synsets. The BCS synsets that remained untranslated by automatic means were translated manually and processed in a similar way. The lexicographers also linked related entries from the EKSZ dictionary to as many synsets as possible, and added definitions, based on EKSZ definitions.

As a starting point, we adopted all the semantic relations among the synsets from PWN 2.0. After the translation of all the BCS synsets to Hungarian was complete, we manually checked all the adopted relations and modified the hierarchies according to specifics of Hungarian lexical semantics.

Following the EuroWordNet methodology, we then added our Local Base Concepts (LBCs), synsets for basic-level and important Hungarian concepts not covered by the common core of the BCS. For this, we used a list of most frequent nouns in the Hungarian National Corpus and those used most frequently as genus terms in the definitions of the EKSZ monolingual dictionary. For each of these, we identified the most frequent sense in the EKSZ, then identified the subset for which no references were made in the Hungarian BCS. For these, we created 250 additional synsets, which constitute the local base concepts for Hungarian.

2.1.2 Concentric extension based on the ILI
After the creation of the concepts of the Base Concept Set and the Local Base Concepts, we decided to extend the Hungarian nominal WordNet concentrically, considering in several iterations the direct descendants of the ILI projection of the actual Hungarian WordNet as candidates. This way, the conceptual density criterion was automatically satisfied during the expansion, and we added general concepts from the upper levels of the concept hierarchy (since we started with the Base Concept Set).

Regarding the fact that upper level synsets usually have more than one hyponym descendants, in each iteration we had to select the 1-2 thousand most promising candidates from 30-40 thousand available. We used four, not necessarily concordant characteristics for ranking:

Translation: The concept candidate was preprocessable with automatic synset translation heuristics ([3]). This way the creation and correct insertion of the concept
to the Hungarian hierarchy was easier to carry out, as one or more literals of the original English synset were available in Hungarian for the linguist expert.

**Frequency:** The concept had high frequency in English corpora (British National Corpus, American National Corpus First Release, SemCor). This usually indicates that the concept itself appears frequently in communication and thus adding it to the WordNet under construction was sensible.

**Overlap with other languages:** The candidate synset was conceptualized in WordNets for several languages besides English. This way we could maximize the overlap between Hungarian and foreign WordNets, that can be beneficial in multilingual applications like Machine Translation, and furthermore we could extend the ontology with such concepts that have been found useful by many other research groups as they added it to their own WordNet.

**Number of relations:** In the initial phases of the extension it made sense to take into account how many new synsets would become reachable by adding the one in question to the ontology. This way we could increase the number of candidates for later phases of the concentric extension.

### 2.1.3 Complete hierarchies for selected domains

As an additional extension method, we chose several domains for which all of the synsets in all of the hyponym subtrees in Princeton WordNet 2.0 were implemented in Hungarian. We did this to try to reach maximum encyclopedic coverage of the following areas:

- Geographic concepts and instances (countries, capitals and major cities, member states, geopolitical and other important regions, continents, names of important bodies of water, mountain peaks and islands)
- Human languages and language families
- Names of people
- Monetary units of the world.

We added 3,200 synsets based on these criteria.

### 2.1.4 Domain synsets

In order to enable the coding of domain relations for synsets to be implemented in the future, we translated all the PWN 2.0 category and region domain synsets. We also extended the set of region domain synsets with a collection of specific Hungarian region names.

We decided to neglect the Princeton WordNet usage domain relationships because of several inconsistencies observed in PWN (e.g. in some cases, the usage classification pertains to all literals in a synset, while in other cases it doesn’t.) Instead, we used a fixed list of our own usage codes, which could be applied individually to each literal using VisDic, providing a more flexible approach.

### 2.1.5 Proper names

National WordNets contain entity names among nominal synsets in a certain proportion. Among these are universal ones, like the world’s countries, capitals, world famous artists, scientists or politicians, and ones that are important for that certain nation/country.
We added a considerable amount of the named entities that were found most useful for the Hungarian WordNet, after the following processing steps:

- Standardization (format and character encoding)
- Selection (selection of categories to incorporate to the ontology and selection of instances for chosen categories)
- Extension (we collected different transliterations, synonyms and paraphrases of the selected entities)

2.2 Verbs and adjectives

In the case of verbs, after an initial phase of applying the expand method, it became obvious that the simple translation of English synsets with the same hierarchical relations between them would not result in a coherent Hungarian semantic network, even if local modifications are allowed. Consequently, we decided to make more extensive use of our monolingual resources, and tried to apply a methodology that would both satisfy the need for alignment with the standard WordNet (at least concerning the core vocabulary) and the need for a representation that does justice to language-specific lexical characteristics of Hungarian.

Lacking frequency data of verb senses, we started out from the frequency data of Hungarian verbal subcategorisation frames, which in Hungarian have specific enough syntactic information to be close to determining sense frequency. We included all the senses of the 800 most frequent Hungarian verbal subcategorisation frames in the Hungarian WordNet and made sure they had English equivalents, but also allowing for approximate interlingual connections (eq_near_synonym relation). If the equivalent of a Hungarian synset was found outside of the range of the BCS, the criterion of conceptual density was followed in all cases.

In order to achieve a more consistent hierarchy of HuWN, we decided that although the Hungarian synsets themselves should be connected to the PWN equivalent synsets, their internal structure should be developed independent of the English one.

In the case of adjectives, the translation of the BCS synsets proved not to present such problems, and concerned only approx. 300 synsets. Given that these were all focal synsets of different descriptive adjective clusters, we followed the expansion method: we added the respective satellite synsets to the translated focal ones, and, if this was necessary, added the antonym half-cluster as well. This work, however, included some minor adjustments, since the lexicalized antonym pairs and their satellite synsets are highly language-specific, which should be reflected in the ontology. Some more structural changes implemented in the adjectival WordNet concerned antonym clusters which were not centered around a bipolar scale, but which had three circular antonym relations ([1]).
2.3 Adverbs

Considering the ratio of the parts of speech observed in corpora, we decided to add about 1,000 adverbial synsets in addition to the synsets of the localized BCS, that did not contain any adverb synsets.

Because of the lack of adverbial sense frequency data for Hungarian, we decided to translate about 1,000 most frequent adverbial senses in PWN 2.0. In order to accomplish this, we first selected PWN synsets containing at least one literal that occurred at least once in that sense in the SemCor sense-tagged corpus. Next, we added up all the frequencies of all the surface forms of all the adverbs in the American National Corpus for each PWN 2.0 adverb synset, and selected synsets with a score of at least 1. The intersection of these two sets formed 1,013 adverbial synsets, which were automatically and manually translated and edited as outlined above.

We then carried out a number of revisions in order to adjust for Hungarian semantics and morphology:

- Separated and added senses for adverbs that have both time and place meaning.
- For adverbs of place, we identified the possible direction subgroups determined by case suffixes, and made each subgroup complete.
- Merged PWN synsets that could be expressed by a single Hungarian adverb sense.

2.4 The financial domain ontology

Besides the construction of general purpose language ontologies, developing domain ontologies for specific terminologies is important, since the vocabularies of general language ontologies are rarely capable of covering the specific language of a special scientific or technical domain. The financial domain ontology connected to the general HuWN ontology served as a basis for information extraction application, described in section 4.1.

We used two different approaches to add domain-relevant terms to the Hungarian WordNet. First, we made use of the high coverage of Princeton WordNet. By manual inspection, we located 32 concepts in PWN that we found to contain relevant terms in the domains of economy, enterprise and commerce. We added the 1,200 synsets that are in the hyponym subtrees of these domain top concepts.

As a second step, we examined a domain corpus consisting of short business articles and collected candidate domain-relevant terms from the text. Those that were not already in HuWN have been added as synsets, along with their synonyms to the ontology. The following table summarizes the distribution of the domain terms as observed in a corpus, over the different parts of speech:
Table 1.

<table>
<thead>
<tr>
<th>POS</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>2835</td>
</tr>
<tr>
<td>Adjective</td>
<td>270</td>
</tr>
<tr>
<td>Adverb</td>
<td>6</td>
</tr>
<tr>
<td>Verb</td>
<td>181</td>
</tr>
<tr>
<td>Overall</td>
<td>3292</td>
</tr>
</tbody>
</table>

3. Validation and Evaluation

3.1 Validation

In the final phase of the project, we focused on merging the parts of the ontology developed at the different project sites and performing several integrity and consistency checks, following [5]. The majority of the most frequent and serious problems were automatically identifiable with simple scripts and were then corrected manually. These included structural problems like:

- invalid sense ids
- same synsets connected with holonym and hypernym relations
- same synsets connected with similar to and near antonym relations
- duplicate synset ids
- duplicated relation between two synsets
- invalid characters in a literal, definition or usage example (character en-coding issues)
- invalid relation types (mostly typos)
- improper linking to the EKSZ monolingual explanatory dictionary
- lexicalized (non-named entity) synset with empty or missing definition/usage example
- mismatching part-of-speech tag and id suffix
- Hungarian local synset with missing external relation
- direct circles in hierarchical relations
- duplicate literals in synsets
- invalid relation (connected synset does not exist or has different POS than required)
- the same definition is used in more than one synsets

We also checked some semantic inconsistencies that required manual inspection of the database by linguist experts, without major computer assistance:

- central synsets of two adjective clusters connected with near antonym relation (we considered these as improper uses of near antonym relation and hanged these to also see relations)
- not reasonable sense distinctions: two synsets could be merged as they represented practically the same concept (here we collected synsets that shared several literals in common.)
3.2 Evaluation method

In order to assess the relevance of synsets added to the Hungarian WordNet, we evaluated random samples from the whole WordNet, from the Base Concept Sets and from the whole hyponym trees we incorporated to the Hungarian Ontology, and compared them to the synsets that received the highest rank during one of the concentric extension phases.

The evaluation was performed in the following way:

1. We generated a random sample of 200 synsets from the concepts we wanted to evaluate.
2. Two native Hungarian speakers independently evaluated the importance of synsets according to their usefulness in a linguistic ontology. They had to assign a score ranging from 1 to 10 to each concept. The higher value they assigned to the concept, the more relevant it was in their point of view. The agreement rate of the annotators leveraged to all the samples was 78.67% (considering the agreement to be 100% in case they assigned the same value to the synset in question and 0% if the difference between their scores was maximal).
3. We took the average of the scores assigned by the two linguists for each synset and then calculated the average and deviance of scores over the 200 element samples.

3.3 Results

The columns of the following two tables represent the segments of the ontology from which we generated the 200 synsets large samples. These were:

- **NONBCS**: the set of English synsets that are not among the base concept sets.
- **BCS1**: 1st Base Concept Set
- **BCS2**: 2nd Base Concept Set
- **BCS3**: 3rd Base Concept Set
- **CONC_1**: a random sample of synsets added during the first concentric extension phase
- **TREE**: a random sample of synsets that were added during the extension of Hungarian WordNet by whole hyponym subtrees
- **CONC_2_CAND**: a random sample of the candidates for the second concentric extension phase
- **LIT_FREQ**: top ranked synsets from the candidates for the second extension phase using frequency-based ranking
- **ILL_OVL**: top ranked synsets from the candidates for the second extension phase according to the number of foreign WordNets they appear in Table 3.
Table 2.

<table>
<thead>
<tr>
<th></th>
<th>NONBCS</th>
<th>BCS1</th>
<th>BCS2</th>
<th>BCS3</th>
<th>CONC_1</th>
<th>TREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4.51</td>
<td>6.56</td>
<td>6.21</td>
<td>5.03</td>
<td>5.71</td>
<td>4.21</td>
</tr>
<tr>
<td>Deviance</td>
<td>2.48</td>
<td>2.78</td>
<td>2.20</td>
<td>2.45</td>
<td>1.71</td>
<td>2.61</td>
</tr>
</tbody>
</table>

Table 3.

<table>
<thead>
<tr>
<th></th>
<th>CONC_2_CAND</th>
<th>LIT_FREQ</th>
<th>ILI_OVL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4.25</td>
<td>5.26</td>
<td>8.32</td>
</tr>
<tr>
<td>Deviance</td>
<td>2.27</td>
<td>1.74</td>
<td>1.25</td>
</tr>
</tbody>
</table>

As a summary we conclude that it is worthy to construct evaluation heuristics for the selection of synset candidates to extend WordNets with. Some heuristics clearly helped to incorporate more useful concepts to the ontology than adding synsets without considering their relevance.

4. Applications

4.1 Information extraction

Our information extraction engine was developed to identify the event type (such as sales, privatisation, litigation etc.) and the participating entities (eg. the seller, buyer and the price in a sale) expressed in short business news texts.

We created so-called event frame descriptions manually after analyzing our collected business news corpus. Each frame description defines an event, and contains participants in specific roles that correspond to the main verb and its typical arguments. In the implementation of the IE engine, a parser first identifies the main syntactic constituents in the input text, and then it tries to match these to the elements of the candidate event frames. There are several kinds of constraints that have to be satisfied for a match. Lexical constraints can either be specified as strings, or as synset ids corresponding to hyponym subtrees of the HuWN ontology. Semantic constraints are expressed by so-called semantic meta-features, or basic semantic categories, such as “human”, “company”, “currency” etc. that are mapped to HuWN synsets and all their hyponyms. There are also syntactic and morphologic constraints, which are checked against the output of the parser and the underlying morphologic analyzer. Finally, the IE engine ranks the candidate event frame matches for the output according to the ratio of event participants matched.

In this approach, the use of ontological categories allows for a simpler and easier to understand layout of the event frames. The main advantage of the use of synset ids
and semantic types (as opposed to bare lexical listings) lies in the fact that the vocabulary of the IE engine can be easily customized and extended by adding new concepts to the ontology, without the need to modify the original event frame descriptions.

4.2 Creating an annotated corpus for WSD

In parallel with the construction of the ontology itself, we selected 39 words that had several commonly used senses and built a lexical sample word sense disambiguation corpus for Hungarian. This corpus is freely available for research and teaching purposes and consists of 350–500 labeled examples for each polysemous lexical item. The sense tags were taken from the synset ids of the senses of the polysemous words in HuWN.

The corpus follows the SensEval lexical sample format in order to ease its use for testing systems developed for other previous lexical sample datasets. The annotation was performed by two independent annotators. The initial annotation had an average inter-annotator agreement rate of 84.78%. Disagreements were later on resolved by consensus of the two annotators and a third independent linguist. The most common sense covers 66.12% of the instances and an average of 4 further senses share the remaining percentage of the labeled examples.

References


2 Please contact the authors for information about obtaining the WSD corpus.
Synset Based Multilingual Dictionary: 
Insights, Applications and Challenges

Rajat Kumar Mohanty¹, Pushpak Bhattacharyya¹, Shraddha Kalele¹, 
Prabhakar Pandey¹, Aditya Sharma¹, and Mitesh Kopra¹

¹ Department of Computer Science and Engineering 
Indian Institute of Technology Bombay, Mumbai - 400076, India 
{rkm, pb, shraddha, pande, adityas, miteshk}@cse.iitb.ac.in

Abstract. In this paper, we report our effort at the standardization, design and partial implementation of a multilingual dictionary in the context of three large scale projects, viz., (i) Cross Lingual Information Retrieval, (ii) English to Indian Language Machine Translation, and (iii) Indian Language to Indian Language Machine Translation. These projects are large scale, because each project involves 8-10 partners spread across the length and breadth of India with great amount of language diversity. The dictionary is based not on words but on WordNet SYNSETS, i.e., concepts. Identical dictionary architecture is used for all the three projects, where source to target language transfer is initiated by concept to concept mapping. The whole dictionary can be looked upon as an M X N matrix where M is the number of synsets (rows) and N is the number of languages (columns). This architecture maps the lexeme(s) of one language—standing for a concept— with the lexeme(s) of other languages standing for the same concept. In actual usage, a preliminary WSD identifies the correct row for a word and then a lexical choice procedure identifies the correct target word from the corresponding synset. Currently the multilingual dictionary is being developed for 11 languages: English, Hindi, Bengali, Marathi, Punjabi, Urdu, Tamil, Kannada, Telugu, Malayalam and Oriya. Our work with this framework makes us aware of many benefits of this multilingual concept based scheme over language pair-wise dictionaries. The pivot synsets, with which all other languages link, come from Hindi. Interesting insights emerge and challenges are faced in dealing with linguistic and cultural diversities. Economy of representation is achieved on many fronts and at many levels. We have been eminently assisted by our long standing experience in building the WordNets of two major languages of India, viz., Hindi and Marathi which rank 5th (~500 million) and 14th (~70 million) respectively in the world in terms of the number of people speaking these languages.

Keywords: Multilingual Dictionary, Dictionary Standardization, Concept Based Dictionary, Light Weight WSD and Lexical Choice, Multilingual Dictionary Database
1 Introduction

In any natural language application, dictionary look-up plays a vital role. We report a model for multilingual dictionary in the context of large scale natural language processing applications in the areas of Cross Lingual IR and Machine Translation. Unlike any conventional monolingual or bilingual dictionary, this model adopts the Concepts expressed as WordNet synsets as the pivot to link languages in a very concise and effective way. The paper also addresses the most fundamental question in any lexicographer’s mind, viz., how to maintain lexical knowledge, especially in a multilingual setup, with the best possible levels of simplicity and economy? The case study of multiple Indian languages with special attention to three languages belonging to two different language groups (such as, Germanic and Indic) within the Indo-European family - English, Hindi and Marathi- throws lights on various linguistic challenges in the process of dictionary development.

The roadmap of the paper is as follows. Section 2 motivates the work. Section 3 is on related work. The proposed synset based model for multilingual dictionary is presented in section 4. Section 5 is on how to tackle the problem of correct lexical choice on the target language side in an actual MT situation through a novel idea of word alignment. Linguistic challenges are discussed in Section 6. Creation, storage and maintenance of the multilingual dictionary is an involved task, and the computational framework for the same is described in section 7. Section 8 concludes the paper.

2 Motivation

Our mission is to develop a single multilingual dictionary for all Indic languages plus English in an effective way, economizing on time and effort. We first discuss the disadvantages of language pair wise conventional dictionaries.

2.1 Disadvantages of Conventional Bilingual Dictionaries

In a typical bilingual dictionary, a word of \( L_1 \) is taken to be a lexical entry and for each of its senses the corresponding words in \( L_2 \) are given. It is possible that one sense of \( W_i \) in \( L_1 \) is exactly the same as one of the senses of \( W_j \) in \( L_1 \). This means that \( W_i \) and \( W_j \) are synonymous for a given sense. An example of this is dark and evil where one of the senses of dark and evil overlaps as for example in dark deeds and evil deeds. This phenomenon is abundant in any natural language. In a conventional dictionary, there is no mechanism to relate \( W_i \) with \( W_j \) in \( L_2 \), though they conceptually express the same meaning. In turn, the corresponding words for \( W_i \) and \( W_j \) in \( L_2 \) are no way related to each other though conceptually they are. That is a major drawback, because of which conventional pair wise dictionaries cannot be used effectively in natural language application, especially when multiple languages are involved.

The other disadvantage of the conventional dictionary is the duplication of manual labor. If an MT system is to be developed involving \( n \) languages, \( n(n-1)/2 \)}
language pair wise dictionaries have to be created. For instance, if we consider 6 languages, 30 bilingual dictionaries have to be constructed. Additionally will be required 15 perfect bilingual lexicographers- by no means an easy condition to meet. Finally, the effort of incorporating semantic features in $O(n^2)$ dictionaries is duplicated by $n/2$ lexicographers- a wastage of manual labor and time.

3 Related Work

Our model has been inspired by the need to efficiently and economically represent the lexical elements and their multilingual counterparts. The situation is analogous to EuroWordNet [1] and Balkanet [2] where synsets of multiple languages are linked among themselves and to the Princeton WordNet ([3], [4]) through Inter-lingual Indices (ILI). Our framework is similar, except for a crucial difference in the form of cross word linkages among synsets (explained in section 5). Another difference is that there are semantic and morpho-syntactic attributes attached to the concepts and their word constituents to facilitate MT. The Verbmobil project [5] for speech to speech multilingual MT had pair wise linked lexicons. To the best of our knowledge, no major machine translation nor CLIR project involving multiple large languages has ever used concept based dictionaries.

The framework has indeed been motivated by our creation of the Marathi WordNet [6] by transferring from the Hindi WordNet [7]. We noticed the ease of linking the concepts when two languages with close kinship were involved ([8], [9]).

4 Proposed Model: Concept-based Multilingual Dictionary

We propose a model for developing a single dictionary for n languages, in which there are linked concepts expressed as synsets and not as words. For each concept, semantic features- which are universal- are worked out only once. As for morph-syntactic features, their incorporation will demand much less effort, if languages are grouped according to their families; in other words we can take advantage of the fact that close kinship languages share morpho-syntactic properties. Table 1 illustrates the concept-based dictionary model considering three languages from two different families.
### Table 1. Proposed multilingual dictionary model

<table>
<thead>
<tr>
<th>Concepts</th>
<th>(L_1) (English)</th>
<th>(L_2) (Hindi)</th>
<th>(L_3) (Marathi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept ID: Concept description</td>
<td>((W_1, W_2, W_3, W_4))</td>
<td>((W_1, W_2, W_3, W_4, W_5, W_6, W_7, W_8))</td>
<td>((W_1, W_2, W_3, W_4, W_5, W_6, W_7, W_8, W_9, W_{10}))</td>
</tr>
<tr>
<td>02038: a typical star that is the source of light and heat for the planets in the solar system</td>
<td>(sun)</td>
<td>(सूर्य, भास्कर, हिदनकर, िदनेश, सिवता, िमिहर, अंशुमान, अंशुमाली)</td>
<td>(सूर्य, भास्कर, हिदनकर, िदनेश, िमिहर, रिव, िदनमणी)</td>
</tr>
<tr>
<td>04321: a youthful male person</td>
<td>(male_child, boy)</td>
<td>(लड़का, बालक, बच्चा, छोकड़ा, छोरा, छोकरा, लौंडा)</td>
<td>(मुलगा, भृद्रा, द्र. रेणे)</td>
</tr>
<tr>
<td>06234: a male human offspring</td>
<td>(son, boy)</td>
<td>(ूग, बेटा, लड़का, बाल, सूत, बच्चा, तनय, िचरंजीव, िचरंजी)</td>
<td>(मुलगा, िचरंजीव, तनय)</td>
</tr>
</tbody>
</table>

Given a row, the first column is the pivot for \(n\) number of languages describing a concept. Each concept is assigned a unique ID. The columns (2–4) show the appropriate words expressing the concepts in respective languages. To express the concept ‘04321: a youthful male person’, there are two lexical elements in English, which constitute a synset. There are seven words in Hindi which form the Hindi synset, and four words in Marathi which constitute the Marathi synset for the same concept, as illustrated in Table 1. The members of a particular synset are arranged in the order of their frequency of usage for the concept in question. The proposed model thus defines an \(M \times N\) matrix as the multilingual dictionary, where each row expresses a concept and each column is for a particular language.

#### 4.1 Advantages of the concept-based multilingual dictionary

(a) The first advantage of the proposed model is economy of labor and storage. Semantic features like \([\pm\text{Animate}, \pm\text{Human}, \pm\text{Masculine}, \text{etc.}]\), are assigned to a nominal concept and not to any individual lexical item of any language. Similarly, the semantic features, such as \([\text{+Stative} (\text{e.g., know}), \text{+Activity} (\text{e.g., stroll}), \text{+Accomplishment} (\text{e.g., say}), \text{+Semelfactive} (\text{e.g., knock}), \text{+Achievement} (\text{e.g., win})\] are assigned to a verbal concept. These semantic features are stored only once for each row and become applicable independent of any language. Consequently, lexical entries with highly enriched semantic features can be added to a dictionary for as many languages as required within a short span of time.

(b) The dictionary developed in this approach also serves all purposes that either a monolingual or bilingual dictionary serves. A monolingual or bilingual dictionary can
automatically be generated from this concept-based multilingual dictionary. The quality of such monolingual or bilingual dictionaries is better than that of any conventional bilingual dictionary in terms of lexical features.

(c) The model admits of the possibility of extracting a domain specific dictionary for all or any specific language pair. This is because the synsets or concepts pertaining to a domain can be selected from among the rows in the $M \times N$ concepts vs. languages matrix.

(d) The language group which lacks competence in the pivot language— which in our case is Hindi— can benefit from the already worked out languages. It may be the case that the lexicographers of language $L_6$ do not have enough competence in the pivot language $L_{pivot}$. They can look for a language $L_n$ which they are comfortable with and use $L_n$ as pivot to link $L_6$. This paves way for the seamless integration of a new language into the multilanguage dictionary.

5 Word-Alignment in the Proposed Model

In an actual MT situation, for every word or phrase in the source language a single word or phrase in the target language will have to be produced. The multilingual dictionary proposed by us links concepts which are sets of synonymous words. This is a major difference from the conventional bilingual dictionary in which a word ($SW_1$) in the source language is typically mapped to one or more words in the target language depending upon the number of senses $SW_1$ has. This implies that for each sense of $SW_1$, there is a single target language word $TW_1$. In our concept-based approach, even if we choose the right sense of a word in the source language ($SW_1$), there is still the hurdle of choosing the appropriate target language word. This lexical choice is a function of complex parameters like situational aptness and native speaker acceptability. For example, the concept of ‘the state of having no doubt of something’ is expressed through the Hindi synset having six members (िनश्शंक, अनाशंिकत, आशंकाहीन, बेखटक, बेिफ़Đ संशयहीन) and through the Marathi synset having four members (िनःशंक, िनधार्स्त, िनःिाɍत, शंकारिहत). However, the third member in the Hindi synset आशंकाहीन is appropriately mapped to the fourth member in the Marathi synset शंकारिहत. Though the mapping of the third member in the Hindi synset (i.e., आशंकाहीन) with the first member of the Marathi synset (i.e., निःशंक) expresses the same meaning, this substitution sounds quite unnatural to the native speakers.

We tackle the problem of correct lexical choice on the target language side by proposing a novel approach of word-alignment across the synsets of languages. Word-alignment refers to the mapping of each member of a synset with the most appropriate member of the synset of another language. For instance, when the word लड़का ‘boy’ in Hindi in the sense of ‘a young male person’ Needs to be lexically transferred to Marathi, there are four choices available in the synset, as illustrated in Figure 1.
Considering Hindi as the pivot, we propose that each of the four words in the Marathi synset be linked to the appropriate Hindi word in the direction Marathi→Hindi and each of the two words in English synset has to be linked with the appropriate Hindi word in the direction English→Hindi. As a result, the first and the third member of the Marathi synset (i.e., मुलगा and रेगा) are mapped to two different Hindi words (i.e., मुलगा→लड़का, रेगा→बच्चा). The second and the fourth member in Marathi synset are linked to one word (i.e., रेगा→छोकरा and रेगे→छोकरा) in the Hindi synset. Three words in Hindi synset (i.e., HW4, HW5, HW7) are left without being linked, as shown in Figure 1. In a situation, when a Marathi word is aligned with a single Hindi word (e.g., मुलगा→लड़का) for a particular concept in the direction Marathi→Hindi, from our past experience we assume that the lexical transfer in the reverse direction (Hindi→Marathi) also holds good, yielding लड़का→मुलगा.

Following this strategy of alignment of synset members of Marathi (or any other language) with the synset members of the pivot (i.e., Hindi in the present scenario), we are having four types of situation to perform a lexical transfer from any language to any language:

- **Situation (1) One-to-One**
- **Situation (2) Many-to-One**
- **Situation (3) One-to-Many**
- **Situation (4) No link**

In situation (1), the source word is found to be linked to a single target word, via a synset member of the pivot if it is neither the source nor the target for any lexical transfer. For instance, the Marathi word मुलगा can be transferred to the Hindi target word लड़का, and the Marathi word मुलगा can be transferred to the English target word 'boy' via the pivot Hindi word लड़का. In situation (1), virtually there is no problem in performing the lexical transfer maintaining the best naturalness to the target language.
speakers. In situation (2), two words from the source language synset are linked to a single word in target language, e.g., छोकरा → छोकरा and छोकरा → छोकरा. Hence, there is no issue involved in lexical transfer maintaining the naturalness. The situation (3) arises when the pivot is taken as the source language in any practical application, e.g., Hindi → Marathi. The lexical transfer involves a puzzle with respect to the naturalness of the target word. Since the members of a synset are ordered according to their frequency of usage for a concept, we are inclined to choose the first member of the target synset as the best in this situation. For instance, the source Hindi word छोकरा ‘boy’ has two choices in the target Marathi synset, i.e., ओरगा and ओरगे, as shown in figure 1. Since ओरगा appears prior to ओरगे, we choose ओरगा for lexical transfer. In situation (4), where no link is available between the source word and the target word, we choose the first member of the target synset for lexical transfer. If we need to transfer the Marathi word ओर to English, there is no consecutive link available, since it stops at बच्छा/HW3 in the pivot (cf. figure 1). However, we choose the first member of the English synset, i.e., boy for Marathi ओर, which is quite appropriate and widely acceptable. Similarly, if the English word boy happens to be the source in the sense of ‘a youthful male person’, the first member of the Marathi synset (i.e., मुलगा) is chosen as the target for lexical transfer, even if its link stops at बालक/HW2 in the pivot (cf. figure 1). In section 8, we present a user-friendly tool to align the members of the synsets across languages with respect to a particular concept. We also present a lexical transfer engine to make the aligned data usable in any system.

6 Linguistics Challenges Involved

In the process of synset based multilingual dictionary development, we face a number of challenges to deal with linguistic and cultural diversity. In this section, we present a few cases that we experienced while dealing with three languages, i.e., English, Hindi and Marathi.

(a) A concept may be expressed using different syntactic category in different languages. For example, the nominal concept कलौंजी ‘stuffed vegetable’ in Hindi is expressed through an adjectival concept भरली ‘stuffed’ in the expression भरलेली भाजी ‘stuffed vegetable’ in Marathi.

(b) It is often the case that a concept is expressed through a synthetic expression in one language, but through a single word expression in the other language. For example, the concept ‘reduce to bankruptcy’ is expressed through a single word in English but through a synthetic expression in Hindi and Marathi, as illustrated in Table 2.
Considering Hindi as the pivot in the process of dictionary development in our approach, one has to deal with two kinds of situation: (i) synthetic expression in the pivot to a single word expression in the other language, (ii) single word expression in the pivot to a synthetic expression in the other language. In situation (ii), the question arises with respect to its morpho-syntactic category in the dictionary. Because, the synthetic element is often constituted of different syntactic categories, as shown in Table 2. In such a situation, we consider the grammatical function of the synthetic element and assign the category accordingly. For example, the Marathi expression विवाहसमयी वराकडन वधुला हादिले जाणारे दागणे ’at-the-time-of-wedding–bridegroom–bride– given–ornament’ refers to an adjectival function at the grammatical level, hence its syntactic category is assigned as ‘adjective’.

(c) When a word expressing the meaning specific to a particular language and culture has to be mapped to another language in the dictionary, we find two ways to express the concept in another language: (i) using a synthetic expression, (ii) using transliteration, if the synthetic expression is larger. For example, the culture specific concept of ‘ornaments and other gifts given to the bride by the bridegroom on the day of wedding’ is lexicalized in Hindi yielding चढ़ावा, but a Marathi speaker has to use a larger synthetic expression विवाहसमयी वराकडन वधुला हादिले जाणारे दागणे to express the same concept. The Hindi word सेहरा ’garland’ is a culture specific word which has no lexical equivalent in Marathi. Even using a large synthetic expression does not express the borrowed concept naturally. In such a situation, we transliterate the culture specific word into Marathi.

It is also the case that a concept is culture specific to a language other than the pivot. For example, the Marathi culture specific concept, e.g., माहेरवाशीण ’a woman who has come to stay at her parents’ place after her marriage’, is not expected to be available in the pivot language dictionary in the initial phase. Therefore, such culture specific concepts are added to the Marathi dictionary in a monolingual manner without being mapped to the pivot language. But those are marked for review using the dictionary development tool. At a later phase, each language specific cultural concepts can be taken, and systematically added to the pivot.

**Table 2. Illustration of single word vs. synthetic expressions**

<table>
<thead>
<tr>
<th>Concept</th>
<th>English</th>
<th>Hindi</th>
<th>Marathi</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘reduce to bankruptcy’</td>
<td>bankrupt</td>
<td>दिवाला निकालना</td>
<td>दिवाले काढणे</td>
</tr>
<tr>
<td></td>
<td>(V)</td>
<td>(N+V)</td>
<td>(N+V)</td>
</tr>
<tr>
<td>‘resulting from careful thought’</td>
<td>considered</td>
<td>विचार देव किया हुआ</td>
<td>विचार देव केलेला</td>
</tr>
<tr>
<td></td>
<td>(ADJ)</td>
<td>(ADV+VERB)</td>
<td>(ADV+VERB)</td>
</tr>
<tr>
<td></td>
<td>‘thoughtfully done’</td>
<td>‘thoughtfully done’</td>
<td></td>
</tr>
<tr>
<td>‘least in age than the other person’</td>
<td>youngest</td>
<td>कलिश</td>
<td>सयोलत लहान</td>
</tr>
<tr>
<td></td>
<td>(ADJ)</td>
<td>(ADJ)</td>
<td>(N+ADJ)</td>
</tr>
<tr>
<td></td>
<td>‘among-all less-in-age’</td>
<td>‘among-all less-in-age’</td>
<td></td>
</tr>
</tbody>
</table>
language to enrich the pivot, and in turn, the whole multilingual dictionary with multicultural concepts.

(d) Given Hindi as the pivot language, when we develop and link the Marathi dictionary, we come across a strange situation. A concept initially recorded in Hindi dictionary, having a singleton member in the pivot synset, can be expressed through more than one finer concept in Marathi. The Hindi word फीका means "the food prepared with less sugar, salt or spice", the equivalent of which is expressed in Marathi through three distinct words expressing three distinct finer concepts, i.e., अगोड 'less sweet', अळणी 'less salty', and िमळिमळत 'less spicy'. These three words cannot be taken as the members of a single synset in Marathi for the concept "the food prepared with less sugar, salt or spice", since the three-way finer meaning distinction is very natural to Marathi speakers. Had it been the case that Marathi were the pivot, we could have been tempted to add three different concepts into Marathi dictionary, and in turn, Hindi dictionary could have included फीका against three concepts implying that फीका has three senses. As long as Hindi is the pivot, the finer concepts found in Marathi (e.g., अगोड 'less sweet', अळणी 'less salty', and िमळिमळत 'less spicy') cannot be mapped to the coarse concept found in Hindi (e.g., फीका ‘the food prepared with less sugar, salt or spice’). However, at a later phase of the dictionary development process, the finer concepts of Marathi (or any other languages) can be identified, and added to the pivot language, i.e., Hindi, after which the other languages can borrow the concepts from the same pivot to enrich their dictionary in the multilingual setting. The computational tool (cf. section 8) provides a support to mark such cases for review, and to retrieve all those when one decides to add to the pivot language synsets.

8 Computational Framework for the Multilingual Dictionary

For effective implementation of our idea of synset based multilingual dictionary, we carefully designed the dictionary development process, which is, in fact, expected to involve a number of human lexicographers. Figure 2 shows the complete semi-automatic data flow in the dictionary development process.
The pivot synsets are extracted from the existing Hindi WordNet along with the concept descriptions, syntactic category and examples. For convenience, an appropriate template is used for multilingual dictionary development, as illustrated in Table 3.

| ID   | :: 02691516 |
| CAT  | :: verb     |
| CONCEPT | :: be in a state of movement or action |
| EXAMPLE | :: “The room abounded with screaming children” |
| SYNSET-ENGLISH | :: (abound, burst, bristle) |

The whole process, shown in figure 2, is implemented using a centralized MYSQL database and a JAVA GUI. The screenshots of the GUI windows are shown in figure 3 and 4. Language and task configuration window is shown Figure 3, and the synset entry interface is shown in figure 4. The tool accepts the data in UNICODE only.
Fig. 3. Language and Task Configuration Window

Fig. 4. Synset entry and word-alignment interface
Once the dictionary is built out of the multilingual data as shown in figure 4, a lexical transfer engine provides the following for various usages:

(i) Given a word in any language, get all the records in the specified template in the same language or in any other language. *(useful for a WSD system)*

(ii) Given a word in any language and its part-of-speech, get all the records in specified template in the same language or in any other language. *(useful for a WSD system)*

(iii) Given a word in any language with respect to a particular concept, get the most appropriate translation of that word in any other language. *(useful for lexical transfer in an MT system, if a WSD system is embedded in the MT system)*

(iv) Given a word in any language, get the most probable translation of that word in any other language. *(useful for lexical transfer in an MT system having no WSD system embedded and in a cross-lingual information retrieval system)*

Using this lexical transfer engine, the multilingual dictionary is accessible online through a user-friendly website having a facility for obtaining feedback from online dictionary users. The feedback obtained from online users is expected to be useful for further development of this invaluable lexical resource.

9 Conclusion and Future Directions

We have reported here our experiences in the construction of a multilingual dictionary framework that is being used across language groups to create large scale MT and CLIR systems. Many challenges are faced on the way, chief amongst them being the one-on-one production of a target language lexeme corresponding to a source language lexeme. On the computational front there are challenges to be tackled for the maintenance of multilingual data, their insertion, deletion and updation in a spatially and temporally distributed situation. Of the many advantages of the framework are:

(i) a linguistically sound basis of the dictionary framework, (ii) economy of representation and (iii) avoidance of duplication of effort. Our future work consists in incorporating domain sensitivity to the framework and also in solving the challenges of the distributed access and storage.

References

Estonian WordNet: Nowadays

Heili Orav, Kadri Vider, Neeme Kahusk, and Sirli Parm

University of Tartu, Institute of Estonian and General Linguistics

Abstract. The Estonian WordNet has been built since 1998. After fin-
ishing EuroWordNet-2, the Estonian team continued with word sense
disambiguation, using Estonian WN as lexicon. Many synsets are im-
proved since then. Nowadays main attention is payed to specific domains
and completing with glosses and examples. Adverbs constitute a totally
new part in EstWN.

1 Introduction

The Estonian team joined the WordNet community (EuroWordNet-2) from the
beginning of January 1998. In the framework of the project of Estonian language
technology the Estonian WordNet has been created during the years 1997–2000.
After some discontinuation this project is awaken again. This year started project
for increasing Estonian WordNet (EstWN) and is supported by Estonian Na-
tional Programme on Human Language Technology.

In our paper we aim to give an overview about development of EstWN and
problems with which we face in every-day work. The Estonian WordNet at the
present stage includes 10372 noun synsets, 1580 adjective synsets and 3252 verb
synsets. Parallel works with thesaurus are increasing in size, adding new semantic
relations and specification of specific domains.

Fig. 1. Current state of Estonian WordNet (September 2007)
2 The Dynamics of Progress

We can consider the EuroWordNet-2 project as first stage in our WordNet building. Then we ended up with ca 9500 synsets. Every synset had to have at least one Language-Internal relation and one InterLingual Index (ILI) relation. Hyperonymy/hyponymy links had first priority, but several other semantic relations were added during more intense insights into specific topics (see Tab. 1).

Table 1. Distribution of language internal relations by part of speech

<table>
<thead>
<tr>
<th>Language-internal relations</th>
<th>Nouns</th>
<th>Verbs</th>
<th>Adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperonymy/hyponym</td>
<td>14752</td>
<td>6444</td>
<td>0</td>
</tr>
<tr>
<td>Near-synonymy</td>
<td>157</td>
<td>145</td>
<td>322</td>
</tr>
<tr>
<td>antonymy</td>
<td>198</td>
<td>122</td>
<td>126</td>
</tr>
<tr>
<td>causation</td>
<td>118</td>
<td>188</td>
<td>2</td>
</tr>
<tr>
<td>Involvement and roles</td>
<td>262</td>
<td>216</td>
<td>1</td>
</tr>
<tr>
<td>subevents</td>
<td>7</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>Holonymy/meronym</td>
<td>294</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In ILI-relations we can see fairly great number of non-equal synonyms. This can have two main reasons: first, near-synonymy relations are result of differences in word-sense distribution. The members of Estonian synsets sometimes do not map precisely into English ones. Second, there are plenty of language-specific concepts that do not have equal match in English – they have equal-hyperonymy/hyponymy relations to describe their exact meaning via ILI.

Table 2. Distribution of interlingual relations by part of speech

<table>
<thead>
<tr>
<th>Interlingual index relations (ILI)</th>
<th>Nouns</th>
<th>Verbs</th>
<th>Adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>eq_synonym</td>
<td>5980</td>
<td>2153</td>
<td>291</td>
</tr>
<tr>
<td>eq_near_synonym</td>
<td>826</td>
<td>1308</td>
<td>25</td>
</tr>
<tr>
<td>eq_has_hyperonym</td>
<td>708</td>
<td>181</td>
<td>0</td>
</tr>
<tr>
<td>eq_has_hyponym</td>
<td>279</td>
<td>136</td>
<td>0</td>
</tr>
<tr>
<td>eq_causes</td>
<td>2</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>eq_is_caused_by</td>
<td>39</td>
<td>83</td>
<td>0</td>
</tr>
<tr>
<td>eq_be_in_state</td>
<td>91</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>eq_involved</td>
<td>144</td>
<td>95</td>
<td>0</td>
</tr>
<tr>
<td>eq_has_holonym</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>eq_has_meronym</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The second stage in Estonian WN development started after end of the EuroWordNet project. Our main focus was concentrated on EstWN applications. The WSD task on SENSEVAL-2 showed that several word senses were missing (Kahusk and Vider 2002). Problems at manual disambiguation revealed the need for more precise sense borders in EstWN, so we added glosses and examples to many synsets. The glosses come mainly from the Explanatory Dictionary of Estonian (EKSS) and examples come from our new WSD corpus.

The third stage in Estonian WN development started on the year 2000. In contrast to the first stage, that involved much manual work, the present day EstWN contains about 4500 noun, verb and adjective synsets that are added from the Estonian dictionary of synonyms (Öim [5]) by automatic extraction. Still, these are only lexical synonym entries without any glosses, examples and semantic relations. We have imported some glosses and examples from EKSS, but language-internal semantic relations and ILI links are provided by lexicographers.

3 Adverbs

Since September 2007 we started to add adverbs into Estonian WordNet. Adverbs clarify or modify the meaning of a sentence. So, they have an important role in sentence meaning. The most common adverbs in Estonian are these of quantity and time, some of them have multiple senses. For example, adverb ‘veel’ (more; still/yet), which means in one context: a greater number or quantity, and another context: the time mentioned previously, however the adverb ‘veel’ is sometimes used in the both sense (as the quantifier and as the time particle).

We are started with adverbs of time, such as ‘täna’ (today), ‘homme’ (tomorrow) etc and polysemous adverbs, such as ‘jälle’ (again), ‘juba’ (already) etc. There are some problems with semantic relations, which we have needed to solve. Estonian adverbs typically express some relation of space, time, manner, degree, cause, inference, condition, exception, purpose, or means. How specifically we need to mark semantic relation between adverbs meanings, such as express some relation of time. An example, in classical semantic analysis the time adverb ‘veel’ (still/yet) is antonym for the time adverb ‘juba’ (already), or at least the near antonym.

In further works, we are continuing to add manner adverbs, which meanings are mostly linked to adjective senses. Estonian manner adverbs are often formed by adding suffix -sti, or -lt, as ‘kiire+sti=kiiresti’, or ‘kiire+lt=kiirelt’ (quickly/rapidly), so these derived adverbs usually inherits the sense of the base adjective and also their semantic relations, in this case, an example ‘kiirelt’ (rapidly) is a antonym for ‘aeglaselt’ (slowly).
4 Adjectives

The most thoroughly examined domain in EstWN are the adjectives of personality traits. This specific domain include in Estonian around 1200 words or expressions, which accordingly form around 400 synsets into Estonian WordNet.

Semantically, words and expressions of character traits converge into certain concept groups. The composition of character vocabulary showed the definitions of intrapersonal or interpersonal qualities to be for the most part broader and more general, it is into these two vast categories the vocabulary is divided into. Based on the material, 55 concept groups of personality traits have been defined, some with subsequent subgroups, and formed mainly on the basis of synonymy/antonymy relationships (Orav 2006). In future we plan to examine more domains which are represented mostly by adjectives, for example colours, weather etc.

5 Specific domains

There are some students who have studied specific domains in language (eg. transportation and motion, see Fig. 2), and increased the number of synsets up to 500 per domain.

Semantic fields which are covered in details at this stage of work are as on Fig. 2.

![Fig. 2. Specific domains in Estonian WordNet.](image-url)
6 Availability

There is an application of EstWN is called TEKsaurus. This is an online service that is based on Estonian WordNet. TEKsaurus is browseable in Internet\(^1\). The engine behind TEKsaurus is a Python script running on server. In first stage, the EstWN export file is used to generate index file of literals. The server-side engine uses the same export file to find offsets, where synset data is found and presented to browser.

For more specific description, see Kahusk and Vider (2005).

The Estonian WordNet source file is available at ELRA.

7 Acknowledgements

The Estonian WordNet is supported by National Programme “Language Technology Support of Estonian Language” projects No EKKTT04-5, EKKTT06-11 and EKKTT07-21 and Government Target Financing project SF0182541s03 “Computational and language resources for Estonian: theoretical and application aspects.”

References


\(^1\) http://www.cl.ut.ee/ressursid/teksaurus
Event Hierarchies in DanNet

Bolette Sandford Pedersen¹ and Sanni Nimb²

¹University of Copenhagen, Njalsgade 80, 2300 S, Denmark,
²Det Danske Sprog- og Litteraturselskab, Christians Brygge 1, 1219 K, Denmark
bolette@cst.dk, sn@dsl.dk

Abstract. The paper discusses problems related to the building of event hierarchies on the basis of an existing lexical resource of Danish, Den Danske Ordbog (DDO). Firstly, we account for the reuse principles adopted in the project where some of the senses given in DDO are either collapsed or readjusted. Secondly, we discuss the semantic principles for building the DanNet event hierarchy. Following the line of Fellbaum, we acknowledge that the manner relation (troponymy) must be defined as the main taxonomical principle for describing verbs, but we observe some complications with this organizing method since many subordinate verbs tend to specify other meaning dimensions than manner. We suggest to encode verbs that do not follow the strict manner pattern as ‘orthogonal’ to the basic hierarchy, a strategy which opens for the compatibility between taxonomical and non-taxonomical sisters of synsets.

Keywords: WordNets, events, event hierarchies, verbs, troponymy.

1 Introduction

Building meaningful event hierarchies proves to be a challenging task, in many respects much harder than building taxonomies over 1st order entities. Firstly, event hierarchies are not quite as intuitive as hierarchies of 1st order entities, and secondly, there seems to be an extra measure of indeterminacy in the meaning of a verb which complicates the issue at several levels. The aim of this paper is to present and discuss some of the principles that we have applied in order to ease the construction of consistent event hierarchies in the DanNet WordNet, basing the encodings partly on a big traditional dictionary of Danish, Den Danske Ordbog (DDO) [1], and partly on encodings from an EU project on semantic computational lexica, SIMPLE (Semantic Information for Multifunctional, Plurilingual Lexica) [2].

It is generally accepted that current WordNets are built on rather heterogeneous subsumption relations, a fact which has been discussed and questioned in literature by both formal ontologists [3] and WordNet builders [4], [5]. The apparently rather messy taxonomical structure of many WordNets should however not be judged as inconsistent or incompetent work, but rather as a result of the fact that they are built on the basis of corpus derived lexical data, and thereby they actually represent the variety and complexity of lexical items with its characteristic heterogeneous mixture.
of types and roles. This can be seen as a contrast to formal ontologies where main attention is paid to types in the ontology skeleton. In DanNet, we distinguish between types and roles for 1st order entities in the sense that we propose to apply Cruses distinctions [6] on nouns between Natural kinds, Functional kinds, and Nominal kinds (see [5]). This distinction is carried out in order to determine when a synset should be categorised as “orthogonal” to the hierarchy (i.e. as a role), as in cases of nominal kinds like for instance climbing trees, and when it is actually a type in the main taxonomy, as is the case for natural kinds like for instance oaks.

In this paper we wish to examine whether similar distinctions can help clear up 2nd order entities in terms of event hierarchies. Fellbaum [4] discusses semantically heterogeneous manner-relations (defined as troponymy) and argues that similar cases do hold between verbs. She gives the examples move and exercise and proposes that parallel hierarchies should be established, allowing verbs like run and jog to act as subordinates of both move and exercise. From a practical viewpoint, we claim that such parallel hierarchies are, however, extremely complicated to build and maintain in a consistent way over a large scale, and we therefore adopt the solution as mentioned above of marking non-taxonomical synsets as “orthogonal”. Such a marking indicates among other characteristics that there is no incompatibility between an orthogonal sister and a taxonomical sister; an oak may be a climbing tree at the same time, as well as practically any moving event could also be seen as an exercising event in a specific context.

Two classical lexical phenomena tend to complicate the establishment of event hierarchies even further, namely those of polysemy and synonymy. In DanNet we take the sense distinctions given by DDO (which are corpus-based) as our starting point, but in the case of the verbs, we have realised that some reconsideration of polysemy and synonymy is necessary when building a WordNet.

Regarding polysemy, some principles for when to split senses and more importantly when to merge polysemous senses are therefore developed. Also the establishment of synonymy calls for further clarification, since it is not always obvious when two verbs denote one or two events. Finally, we observe that within some domains, the manner relation is not the main organizing principle whatsoever. In these cases, we propose an under-specification of the taxonomical description, but suggest to specify the particular meaning dimension via other specific relations, if possible.

The paper is composed as follows: we start in Section 2 with a brief introduction to the DanNet project as a whole, a WordNet project which relies heavily on reuse of existing lexical data. Then we present some problems especially connected to the reuse of verb entries from DDO (Section 3). In Section 4 we discuss the building of DanNet verb hierarchies and introduce the orthogonal hyponym as a way of dealing with a series of non-typical cases of troponymy. Finally in Section 5 we conclude and turn to future work, where we plan to combine DanNet with a deeper FrameNet-like description of verbs.
2 DanNet: Background

DanNet [7] is a collaborative project between a research institution, Center for Sprogteknologi, University of Copenhagen, and a literary and linguistic society, Det Danske Sprog- og Litteraturselskab under The Danish Ministry of Culture. In the project we exploit the large dictionary, DDO (approx. 100,000 senses) and the ontological resource SIMPLE-DK (11,000 senses), i.e. the Danish part of the EU-project Semantic Information for Multifunctional, Plurilingual Lexica [2].

The first phase of the project is coming to an end, and the DanNet database has now reached a size of 40,000 synsets, of which 6,000 are verb senses. In the second phase of the project, the goal is to achieve the complete coverage of DDO, namely approx. 65,000 senses disregarding in this context most multiword expressions.

3 Reuse Perspectives

DDO contains 6600 Danish verb lemmas amounting to 19,000 senses in all. For verbs, as for all other word classes, genus proximum information is assigned in a specific field in DDO, coinciding with a superordinate verb which is already a part of the word definition. At a first glance, it therefore seems straightforward to reuse this information on genus proximum when building the event taxonomy in DanNet, as well as it is done in the case of nouns. When we look closer into the data, we see that many verb senses share the same genus proximum and that the verbs which are most frequently used as genus proximum often have very vague meanings. As an example, 4755 verbs senses (25 % of the total number of senses) share the same 15 verbs as genus proximum. These 15 verbs are all extremely polysemous lemmas, in average they have in fact 22 main senses and sub-senses each.

Given the fact that the genus proximum in DDO is not marked with respect to sense distinctions in the many cases of polysemy, it becomes clear that the genus proximum information given for verbs in DDO does not automatically indicate a reusable structure of a general hierarchy, especially not at the top level of the network. We have therefore also taken in information from the network of SIMPLE-DK and have been forced to manually adjust a large set of the hyponymy relations given in DDO (see Section 4 for further details).

These adjustments regarding the event taxonomy are furthermore challenged by the classical dilemma of when to merge and when to split senses, see e.g. [8], [9] and [10]. Frequent verbs are described in DDO at a very detailed level with many senses and sub-senses (see Table 1). The question is whether we necessarily want to maintain the fine-grainedness of DDO. Is it at all manageable in a semantic net meant for computer systems? And if not, how do we ensure a systematic reduction of senses? And vice versa: are there cases where we need to split DDO senses in order to capture important ontological differences?
Table 1. The distribution and the polysemy of verb genus proxima in DDO

<table>
<thead>
<tr>
<th>Genus proximum</th>
<th>number of main senses and sub-senses of this genus proximum lemma in DDO (without phrasal verb senses and idiom senses)</th>
<th>number of verb senses described by this genus proximum in DDO (total number of verb senses in DDO: 19,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gøre (to do)</td>
<td>25</td>
<td>743</td>
</tr>
<tr>
<td>være (to be)</td>
<td>20</td>
<td>580</td>
</tr>
<tr>
<td>give (to give)</td>
<td>17</td>
<td>506</td>
</tr>
<tr>
<td>få (to get/have)</td>
<td>26</td>
<td>413</td>
</tr>
<tr>
<td>bevæge (to move)</td>
<td>4</td>
<td>391</td>
</tr>
<tr>
<td>have (to have)</td>
<td>23</td>
<td>376</td>
</tr>
<tr>
<td>blive (to become)</td>
<td>11</td>
<td>329</td>
</tr>
<tr>
<td>fjerne (to remove)</td>
<td>5</td>
<td>229</td>
</tr>
<tr>
<td>lade (to let)</td>
<td>11</td>
<td>195</td>
</tr>
<tr>
<td>tage (to take)</td>
<td>71</td>
<td>187</td>
</tr>
<tr>
<td>komme (to come)</td>
<td>23</td>
<td>187</td>
</tr>
<tr>
<td>bringe (to bring)</td>
<td>8</td>
<td>182</td>
</tr>
<tr>
<td>gå (to go)</td>
<td>35</td>
<td>171</td>
</tr>
<tr>
<td>sætte (to put)</td>
<td>30</td>
<td>161</td>
</tr>
<tr>
<td>holde (to hold)</td>
<td>24</td>
<td>105</td>
</tr>
<tr>
<td><strong>total 15 verbs</strong></td>
<td><strong>total 333 senses</strong></td>
<td><strong>4755 = 25 % of all verb senses</strong></td>
</tr>
</tbody>
</table>

Starting with the latter case, we sometimes find verb definitions in DDO covering what we would in DanNet consider as two different senses with different hyperonyms. One example is: *krumme* (to bend) which is defined as: ‘to be or to become curved’. The definition covers two types of telicity, and therefore represents two different ontological types in DanNet, resulting therefore in a splitting into two synsets. Another example is *afbilde* (to depict), where one definition in DDO in fact covers two DanNet senses: 1) somebody illustrates something by producing a mathematic figure; 2) a mathematic figure shows something. The first part of the definition describes an act by a person, whereas the second rather denotes a state. Summing up, the following procedure is followed:

- **split senses** when a sense in DDO covers both an activity and a state.

If we now move to the verbs that are described as polysemous in DDO, it turns out that 1500 of the 6600 verbs in DDO have more than one sense, meaning that approx 14,000 senses come from polysemous verbs. In other words, each polysemous verb has an average of almost 10 senses. The general assumption in DanNet is that we maintain the main sense divisions given in DDO since we rely on them as being actually corpus-based and therefore relevant for our purpose. For instance, we maintain the four sense distinctions given for *lukke* (to close): 1) to close a window, the eyes, a door 2) to close a bag 3) to close a road, a passage and 4) to stop a function (e.g. the television). Nevertheless, for some of the more systematic main sense cases, we have adopted two merging strategies:
- *merge senses* describing a certain (physical) act being performed by either a human being or another living entity. Example: *æde* (to eat) which has two main senses in DDO, one for animals and one for humans.

- *merge senses* describing different valency patterns, but with the same meaning, such as ergative verbs like *geare ned* (gear down (fig.)) which can either be intransitive or transitive.

When it comes to the many subsenses of verbs in DDO, we generally find far more cases of potential merging. A main principle is to merge a sub-sense with its main sense when the sub-sense represents (i) a more restricted sense, or (ii) an extended sense. In contrast, figurative senses, these are generally maintained, belonging often to different ontological type. See Fig. 1.

![Fig. 1. Merging sub-senses from DDO with the main sense](image)

### 4 Verb Descriptions in DanNet

A preliminary encoding of the first 6,000 verb synsets has recently been completed in the project. Although highly inspired by the SIMPLE-DK descriptions of events (built partly on Levin classes, cf. [11] and [12]), we have chosen to apply the EWN Top Ontology of 2nd Order Entities (see Figure 2) in order to be compatible with other WordNets developed within this framework. In order to guide the encoding work, approx. 60 event templates have been established, combining situation types and situation components in different sets. The main dividing principle is that of telicity, as reflected in the situation types BoundedEvent and UnboundedEvent. However, in Danish, telicity is in most cases specified by means of verb particles - and not as in Romance languages - given in the verbal root. This can be seen for instance for the verb *spise* (eat) which seen in isolation denotes an atelic, unbounded event as opposed
to the phrasal verb *spise op* (finish ones food) which denotes a telic, bounded event. Phrasal verbs in general constitute a large part of the encoded senses in DanNet, and many verbs have parallel encodings as bounded and unbounded events depending on the presence or absence of a phrasal particle.

Fig. 2: The EWN Top Ontology, cf. [13]

The building process was initiated by working with the genus proximums in DDO that denote physical events, and where the encoded hyperonym proved to be more or less reliable, such as *bevæge sig* (move), *fjerne* (remove), *stille* (place), *ytre* (evince, express) etc. A feature in the DanNet tool enables us to identify such groups directly from DDO, analyzing thereby where to find larger groups of more or less homogeneous verbs.

The encodings of physical as well as communicative verbs served as the first building blocks for the event hierarchy. When about half of the verb vocabulary had been provisionally encoded, the need emerged to work top-down in order link the
verb groups together in a joint network. For this purpose 24 Danish top-ontological verbs were identified forming thereby a language-specific parallel to the EWN Top Ontology onto which all other verb senses are subsequently linked.

### 4.1 Determining the Taxonomical Structure of Events

Where the main organizing mechanism behind 1st and 3rd Order Entities is constituted by the hyponymy relation, events seem to be better organized along the dimensions of the manner relation – or troponymy relation – as proposed by Fellbaum. To give an example from DanNet, *guffe* (scoff) is a (quick and rough) way of *spise* (eat) which again is a way of *indtage* (consume), which again is a way of *handle* (act) etc as depicted in Figure 3.

![Diagram](image)

**Fig. 3.** *Indtage* (consume) with some of its hyponyms

Some verbs in the domain, however, tend to fall out of the pattern of troponymy, or at least they denote another dimension of the manner relation. The verb *trøstespise* (eat for comfort, i.e. be a compulsive eater), which is composed by the two verbs *trøste* (comfort) and *spise* (eat), respectively, is an example of a verb which does not relate to the physical manner of eating (fast, slow, nice, ugly, large amounts, small amounts), but rather to a psychological dimension of eating. As seen in Figure 4, we thus assign the hyponym as orthogonal to *spise*, depicted in the figure by a rhomb. Another solution would have been to follow Fellbaum, and establish parallel hierarchies by encoding *trøstespise* both as an eating event and as a comforting event, i.e. as troponym to *trøste_sig*. Such multiple inheritance is possible in the DanNet framework, but for pragmatic reasons we have decided not to establish such parallel hierarchies if they can be avoided, since they prove hard to encode in a consistent way.
A similar situation arises when encoding near synonyms. Synonymy further complicates the establishment of event hierarchies since, compared to 1st order entities, it is often much more unclear when two verbs actually refer to the same event. Figure 5 is a screen shot from the DanNet encoding tool which illustrates the problem with the synset *tilberede* (prepare (food)) together with its co-synonyms *tillave*, *preparere* and *lave*. *Kokkerere* (to perform finer cooking) in Danish is another word for preparing food, but it gives a specific association to finer cooking. Therefore it has been placed as a subordinate to {*tilberede, tillave, lave, preparere*}. This is at first glance unproblematic, only it specifies another semantic dimension than *bage* (bake), *pochere* (poach), *spejle* (fry (an egg)), *koge* (boil) etc. where the manner component is clearly in focus describing exactly what kind of heating process the food undergoes. Therefore, we encode *kokkere* as orthogonal to the rest of the hyponyms to cook, again visualized in the figure by a rhomb. Note that the orthogonal synset is characterized by being compatible with its sisters (unlike taxonyms); while performing finer cooking the ingredients may actually both undergo frying, baking and boiling.

4.2 Other Organizing Principles than Manner?

During our work, we have observed that within some physical domains, the main organizing principle cannot actually be characterized as a manner relation. Under the verb *fjerne* (remove), we find a series of verbs such as *afkalke* (decalcify, descale), *afluse* (delouse), *affugte* (dehydrate) and *affarve* (discolour, bleach) which specify *what* is removed from the object and not *how* it is removed. Likewise, in the domains
of most mental verbs, it proves to be the case that more subtle meaning dimensions are specified in the different hyponyms. Under the verb tenke (think) we find verbs like dagdrømme (day dream), bekymre sig (worry), forske (investigate) and mindes (recall), a very heterogeneous group of verbs organized along different dimensions of meaning that are not satisfactorily labeled as manner-relations.

5 Conclusions and Future Work

In this paper we have discussed some problems related to the building of event hierarchies on the basis of existing lexical resources of Danish. Following the line of Fellbaum, we acknowledge that the manner relation (troponymy) must be defined as the main taxonomical principle for describing verbs, but we also observe from the practical encoding that there are several complications with this organizing method since many subordinate verbs tend to specify other meaning dimensions than manner. If we look at verbs denoting physical events, which are actually the less complicated to work with, the manner relation is by far the most frequent relation, but there are many exceptions where a verb denotes a slightly different semantic dimension. In several of these cases, the verb could also be organized under another hyperonym building thereby parallel hierarchies, a strategy, however, that we have abandoned for maintenance reasons. Instead we suggest to mark verbs that do not follow the strict manner pattern with a feature stating that it denotes an orthogonal dimension of meaning to the basic hierarchy. This can be seen as parallel to the way that we encode 1st Order Entities, where we distinguish taxonomical and non-taxonomical hyponymy relations. We believe that by introducing this division, we obtain cleaner event hierarchies, and we allow for the compatibility between taxonomical and non-taxonomical synsets (i.e. for kikkere (to perform finer cooking) to take place by means of frying and baking etc).

Future plans regarding semantic verb descriptions in DanNet include to combine the resource with the already existing syntactic lexical database STO [14] in order to relate each verb sense to its corresponding valency pattern. We also intend to further specify the semantics of Danish verbs in a FrameNet-like project for which we are currently applying for funding. The hypothesis is that groups of verbs sharing the same hyperonym in DanNet as well as the same ontological type, are also candidates to be members of the same Semantic Frames in a Danish FrameNet, meaning that they will share the same semantic roles and to some degree also similar selectional restrictions.

References

1. DDO = Hjorth, E., Kristensen, K. et al. (eds.): Den Danske Ordbog 1-6 (‘The Danish Dictionary 1-6’). Gyldendal & Society for Danish Language and Literature (2003–2005)


Building Croatian WordNet

Ida Raffaelli¹, Marko Tadić¹, Božo Bekavac¹, and Željko Agić²

¹Department of Linguistics
²Department of Information Sciences
Faculty of Humanities and Social Sciences
University of Zagreb, Ivana Lučića 3, Zagreb, Croatia
{ida.raffelli, marko.tadic, bbekavac, zeljko.agic}@ffzg.hr

Abstract. This paper reports on the prototype Croatian WordNet (CroWN). The resource has been collected by translating BCS1 and 2 from English, but also by usage of machine readable dictionary of Croatian language which was used for automatical extraction of semantic relations and their inclusion into CroWN. The paper presents the results obtained, discusses some problems encountered along the way and points out some possibilities of automated acquisition and populating synsets and their refinement in the future. In the second part the paper discusses the lexical particularities of Croatian, which are also shared between other Slavic languages (verbal aspect and derivation patterns), and points out the possible problems during the process of their inclusion in CroWN.

Keywords: WordNet, Croatian language, lexical semantics.

1 Introduction

WordNet has become one of the most valuable resources in any language for which the language technologies are tried to be built. One could say that having in mind the state-of-the-art in LT, a WordNet for a particular language could be considered as one of the basic lexical resources for that language. Semantically organized lexicons like WordNets can have a number of applications such as semantic tagging, word-sense disambiguation, information extraction, information retrieval, document classification and retrieval, etc. In the same time carefully designed and created WordNet represents one of possible models of a lexical system of a certain language and this pure linguistic value is sometimes being neglected or forgotten.

Following, but also widening the original Princeton design of WordNet for English [7], since EuroWordNet [18], a multilingual approach in building WordNets has taken the ground resulting in number of coordinated efforts for more than one language such as BalkaNet [17], MultiWordNet [9]. A comprehensive list of WordNet building initiatives is available at Global WordNet Association web-site¹.

In spite of efforts to coordinate building of WordNets for Central European languages (Polish, Slovak, Slovenian, Croatian, Hungarian) since 2nd GWC in Brno

from 2004, building WordNets for these particular languages have started separately by respective national teams. The Croatian WordNet (CroWN from now on) is being built at the Institute of Linguistics, Faculty of Humanities and Social Sciences at the University of Zagreb. This paper represents the first report on the work-in-progress and the results that it presents are by all means preliminary.

The second section of the paper deals with the method of creating CroWN, dictionaries and corpora used. The third section discusses some particularities of Croatian lexical system that have been observed and which has to be taken into consideration while building the CroWN. The paper ends with future plans and concluding remarks.

2 The Process of Building

2.1 Method

To build a WordNet for a language there are two methods to choose from: 1) expand model [19], which in essence takes the source WordNet (usually PWN) and translates the selected set of synsets into target language and then later expands it with its own lexical semantic additions; and 2) merge model [19], where different separate (sub-)WordNets are being built for specific domains and later merged into a single WordNet. Both approaches have pros and cons with the former being simpler, less time- and man-months (i.e. also financially) consuming, while the latter is usually quite the opposite. On the other hand the results of the former approach are WordNets that are to a large extent at the upper hierarchy levels isomorphous with the source WordNet thus possibly deviating from the real lexical structure of a language. This can be noted particularly in the case of typologically different languages when number of discrepancies starts to grow. The latter approach reflects the lexical semantic structure more realistically but it can be hard to connect it with other WordNets and to make this resource usable for multilingual applications as well.

Having no semantically organized lexicons for Croatian except the [13] which exists only on paper, for initial stages of building CroWN we were forced to use existing monolingual Croatian lexicons which we had in digital form i.e. [1]. Also having very limited human and financial resources we were also forced to opt for expand model but we wanted to keep in mind all the time that it should not be reduced to a mere “copy, paste and translate” operation and that one should always take care about the differences in lexical systems. The expand model has being successfully used in a number of multilingual WordNet projects so we believed that this direction could not be wrong if we also consider thorough manual checking as well.

Up until now our top-down approach has been limited to the translation of BCS1, 2 and 3 from BalkaNet and additional data collecting from dictionary and corpora. The more specialized and more language-specific concepts will be added in further phases of creating CroWN. Table 1. shows basic statistics of POS in BCS1 and BCS2 of CroWN. The BCS3 is not included since it has not been completely adapted.
Table 1. Basic statistics on POS in BCS1 and BCS2 of CroWN.

<table>
<thead>
<tr>
<th></th>
<th>BCS1</th>
<th>BCS2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>965</td>
<td>2245</td>
<td>3210</td>
</tr>
<tr>
<td>Verbs</td>
<td>254</td>
<td>1188</td>
<td>1442</td>
</tr>
<tr>
<td>Adjectives</td>
<td>0</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>1219</td>
<td>3469</td>
<td>4688</td>
</tr>
</tbody>
</table>

2.2 Dictionary and its processing

The only dictionary resource we had available in machine readable form thus usable for populating the CroWN was [1]. Printed and CD-ROM edition of the dictionary contains approximately 70,000 dictionary entries. The right-side of lexicographic articles was divided into several subsections: part-of-speech and other grammatical information, domain abbreviations (e.g. anat. for anatomy), a number of entry definitions (containing various examples and synonyms), syntagms and phraseology, etymology and onomastics. Each of the subsections was labeled in original dictionary using a special symbol, making the dictionary easily processible. After extracting dictionary data and resolving some technical issues, we were left with 69,279 entries as candidates for the first phase of CroWN population. At this step, we omitted grammatical and lexicographic category in formation, phraseology, etymology and onomastics from articles but this information can be easily added later. In Figure 1 both original and simplified dictionary entries are shown:

```plaintext
postanak m
1. pojava, pojavljivanje, nastanak čega
2. prvi trenutak u razvoju čega; postanje
△ Knjiga ~ka prva biblijska knjiga, govori o postanku svijeta
<ENTRY>
postanak
postanak DEF pojava, pojavljivanje, nastanak čega
postanak DEF prvi trenutak u razvoju čega; postanje
postanak SINT Knjiga ~ka prva biblijska knjiga, govori o postanku svijeta
</ENTRY>
```

Fig. 1. Original and reduced dictionary entry.

Each processed lexicographic element in reduced dictionary entry was tagged by the corresponding tag for definition, example and syntagm. Each headword was repeated before DEF and SINT tags, indicating that the definition and syntagm sections are linked to the entry. This redundant form was easily processed with regular patterns (local grammars) using NooJ environment [11]. The starting 69,279 entries now contained 88,352 different definition tags and 7,788 syntagm tags.²

² Note that the overall number of definitions is even bigger since we omitted as redundant the tags in single-line entries, i.e. those entries that contain only the headword and its right-side definition – their processing is trivial.
In this first extraction step we aimed at two things: 1) automatic linking of headwords to their definitions; and 2) creation of a set of well-defined lexical patterns which will be used to acquire additional knowledge from entries using information available in definitions and syntagms sections. We chose definitions and syntagms over all other lexicographic elements as definitions are more likely to contain well-formed word links than phraseology: e.g. the entry crn (en. black) has seven definitions in the dictionary and all of them are starting with koji je (en. which is, that is), providing a constant data extraction pattern. The same procedure is applicable to syntagms – crni humor (en. black humor), crna lista (en. black list), etc.

In dictionary filtering and pattern design, it was our intention to create correct and reliable set of WordNet entries containing basic information – their nearest hypo- and hyperonym classes, basic definitions and possible links to other entries.

In the preliminary test, which was used to determine whether the pattern method is feasible or not, we defined several lexical patterns and using NooJ tested them on our tagged and filtered dictionary. The simple patterns were defined in order to separate animate and inanimate nouns and also to try and link these nouns to other entry types similar in meaning. Some results are given in Table 2.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Extracted</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>onaj koji (en. the one who)</td>
<td>2138</td>
<td>brojač PATTERN broji (en. counter PATTERN counts)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>psiholog PATTERN se bavi psihologijom (en. psychologist PATTERN does psychology)</td>
</tr>
<tr>
<td>osoba koja (en. the person that)</td>
<td>90</td>
<td>korisnik PATTERN se koristi računalom (en. user PATTERN uses a computer)</td>
</tr>
<tr>
<td>osobina onoga koji (je) (en. property of one who (is))</td>
<td>170</td>
<td>aktivnost PATTERN aktivan (en. activity PATTERN active)</td>
</tr>
<tr>
<td>odlika onoga koji (je) (en. quality of one who (is))</td>
<td></td>
<td>budnost PATTERN budan</td>
</tr>
</tbody>
</table>

We can make several conclusions from results of the type given in this table. The first one is that the pattern itself, if well-defined, can provide us with an insight on resulting entries; for example, onaj koji (en. the one who) clearly indicates a person, while osobina onoga koji (en. a quality of one who) indicates a property of an entity. Furthermore, although the [1] dictionary was written using a fairly controlled language subset, our patterns should still undergo parallel expansions in order to handle language variety that occurs in definitions (in Table 1: property, quality could be expanded with feature, attribute, etc.). Patterns should also be tuned with regard to article tokens occurring on its right sides; some of them could capture related nouns (psychologist – psychology) while others could link nouns to adjectives (awakeness – awake). Another possible enhancement to these patterns could be token sensitivity; if the dictionary were to be preprocessed with a PoS/MSD tagger or a morphological lexicon [16], pattern surroundings could be inspected and tokens collected with regard of their MSD and other properties (e.g. obligatory number, case and gender agreement in attribute constructions). Given these facts, we can come out with a conclusion to test: if carefully designed and paired with large, reliable dictionaries and MSD
tagging, pattern detection using local grammars could prove a good method for semi-automated construction of CroWN. Therefore, future dictionary processing and data acquiring tasks will include: enhancing all processing stages in order to collect even more definitions and syntagms that were left behind this first attempt in automatic CroWN population.

2.3 Corpora

We were aware that harvesting semantic relations encoded in the existing machine-readable dictionary, would still not be sufficient for building the exhaustive semantic net as WordNet should be. Therefore we also turned our attention to Croatian corpora and text collections in order to detect more examples and validate the existing ones.

As the treatment of compound words in WordNet from version 1.6. became more important, and since we had developed a system for detecting, collecting and processing compounds words (i.e. syntagms) [5], we decided to include them in CroWN right after completing the translation of BCS1-3. Overview of the compound words in the WordNet and their treatment is described in [10] so we will not go into details here.

When building an ontology from the scratch it is very useful to have a huge source of potential candidates for ontology population. For this task we used the downloaded Croatian edition of Wikipedia (http://hr.wikipedia.org) which at that time comprised 30,985 articles. For identification of distinctive compounds we extracted all explicitly tagged Wikipedia links, that undoubtfully point to a concept which was worded with at least two lower case words. The example can be seen in Figure 2.

![Fig. 2. Example of targeted compound from Wikipedia (circled text ekumenske teologije).](image)

Definition of internal compound structures serves as filter for elimination of unwanted candidates. Examples of such patterns are combinations of MSDs like Adjective + Noun: e.g. plava zastava (en. blue flag); Noun + Noun-in-Genitive: e.g. djeca branitelja (en. children of defenders); Noun+Preposition+Noun-in-case-
governed-by-preposition: e.g. hokej na ledu (en. literally hockey on ice) etc. The compound dictionary collected in this way has also been included in lexical pattern processing of dictionary text described in the previous section.

Since we are still in the process of collecting and processing basic resources to create CroWN, we have not used Croatian National Corpus (HNK) for collecting literals. However it will be used in the process of corpus evidence and validation of literals within synsets used in CroWN.

Of course the last step before the inclusion of new items in CroWN is always the human checking and postprocessing of retrieved candidates where the final judgment about their inclusion and position in CroWN is taking place.

3. Particularities of Croatian

In this part of the paper we would like to discuss some underlying problems that we have detected while we were examining the structure of the Croatian lexical system which could, we believe, be relevant for building WordNets of other languages.

Except the necessity to be compatible with other WordNets, CroWN should preserve and maintain language specificity of Croatian lexical system in order to be a computational lexical database which reflects all semantic specifics of lexical structures in Croatian. Specifics of semantic and lexical structures in Croatian will especially become relevant in the construction of synsets at deeper hierarchical levels. Beside linking synsets with basic relations such as (near)synonyms, hypo/hypernyms, antonyms and meronyms, some of morphosemantic phenomena typical not only for Croatian, but also for other Slavic languages, should be taken into consideration and integrated in the construction of synsets and linking lexical entries within a synset.

Two of the most problematic language-specific phenomena of Croatian (which are shared with other Slavic languages) that should have inevitable impact on creating CroWN are: 1) verbal aspect and 2) derivation. Although these phenomena are traditionally considered as morphological processes, their impact to the semantic structure of a lexical unit should not be neglected in labeling lexical entries in CroWN. Moreover, as we will try to show all of these two morphological processes exhibit some regularity in patterns in Croatian derivation which could be exploited for automatic labeling of lexical entries. Regular derivational patterns characteristic for each morphological category should not be considered without close examination of their role in changing the semantic structure of a certain lexical entry in the CroWN. In other words, regularity of morphosemantic or derivational patterns could be useful for automatic labeling of senses in the CroWN, but at the same time there are many cases in the lexical system where some of these patterns considerably have motivated the change of the meaning from the basic lexical item.

3.1 Verbal Aspect

In one of the most recent Croatian grammars [12] aspect is defined as an instrument to express a difference between an ongoing action (imperfective aspect) and an action that has already been finished (perfective aspect). The category of aspect enables the
division of verbs in Croatian into perfective verbs and imperfective verbs which stand in binary opposition. Perfective verbs could be derived from the imperfective verbs and vice versa, imperfective verbs could be derived from the perfective verbs. Traditionally, aspectual verbal pairs are treated as separate lexical entries and in lexica they are sometimes listed as separate headwords and sometimes under the same headword (usually imperfective). Both practices can exist in the same dictionary in parallel. Some of the most prominent derivational patterns in the formation of both, perfective and imperfective verbs are the following:

1) Perfective verbs could be formed from imperfective verbs by substitution of the suffix of the verbal stem of an imperfective verb. The perfective verb e.g. baci ti (en. to throw) is formed by substitution of the suffix -a of the verbal stem bac a ti (en. to throw as imperfective verb) with the suffix –i. Similar substitutional patterns cover other suffixes.

2) Perfective verbs could be formed by adding the prefix (e.g. pre-, na-, u-, pri-, do-, od-, pro-, etc.) to the verbal stem of an imperfective verb. Many perfective verbs are formed this way: gledati (en. to look) – pregledati (en. to look over, to examine), hodati (en. to walk) – prehodati (en. to walk a distance, used often in a metaphorical sense, meaning to walk a flu over), pisati (en. to write) – prepisati (en. to copy in writing) and many others.

As it could be observed from the previous examples, adding the prefix pre- to the verbal stem of an imperfective verb enables the formation of the perfective verb using regular and frequent derivational pattern, but it also triggers some of not negligible changes of the semantic structure of the basic verbal meaning. If we take the example of the aspect pair pisati (en. to write) – prepisati (en. to copy in writing) the semantic change of the perfective verb prepisati is quite significant with respect to the imperfective verb pisati. Though, there is another derivational pattern for the formation of a perfective verb from the imperfective pisati i.e. it is possible to add the prefix na- to the same verbal stem. The aspect pair pisati – napisati does not exhibit a significant semantic shift of the derived verb toward a new meaning as in the previous case. Moreover, the derivational pattern has been introducing only the distinction between an ongoing and an already finished action.

The same pattern exhibit the aspect pair gledati (en. to look) – pregledati (en. to examine) pointing again to the significant semantic shift of the perfective verb, whereas the aspect pair gledati – pogledati (perfective verb formed by adding the prefix po-) is exclusively related with respect to the differentiation of the type of action.

3) The most prominent pattern of the formation of the imperfective verbs from the perfective ones is the substitution of the suffixes of the verbal stem with derivational morphemes such as -a-, -ava-, and -iva- like in examples: preporuč-i-ti › preporuč-a-ti, prouč-i-ti › prouč-ava-ti and uključ-i-ti › uključ-iva-ti.

It is necessary to point out that this kind of formational pattern does not trigger significant semantic changes of the formed (imperfective) verb. The aspect pair is in binary opposition only with respect to the type of action (prefective or imperfective) they are referring to.

Basically, in Croatian grammars [3] and [12] verbs which differentiate with respect to the type of the action are considered as aspect pairs. However, aspect pairs could also differentiate with respect to the nature of the action or the way the action is
This way of differentiating verbs which form an aspect pair is highly semantically motivated and should be taken into consideration when placing the lexical entries within a synset. For example the aspect pairs *kopati* (en. *to dig*) – *otkopati* (en. *to dig up*), *kopati* – *zakopati* and *kopati* – *pokopati* semantically differentiate primarily with respect to the nature of the action. In the first aspect pair the perfective verb exhibits the beginning of the action (inchoative meaning), the two other pairs exhibit the end of the action (finitive meaning). It should also be pointed out that perfective verbs *zakopati* and *pokopati* do not have the same meaning. The verb *pokopati* means “to bury”, whereas *zakopati* could mean “to bury” but also “to cover with something”.

Grammar [12] distinguishes 11 different meanings of the aspect pairs with respect to the nature of the action and it is clear that this task will not be simple and without problems. The main issue is could we differentiate between these subtle senses using automatic techniques instead of tedious manual validation against the corpora.

### 3.2 Derivation

As Pala and Hlaváčková in [8] point out, derivational relations in highly inflectional languages represent a system of semantic relations that definitely reflects cognitive structures that may be related to language ontology. Derivational processes are deeply integrated in language knowledge of every speaker and represent a system which is morphologically and semantically highly structured. Therefore, as it is stressed in [8], derivational processes can not be neglected in building Czech Wordnet as well as any other Slavic language WordNet.

As already mentioned derivations in Slavic languages are highly regular and are suitable for automatic processing. In the paper [8] 14 (+2) derivational patterns have been adopted as a starting point for the organization of so-called derivational nests of the Czech Wordnet. They are aware of the main problem considering the derivational patterns and relations. Although there exists a significant number of cases where affixes preserve their meaning in Czech as well as in Croatian, it should be taken into consideration that there is also many cases where affixes do not preserve their prototypical meaning and become semantically opaque. This certainly poses a problem for automatic processing of derivational patterns and relations.

If we consider perfixation as one of possible derivational processes in Croatian as in Czech [8] as well, it is without any doubt that prefixes denote different relations such as time, place, course of action, and other circumstances of the main action. There are many cases where prefixes preserve their prototypical meaning, often related to its prototypical meaning as prepositions since most of them developed from prepositions. For example Croatian prefix *na-* has been developed from the preposition *na* with a prototypical meaning referring to the process of directing an object X on the surface of an object Y. There are many verbs in Croatian formed with the prefix *na-* where the prefix has preserved this meaning: *baciti* (en. *to throw*) – *nabaciti* (en. *to throw on sth/smb*), *lijepiti* (en. *to stick*) – *nalijepiti* (en. *to stick sth. on sth*), *skočiti* (en. *to jump*) – *naskočiti* (en. *to jump on sth*).

Unfortunately, this is not the only meaning of the prefix *na-* in Croatian. It also serves for derivation of a large number of verbs meaning *to do sth. to a large extent.*
For example: *krasti* (en. *to steal*) – *nakrasti* (en. *to steal heavily*), *kuhati* (en. *to cook*) – *nakuhati* (en. *to cook lot of food, or to cook for a long time*). In [2] there are three more meanings of the prefix *na-* and they should be integrated in any kind of automatic processing of prefixation in CroWN. Though in our opinion the greatest problem would represent some cases where the same verb, as a result of a prefixation, changes a meaning towards a completely new domain but still preserving some of possible meanings of the prefix.

Such an example is the verb *napustiti*. The verb *pustiti* means *to drop, to let go/loose* while *napustiti* has two semantic cores or two basic meanings. One is related to the first meaning of the prefix *na-* (*to put X on the surface of Y*) and it is *to drop X on the surface of Y*. The other meaning is related to another possible meaning of *na-*; *lead to a result*. So *napustiti* could also mean *to abandon, to quit, to give up*. The connection between two semantic cores is hard to grasp for an average speaker of Croatian, but it could be explained with respect to different meanings of the prefix *na-* in the CroWN the verb *napustiti* should be linked to the verb *pustiti* and its (near)synonyms, as well to verbs such as *ostaviti, odustati* which are both (near)synonyms of *napustiti*. What co-textual patterns will be detected in the corpus and will there be any explicit means to univocally differentiate between these senses remains to be seen.

As shown from the previous examples, derivational patterns such as suffixation and prefixation could not be considered as formal processes using affixes with simple and unique semantic value. Moreover, in highly grammatically motivated languages such as Croatian, as well as in any other Slavic language, suffixation and prefixation should not be regarded as grammatical processes which always result in same transparent and regular semantic changes of the basic lexical item. In many cases affixes used in derivational patterns lose their prototypical meaning enabling significant changes of the semantic structure of the basic lexical item thus influencing the organisation of highly structured morphosemantic relations.

### 4 Future Plans and Concluding Remarks

Being at the very beginning of creating CroWN, this section could be expected to be quite extensive. In order to keep things moderate, we will list only the most imminent future plans to develop CroWN.

The first step would be the digitalization of [13] dictionary and its preprocessing for later usage. Being a lexicographically well-formed dictionary of synonyms in Croatian, this resource would provide us with huge amount of reliable data for direct CroWN synset acquisition and refinement.

The next step is refining and elaborating patterns for extraction of semantic relations from the dictionaries and corpora. This does not only include more complex lexical patterns but also additional dictionaries and corpora including mono- and multilingual such as Croatian-English Parallel Corpus [14] etc.

Particularly important for quality checking of CroWN will be proving the frequency data of literals and their meanings with Croatian reference corpus, namely Croatian National Corpus [15].
We expect to gain some insight also from checking correspondence with WordNets of genetically close languages (Slovenian, Serbian) [6,17] as well as culturally close languages (Slovenian, Czech, Hungarian, German, Italian), particularly at the level of culturally motivated concepts.

In this paper we have presented the first steps in creating Croatian WordNet which consist of translating BCS1, 2 and 3 from English into Croatian. Also we have described procedures for additional synset population from a machine-readable monolingual Croatian dictionary using lexical patterns and regular expressions. Similar procedure has been applied for compound words collection from a semistructured corpus of Croatian Wikipedia articles. Particularities of Croatian and possible problematic issues for defining synset structures are being discussed at the end of the paper with the hope that their solving will lead to a more thorough and precise semantic network of Croatian language.

Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-0645, 130-1300646-1002, 130-1300646-1776 and 036-1300646-1986.

References

Towards Automatic Evaluation of WordNet Synsets

J. Ramanand and Pushpak Bhattacharyya

Department of Computer Science and Engineering,
Indian Institute of Technology, Bombay
{ramanand@it, pb@cse}.iitb.ac.in

Abstract. Increasing and varied applications of WordNets call for the creation of methods to evaluate their quality. However, no such comprehensive methods to rate and compare WordNets exist. We begin our search for WordNet evaluation strategies by attempting to validate synsets. As synonymy forms the basis of synsets, we present an algorithm based on dictionary definitions to verify that the words present in a synset are indeed synonymous. Of specific interest are synsets in which some members “do not belong”. Our work, thus, is an attempt to flag human lexicographers’ errors by accumulating evidences from myriad lexical sources.

1 Introduction

Lexico-semantic networks such as the Princeton WordNet ([1]) are now considered vital resources in several applications in Natural Language Processing and Text mining. Wordnets are being constructed in different languages as seen in the EuroWordNet ([2]) project and the Hindi WordNet ([3]). Competing lexical networks, such as ConceptNet ([4]), HowNet ([5]), MindNet ([6]), VerbNet [7], and FrameNet ([8]), are also emerging as alternatives to WordNets. Naturally, users would be interested in knowing not only the relative merits from among a selection of choices, but also the intrinsic value of such resources. Currently, there are no measures of quality to evaluate or differentiate these resources.

A study of lexical networks could involve understanding the size and coverage, domain applicability, and content veracity of the resource. This is especially critical in cases where WordNets will be created by automated means, especially to leverage existing content in related languages, in contrast to the slower manual process of WordNet creation which has been the traditional method.

The motivation for evaluating WordNets is to help answer questions such as the following:

1. How to select one lexico-semantic network over another?
2. Is a given WordNet sound and complete?
3. Is this resource usable, scalable, and deployable?
4. Is this WordNet suitable for a particular domain or application?
A theory of evaluation must address the following issues:

1. Establishing criteria to measure intrinsic quality of the content held in these lexical networks.
2. Establishing criteria to make useful comparisons between different lexico-semantic networks.
3. Methods to check if a network's quality has improved or declined after content updates.
4. Quality of content in the synsets and relationships between synsets

This paper is organized as follows: In Section 2, we briefly survey work related to the area of ontology evaluation. This is followed in Section 3 by an introduction to the novel problem of validating synonyms in a synset. In Section 4, we describe our dictionary-based algorithm in detail. We discuss the experimental setup and results in Section 5. Finally, in Section 6, we present the key conclusions from our work.

2 Related Work

2.1 Evaluations of Lexico-Semantic Networks

Our literature survey revealed that, to the best of our knowledge, there have been no comprehensive efforts to evaluate WordNets or other lexico-semantic networks on general principles. [9] describes a statistical survey of WordNet v1.1.7 to study types of nodes, dimensional distribution, branching factor, depth and height. A syntactic check and usability study of the BalkaNet resource (WordNets in Eastern European languages) has been described in [10]. The creators of the common-sense knowledge base ConceptNet carried out an evaluation of their resource based on a statistical survey and human evaluation. Their results are described in [4]. [11] discuss evaluations of knowledge resources in the context of a Word Sense Disambiguation task. [12] apply this in a multi-lingual context. Apart from these, we are not aware of any other major evaluations of any lexico-semantic networks.

2.2 Evaluations of Ontologies

In the related field of ontologies, several evaluation efforts have been described. As lexical networks can be viewed as common-sense ontologies, a study of ontology evaluations may be useful. [13] describes an attempt at creating a formal model of an ontology with respect to specifying a given vocabulary's intended meaning. The paper provides an interesting theoretical basis for evaluations. [14] provides a classification of ontology content evaluation strategies and also provides an additional perspective to evaluation based on the “level” of appraisal (such as at the lexical, syntactic, data, design levels). [15] describes some metrics which have been used in the context of ontology evaluation.
Some ontology evaluation systems have been developed and are in use. One of these is OntoMetric ([16]), a method that helps users pick an ontology for a new system. It presents a set of processes that the user should carry out to obtain the measures of suitability of existing ontologies, regarding the requirements of a particular system. The OntoClean ([17]) methodology is based on philosophical notions for a formal evaluation of taxonomical structures. It focuses on the cleaning of taxonomies. [18] describes a task-based evaluation scheme to examine ontologies with respect to three basic levels: vocabulary, taxonomy and non-taxonomic semantic relations. A score based on error rates was designed for each level of evaluation. [19] describes an ontology evaluation scheme that makes it easier for domain experts to evaluate the contents of an ontology. This scheme is called OntoLearn.

We felt that none of the above methods seemed to address the core issues particular to WordNets, and hence we approached the problem by looking at synsets.

3 Synset Validation

3.1 Introduction

Synsets are the foundations of a WordNet. A WordNet synset is constructed by putting together a set of synonyms that together define a particular sense uniquely, as given by the principles of minimality and coverage described in the previous section. This sense is explicitly indicated for human readability by a gloss. For instance, the synset \{proboscis, trunk\} represents the sense of “a long flexible snout as of an elephant”, as opposed to the synset \{luggage compartment, automobile trunk, trunk\} which is “a compartment in an automobile that carries luggage or shopping or tools”. Words with potentially multiple meanings are associated together, out of which a single sense emerges. To evaluate the quality of a synset, we began by looking at the validity of its constituent synonyms.

Before the validation, the following theoretical questions must be addressed:

1. What is the definition of a synonym?
2. What are the necessary and sufficient conditions to determine that synonymy exists among a group of words?

Intuitively, synonymy exists between two words when they share a similar sense. This also implies that one word can be replaced by its synonym in a context without any loss of meaning. In practice, most words are not perfect replacements for their synonyms i.e. they are near synonyms. There could be contextual, collocational and other preferences behind replacing synonyms. [20] describes attempts to mathematically describe synonymy. To the best of our knowledge, no necessary and sufficient conditions to prove that two words are synonyms of each other have been explicitly stated.
The foundation of our work is the following: we conjecture that

1. if two words are synonyms, it is necessary that they must share one common meaning out of all the meanings they could possess.
2. A sufficient condition could be showing that the words replace each other in a context without loss of meaning.

The task of synset validation has the following subtasks:

1. Are the words in a synset indeed synonyms of each other?
2. Are there any words which have been omitted from the synset?
3. Does the combination of words indicate the required sense?

In this paper, we attempted to answer the first question above i.e. given a set of words, could we verify if they were synonyms? Our literature survey revealed that though much work had been done in the automated discovery of synonyms (from corpora and dictionaries), no work had been done in automatically verifying whether two words were synonyms. Nevertheless, we began by studying some of the synonym discovery methods available.

3.2 Related Work on Automatic Synset Creation

All these methods are based on web and corpora mining. [21] describes a method to collect synonyms in the medical domain from the Web by first building a taxonomy of words. [22] provides an unsupervised learning method for extracting synonyms from the Web. [23] shows an interesting topic signature method to detect synonyms using document contexts and thus enrich large ontologies. Finally, [24] is a survey of different synonym discovery methods, which also proposes its own dictionary-based solution for the problem. Its dictionary based approach provides some useful hints for our own experiments in synonymy validation.

3.3 Our Approach

We focus only on the problem of checking whether the words in a synset can be shown to be synonyms of each other and thus correctly belong to that synset. As of now, we do not flag omissions in the synsets. It is to be also noted that failure to validate the presence of a word in a synset does not strongly suggest that the word is incorrectly entered in the synset - it merely raises a flag for human validation.

The input to our system is a WordNet synset which provides the following information:

1. The synonymous words in the synset
2. The hyponym(s) of the synset
3. Other linked nodes, gloss, example usages
The output consists of a verdict on each word as to whether it fits in the synset, i.e. whether it qualifies to be the synonym of other words in the synset, and hence, whether it expresses the sense represented by the synset. A block diagram of the system is shown in Fig.1.

![Block Diagram for Synset Synonym Validation](image)

**Fig. 1.** Block Diagram for Synset Synonym Validation

## 4 Our Dictionary-based Algorithm

### 4.1 The Basic Idea

In dictionaries, a word is usually defined in terms of its hypernyms or synonyms. For instance, consider definitions of the word *snake*, whose hypernym is *reptile*, and its synonyms *serpent* and *ophidian* (obtained from the website Dictionary.com [22]):

*snake:* any of numerous limbless, scaly, elongate *reptiles* of the suborder Serpentes, comprising venomous and non-venomous species inhabiting tropical and temperate areas.

*serpent:* a *snake*

*ophidian:* A member of the suborder Ophidia or Serpentes; a *snake.*

This critical observation suggests that dictionary definitions may provide useful clues for verifying synonymy.

We use the following hypothesis:

*if a word is present in a synset, there is a dictionary definition for it which refers to its hypernym or to its synonyms from the synset.*

Instead of matching synonyms pair-wise, we try to validate the presence of the word in the synset using the hypernyms of the synset and the other synonyms in the synset. A given word belongs to a given synset if there exists a definition for that word, which refers to one of the given hypernym words or one of the synonyms. We
use the hypernyms and synonyms to validate other synonyms by mutual reinforcement.

4.2 Algorithm Description

The dictionary-based algorithm consists in applying three groups of rules in order. The first group applies to each word individually, using its dictionary definitions. The second group relies on a set of words collected for the entire synset during the application of the first group. The final group consists of rules that do not use the dictionary definitions. (All definitions in this section are from the website Dictionary.com [25].)

In this section, we describe the steps of the algorithm with examples. The Algorithm has been stated in Section 4.3.

Group 1

Rule 1 - Hypernyms in Definitions

Definitions of words for particular senses often make references to the hypernym of the concept. Finding such a definition means that the word's placement in the synset can be defended.

e.g.
Synset: {brass, brass instrument}
Hypernym: {wind instrument, wind}
Relevant Definitions:
brass instrument: a musical wind instrument of brass or other metal with a cup-shaped mouthpiece, as the trombone, tuba, French horn, trumpet, or cornet.

Rule 2 - Synonyms in Definitions

Definitions of words also make references to fellow synonyms, thus helping to validate them.

e.g.
Synset: {anchor, ground tackle}
Hypernym: {hook, claw}
Relevant Definitions:
ground tackle: equipment, as anchors, chains, or windlasses, for mooring a vessel away from a pier or other fixed moorings.

Rule 3 - Reverse Synonym Definitions

Definitions of synonyms may also make references to the word to be validated.

e.g.
Synset: {Irish Republican Army, IRA, Provisional Irish Republican Army, Provisional IRA, Provos}
Hypernym: {terrorist organization, terrorist group, foreign terrorist organization, FTO}
Relevant Definitions:

*Irish Republican Army*: an underground Irish nationalist organization founded to work for Irish independence from Great Britain: declared illegal by the Irish government in 1936, but continues activity aimed at the unification of the Republic of Ireland and Northern Ireland.

*Provos*: member of the Provisional wing of the *Irish Republican Army*.

Here *Irish Republican Army* can be validated using the definition of *Provos*.

**Rules 4 and 5 - Partial Hypernyms and Synonyms in Definitions**

Many words in the WordNet are multi-words, i.e., they are made up of more than one word. In quite a few cases, such multi-word hypernyms are not entirely present in the definitions of words, but parts of them can be found in the definition.

e.g.
Synset: *{fibrinogen, factor I}*
Hypernym: *{coagulation factor, clotting factor}*
Relevant Definitions:
*fibrinogen*: a globulin occurring in blood and yielding fibrin in blood *coagulation*.

**Group 2**

**Rule 6 – Bag of Words from Definitions**

In some cases, definitions of a word do not refer to synonyms or hypernym words. However, the definitions of two synonyms may share common words, relevant to the context of the sense. This rule captures this case.

When a word is validated using Group 1 rules, the words of the validating definition are added to a collection. After applying Group 1 rules to all words in the synset, a bag of these words (from all validating definitions seen so far) is now available. For each remaining synonym yet to be validated, we look for any definition for it which contains one of the words in this bag.

e.g.
Synset: *{serfdom, serfhood, vassalage}*
Hypernym: *{bondage, slavery, thrall, thralldom, thraldom}*
Relevant Definitions
*serfdom*: person (held in) *bondage*; *servitude*
*vassalage*: dependence, subjection, *servitude*

*serfdom* is matched on account of its hypernym *bondage* being present in its definition. So the Bag of Words now contains "*person, bondage, servitude*".

No definition of *vassalage* could be matched with any of the rules from 1 to 5. But Rule 6 matches the word *servitude* and so helps validate the word.
Group 3

Rules 7 and 8 - Partial Matches of Hypernyms and Synonyms

Quite a few words to be validated are multi-words. Many of these do not have definitions present in conventional dictionaries, which make the above rules inapplicable to them. Therefore, we use the observation that, in many cases, these multi-words are variations of their synonyms of hypernyms i.e. the multi-words share common words with them. Examples of these are synsets such as:

1. \{dinner theater, dinner theatre\}: No definition was available for dinner theatre, possibly because of the British spelling.
2. \{laurel, laurel wreath, bay wreath\}: No definitions for the two multi-words.
3. \{Taylor, Zachary Taylor, President Taylor\}: No definition for the last multi-word.

As can be seen above, the multi-word synonyms do share partial words. To validate such multi-words without dictionary entries, we check for the presence of partial words in their synonyms.

e.g.
Synset: \{Taylor, Zachary Taylor, President Taylor\}
Hypernym: \{President of the United States, United States President, President, Chief Executive\}
Relevant Definitions:
Taylor, Zachary Taylor. (1784-1850) the 12th President of the United States from 1849-1850.
President Taylor: - no definition found -

The first two words have definitions which are used to easily validate them. The third word has no definition, and so rules from Group 1 and 2 do not apply to it. Applying the Group 3 rules, we look for the component words in the other two synonyms. Doing this, we find “Taylor” in the first synonym, and hence validate the third word.

A similar rule can be defined for a multi-word hypernym, wherein we look for the component word in the hypernym words. In this case, we would match the word “President” in the first hypernym word.

We must note that, in comparison to the other rules, these rules are likely to be susceptible to erroneous decisions, and hence a match using these rules should be treated as a weak match. The reason for creating these two rules is to overcome the scarcity of definitions for such multi-words.
4.3 Algorithm Statement

Algorithm 1 – Validating WordNet synsets using a dictionary

1: Input: synset $S$, words $W$ in synset $S$, Dictionary of definitions

2: For each word $w$ belonging to $W$ do

3: Apply rules in Group 1:
   - 3.1: (Rule 1) Find a definition for $w$ in the dictionary such that it contains a hypernym word $h$ (repeat with other hypernyms if necessary)
   - 3.2: (Rule 2) Else, find a definition for $w$ containing any synonym of $w$ from the synset
   - 3.3: (Rule 3) Else, find a synonym's definition referring to $w$
   - 3.4: (Rule 4) (applicable to multi-words in the hypernym) Else, find a definition of $w$ referring to a partial word from a multi-word in the hypernym
   - 3.5: (Rule 5) (applicable to synonyms that are multi-words) Else, find a definition for $w$ referring to a partial word from a multi-word synonym

4: Apply the rule 6 in Group 2:
   - 4.1: For every word $m$ from the synset that was matched by one of the above rules, add the words in the validating definition for $m$ to a collection of words $C$.
   - 4.2: For each word $w$ in the synset that has not been validated, find a definition $d$ of $w$ such that $d$ has a word appearing in $C$.

5: Apply rules in Group 3 to each remaining unmatched word $w$:
   - 5.1: (Rule 7) See if a partial word from a multi-word $w$ is found in the synonym to be matched
   - 5.2: (Rule 8) Else, see if a partial word from the multi-word $w$ is found in a hypernym word $h$.

6: end for
5 Experimental Results

5.1 Setup

The validation was tested on the Princeton WordNet (v2.1) noun synsets. Out of the 81426 noun synsets, 39840 are synsets with more than one word – only these were given as input to the validator. This set comprised of a total of 103620 words.

One of the contributions of our work is the creation of a super dictionary which consists of words and their definitions constructed by automatic means from the online dictionary service Dictionary.com ([25]) (which aggregates definitions from various sources such as Random House Unabridged Dictionary, American Heritage Dictionary, etc.) Of these, definitions from Random House and American Heritage dictionaries were identified and added to the dictionary being created. English stop words were removed from the definitions, and the remaining words were stemmed using Porter's stemmer [26]. The resulting dictionary had 463487 definitions in all for a total of 49979 words (48.23% of the total number of words).

5.2 Results and Discussions

Figs. 2, 3, and 4 summarise the main results obtained by running the dictionary-based validator on 39840 synsets. As shown in Fig. 2, 14844 out of the 18322 unmatched words did not have definitions in the dictionary. Therefore, there are 88776 words which either have definitions in the dictionary, or are referenced in the dictionary, or are matched by the partial rules. So, considering only these 88776 words, there are 85298 matched words, i.e. a validation value of 96.08%.

In about 9% of all synsets, none of the words in the synset could be verified. Of these 3660 synsets, 2952 (80%) had only 2 words in them. The primary reason for this typically was one member of the synset not being present in the dictionary, and hence reducing the number of rules applicable to the other word.

Failure to validate a word does not mean that the word in question is incorrectly present in the synset. Instead, it flags the need for human intercession to verify whether the word indeed has that synset's sense. The algorithm is not powerful enough to make a firm claim of erroneous placement. In an evaluation system, the validator can serve as a useful first-cut filter to reduce the number of words to be scrutinised by a human expert. In some cases, the non-matches did raise some interesting questions about the validity of a word in a synset. We discuss some examples in the next section.
Fig. 2. The Dictionary Approach: Summary of results

Fig. 3. The Dictionary Approach: A synset perspective

Fig. 4. The Dictionary Approach: Rule-wise summary
5.3 Case Studies

(All sources for definitions in the following examples are from the website Dictionary.com [25])

5.3.1 Possible True Negatives flagged by the validator

The validator could not match about 18% of all words. In most of these cases, the words are indeed correctly placed (as one would expect of a resource manually created by experts) but are flagged incorrectly by the validator as it is as yet not powerful enough to match them. However, consider the following cases of words where non-matches are interesting to study.

Instance 1:
Synset: {visionary, illusionist, seer}
Hypernym: {intellectual, intellect}
Gloss: a person with unusual powers of foresight

The word “illusionist” was not matched in this context. This seems to be a highly unusual sense of this word (more commonly seen in the sense of “conjuror”). None of the dictionaries consulted provided this meaning for the word.

Instance 2:
Synset: {bobby pin, hairgrip, grip}
Hypernym: {hairpin}
Gloss: a flat wire hairpin whose prongs press tightly together; used to hold bobbed hair in place

It could not be established from any other lexical resource whether grip, though a similar sounding word to hairgrip, was a valid synonym for this sense. Again, this could be a usage local to some cultures, but this was not readily supported by other dictionaries.

5.3.2 True Positives correctly flagged by the validator

Here are examples of the validator correctly flagging matches.

Instance 1
Synset: {smokestack, stack}
Word to be validated: smokestack
Hypernym: {chimney}
Relevant Definitions:
smokestack: A large chimney or vertical pipe through which combustion vapors, gases, and smoke are discharged.
Instance 2
Synset: {zombi, zombie, snake god}
Word to be validated: snake god
Hypernym: {deity, divinity, god, immortal}
Relevant Definitions:
  zombie: a snake god worshiped in West Indian and Brazilian religious practices of African origin.

5.3.2 False Negatives flagged by the validator
Here are examples of the validator being unable to match words, despite definitions being present:

Instance 1
Synset: {segregation, separatism}
Word to be validated: segregation
Hypernym: {social organization, social organisation, social structure, social system, structure}
Relevant Definitions:
  segregation: The act or practice of segregating
  segregation: the state or condition of being segregated

Noun forms of such verbs typically refer to the act, which makes it hard to validate using other words.

Instance 2
Synset: {hush puppy, hushpuppy}
Word to be validated: hush puppy
Hypernym: {cornbread}
Relevant Definitions:
  Hush puppy: a small, unsweetened cake or ball of cornmeal dough fried in deep fat.

Establishing the similarity between cornmeal and cornbread would have been our best chance to validate this word. Currently, we are unable to do this.

6 Conclusions and Future Work

Our observations show that the intuitive idea behind the algorithm holds well. The algorithm is quite simple to implement. No interpretation of numbers is required; the process is just a simple test. The algorithm is heavily dependent on the depth and quality of dictionaries being used. WordNet has several words that were not present in conventional dictionaries available on the Web. Encyclopaedic entries such as Mandara (a Chadic language spoken in the Mandara mountains in Cameroon), domain-specific words, mainly from agriculture, medicine, and law, such as ziziphus jujuba (spiny tree having dark red edible fruits) and pediculosis capitis (infestation of
the scalp with lice), phrasal words such as caffeine intoxication (sic) were among those not found in the collected dictionary.

Since the Princeton WordNet is manually crafted by a team of experts, we do not expect to find too many errors. However, many of the words present in the dictionary and not validated were those with rare meanings and usages. Our method makes it easier for human validators to focus on such words. This will especially be useful in validating the output of automatic WordNet creations.

The algorithm cannot yet detect omissions from a synset, i.e. the algorithm does not discover potential synonyms and compare them with the existing synset.

Possible future directions could be expanding the synset validation to other parts of a synset such as the gloss and relations to other synsets. The results could be summarized into a single number representing the quality of the synsets in the WordNet. The results could then be correlated with human evaluation, finally converging to a score that captures the human view of the WordNet.

The problem of scarcity of definitions could be further addressed by adding more dictionaries and references to the set of sources.

The presented algorithm is available only for English WordNet. However, the approach should broadly apply to other language WordNets as well. The limiting factors are the availability of dictionaries and tools like stemmers for those languages. Similarly, the algorithm could be used to verify synonym collections such as in Roget's Thesaurus and also other knowledge bases. The algorithm has been executed on noun synsets; they can also be run on synsets from other parts of speech.

We see such evaluation methods becoming increasingly imperative as more and more WordNets are created by automated means.

Acknowledgments

We would like to express our gratitude to Prof. Om Damani (CSE, IIT Bombay) and Sagar Ranadive for their valuable suggestions towards this work.

References

Lexical Enrichment of a Human Anatomy Ontology using WordNet

Nils Reiter\textsuperscript{1} and Paul Buitelaar\textsuperscript{2}

\textsuperscript{1} Department of Computational Linguistics, Heidelberg University, Heidelberg, Germany
\texttt{reiter@cl.uni-heidelberg.de}

\textsuperscript{2} Language Technology Lab & Competence Center Semantic Web, DFKI, Saarbrücken, Germany
\texttt{paulb@dfki.de}

Abstract. This paper is concerned with lexical enrichment of ontologies, i.e. how to enrich a given ontology with lexical entries derived from a semantic lexicon. We present an approach towards the integration of both types of resources, in particular for the human anatomy domain as represented by the Foundational Model of Anatomy (FMA). The paper describes our approach on combining the FMA with WordNet by use of a simple algorithm for domain-specific word sense disambiguation, which selects the most likely sense for an FMA term by computing statistical significance of synsets on a corpus of Wikipedia pages on human anatomy. The approach is evaluated on a benchmark of 50 ambiguous FMA terms with manually assigned WordNet synsets (i.e. senses).

1 Introduction

This paper is concerned with lexical enrichment of ontologies, i.e. how to enrich a given ontology with lexical entries derived from a semantic lexicon. The assumption here is that an ontology represents domain knowledge with less emphasis on the linguistic realizations (i.e. words) of knowledge objects, whereas a semantic lexicon such as WordNet defines lexical entries (words with their linguistic meaning and possibly morpho-syntactic features) with less emphasis on the domain knowledge associated with these.

1.1 Ontologies

An ontology is an explicit and formal description of the conceptualization of a domain of discourse (see e.g. Gruber [1], Guarino [2]). In its most basic form an

\textsuperscript{**} This work has been done while the first author was affiliated to the Department of Computational Linguistics at Saarland University.
ontology consists of a set of classes and a set of relations that describe the properties of each class. Ontologies formally define relevant knowledge in a domain of discourse and can be used to interpret data in this domain (e.g. medical data such as patient reports), to reason over knowledge that can be extracted or inferred from this data and to integrate extracted knowledge with other data or knowledge extracted elsewhere. With recent developments towards knowledge-based applications such as intelligent Question Answering, Semantic Web applications and semantic-level multimedia indexing and retrieval, the interest in large-scale ontologies has increased. Here we use a standard ontology in human anatomy, the FMA: Foundational Model of Anatomy\(^3\) ([3]). The FMA describes the domain of human anatomy in much detail by way of class descriptions for anatomical objects and their properties. Additionally, the FMA lists terms in several languages for many classes, which makes it a lexically enriched ontology already. However, our main concern here is to extend this lexical representation further by automatically deriving synonyms from WordNet.

1.2 Lexicons

A lexicon describes the linguistic meaning and morpho-syntactic features of words and possibly also of more complex linguistic units such as idioms, collocations and other fixed phrases. Semantically organized lexicons such as WordNet and FrameNet define word meaning through formalized associations between words, i.e. in the form of synsets in the case of WordNet and with frames in the case of FrameNet. Although such a representation defines some semantic aspects of a word relative to other words, it does not represent any knowledge about the objects that are referred to by these words. For instance, the English noun “ball” may be represented by two synsets in WordNet (\{ball, globe\}, \{ball, dance\}) each of which reflects another interpretation of this word. However, deeper knowledge about what a “ball” in the sense of a “dance” involves (a group of people, a room, music to which the group of people move according to a certain pattern, etc.) cannot be represented in this way. Frame-based definitions in FrameNet allow for such a deeper representation into some respect, but also in this case the description of word meaning is concerned with the relation between words and not so much with object classes and their properties that are referred to by these words. However, for instance in the case of BioFrameNet ([4]), an extension of FrameNet for use in the biomedical domain, such an attempt has been made and is therefore much in line with our work described here.

1.3 Related Work

Other related work is on word sense disambiguation (WSD) and specifically domain-specific WSD as this is a central aspect of our algorithm in selecting the

\(^3\) See [http://sig.biostr.washington.edu/projects/fm/AboutFM.html](http://sig.biostr.washington.edu/projects/fm/AboutFM.html) for more details on the FMA
most likely sense of words occurring in FMA terms. The work presented here is based directly on [5] and similar approaches ([6, 7]). Related to this work is the assignment of domain tags to WordNet synsets ([8]), which would obviously help in the automatic assignment of the most likely sense in a given domain – as shown in [9]. An alternative to this idea is to simply extract that part of WordNet that is directly relevant to the domain of discourse ([10, 11]). However, more directly in line with our work on enriching a given ontology with lexical information derived from WordNet is presented in [12], but the main difference here is that we use a domain corpus as additional evidence for statistical significance of a selected word sense (i.e. synset). Finally, also some recent work on the definition of ontology-based lexicon models ([13–15]) is of (indirect) relevance to the work presented here as the derived lexical information needs to be represented in such a way that it can be easily accessed and used by NLP components as well as ontology management and reasoning tools.

2 Approach

Our approach to lexical enrichment of ontologies consists of a number of steps, each of which will be addressed in the remainder of this section:

1. extract all terms from the term descriptions of all classes in the ontology, lookup of terms in WordNet
2. for ambiguous terms: apply domain-specific WSD by ranking senses (synsets) according to statistical relevance in the domain corpus
3. select most relevant synset and add the synonyms of this synset to the corresponding term representation

2.1 Term Extraction and WordNet Lookup

Ontologies, such as the FMA, describe objects and their relations to each other. Additionally, each such object (or rather the class descriptions for such objects) may carry terminological information in one or more languages. In the FMA, terms for classes are defined in several languages, i.e. 100,000 English terms, 8,000 Latin, 4,000 French, 500 Spanish and 300 German terms. Terms in the FMA can be simple, consisting of just one word, or complex multiword terms, e.g. “muscular branch of lateral branch of dorsal branch of right third posterior intercostal artery”. In our approach we considered simple as well as complex terms although only a small number of such domain-specific terms will actually occur in WordNet as will be reported below in section 3.

2.2 WSD Algorithm

The core of our approach is the word sense disambiguation algorithm as shown in figure 1. The algorithm iterates over every synonym of every synset of the
term in question. It calculates the $\chi^2$ values of each synonym and adds them up for each synset.

```javascript
function getWeightForSynset(synset) {
    synonyms = all synonyms of synset
    weight = 0
    foreach synonym in synonyms
        c = chi-square(synonym)
        weight = weight + c
    end foreach
    return weight
}
```

$s = \text{synsets to which } t \text{ belongs}$

```javascript
highest_weight = 0
best_synsets = {}
foreach synset in $s$
    synonyms = all synonyms of synset
    weight = getWeightForSynset(synset)
    if (weight == highest_weight)
        best_synsets = best_synsets + { synset }
    else if (weight > highest_weight)
        best_synsets = { synset }
    end if
end foreach
return best_synsets
```

Fig. 1. Algorithm for the sense disambiguation of the term $t$

Using the $\chi^2$-test (see, for instance, [16, p. 169]), one can compare the frequencies of terms in different corpora. In our case, we use a reference and a domain corpus and assume that the terms occurring (relatively) more often in the domain corpus than in the reference corpus are “domain terms”, i.e., are specific to this domain. If it is a domain term, it should be defined in the ontology.

$$\chi^2(t) = \frac{N \cdot (O_{11}^t O_{22}^t - O_{12}^t O_{21}^t)^2}{(O_{11}^t + O_{12}^t)(O_{11}^t + O_{21}^t)(O_{12}^t + O_{22}^t)(O_{21}^t O_{22}^t)}$$

(1)

$\chi^2$, calculated according to formula 1, allows us to measure exactly this. $O_{11}^t$ and $O_{12}^t$ denote the frequencies of the term $t$ in the domain and reference corpora.
while $O_{11}^t$ and $O_{22}^t$ denote the frequency of any term but $t$ in the domain and reference corpora:

\[
\begin{align*}
O_{11}^t &= \text{frequency of } t \text{ in the domain corpus} \\
O_{12}^t &= \text{frequency of } t \text{ in the reference corpus} \\
O_{21}^t &= \text{frequency of } \neg t \text{ in the domain corpus} \\
O_{22}^t &= \text{frequency of } \neg t \text{ in the reference corpus} \\
N &= \text{Added size of the two corpora}
\end{align*}
\]

The algorithm finally chooses the synset with the highest weight as the appropriate one.

The term “gum”, for instance, has six noun senses with on average 2 synonyms. The $\chi^2$ value of the synonym “gum” itself is 6.22. Since this synonym occurs obviously in every synset of the term, it makes no difference for the rating. But the synonym “gingiva”, which belongs to the second synset and is the medical term for the gums in the mouth, has a $\chi^2$ value of 20.65. Adding up the domain relevance scores of the synonyms for each synsets, we find that the second synset gets the highest weight and is therefore selected as the appropriate one.

Relations The algorithm as shown in figure 1 uses the synonyms found in WordNet. However, other relations that are provided by WordNet can be used as well. Figure 2 shows the improved algorithm. The main difference is that we calculate and add the weights for each synonym of each synset to which a synset of the original term is related.

2.3 Lexical Representation

Finally, after the synsets for an ambiguous term $t$ have been ranked according to relevance to the domain, we can select the top one or more to be included as (additional) lexical/terminological information in the ontology, i.e., the synonyms that are contained in this synset can be added as (further) terms for the ontology class $c$ that corresponds to term $t$.

Here, we actually propose to extend the ontology with an ontology-based lexicon format, LingInfo, which has been developed for this purpose in the context of previous work ([15]). By use of the LingInfo model we will be able to represent each synonym for $t$ as a linguistic object $l$ that is connected to the corresponding class $c$. The object $l$ is an instance of the LingInfo class of such linguistic objects that cover the representation of the orthographic form of terms as well as relevant morpho-syntactic information, e.g., stemming, head-modifier decomposition, part-of-speech. The implementation of a LingInfo-based linguistic knowledge base for the FMA is ongoing work, but a first version of a similar
3 Experiment

In an empirical experiment, we enrich the FMA ("Foundational Model of Anatomy") ontology with lexical information (synonyms) derived from WordNet using Wikipedia pages on human anatomy as domain corpus.

3.1 Data Sources

Ontology: Foundational Model of Anatomy  "The Foundational Model of Anatomy (FMA) ontology was developed by the Structural Informatics Group at the University of Washington. It contains approximately 75,000 classes and over 120,000 terms; over 2.1 million relationship instances from 168 relationship types link the FMA’s classes into a coherent symbolic model. The FMA is one of the largest computer-based knowledge sources in the biomedical sciences. The most comprehensive component of the FMA is the Anatomy taxonomy" (FMA

4 http://sig.biostr.washington.edu/index.html
Lexical Enrichment of a Human Anatomy Ontology using WordNet

Semantic Lexicon: WordNet  
The most recent version of WordNet (3.0) was used in our experiment. As an interface to our own implementation, we use the Java WordNet interface\textsuperscript{5}. The number of English terms (simple and complex) we were able to extract from the FMA was 120,417, of which 118,785 were not in WordNet. This left us with a set of 1,382 terms that were in WordNet but only 250 of these were actually ambiguous and therefore of interest to our experiment. Interestingly, 10 of these were in fact multiword terms. The experiment as reported below is therefore concerned with the disambiguation of these 250 FMA terms, given their sense assignments in WordNet.

Medical Corpus: Wikipedia Pages on Human Anatomy  
Our approach requires the use of a domain corpus. As corpus from the anatomy domain, we use the Wikipedia pages from the category “Human Anatomy” and all its subordinated categories\textsuperscript{6}. These are 7,251 single pages, containing over 4.4 million words.

We removed the meta information (categories, tables of contents, weblinks, ...) using heuristic methods. By use of part of speech tagging with the TreeTagger ([18]), we automatically extracted all nouns from this corpus, resulting in 1.3 million noun tokens and 92,927 noun types.

Reference Corpus: British National Corpus  
Our ranking of the domain relevance of a synset is based on comparing the frequencies of its synonyms in a domain corpus and a reference corpus. The reference corpus we use is the British National Corpus (BNC). Since we were only interested in the frequencies, we used the frequency lists provided by [19].

3.2 Benchmark  
Our benchmark (gold standard) consists of randomly selected 50 ambiguous terms from the ontology. Four terms have been removed from the test set because none of their senses belong to the domain of human anatomy.

\textsuperscript{5} http://www.mit.edu/~markaf/projects/wordnet/  
\textsuperscript{6} http://en.wikipedia.org/wiki/Category:Human_anatomy
Two annotators manually disambiguated them according to the domain of human anatomy. Each term is associated with one (or more) WordNet synsets. More than one synset is used due to the high granularity of WordNet. The agreement between the two annotators is, generally speaking, high. In every single case, there is an overlap in the associated synsets, i.e., for every term, there is at least one synset chosen by both annotators. If we count only a perfect match, i.e., both annotators chose exactly the same set of senses, the kappa value $\kappa$ according to [20] is still $\kappa = 0.71$.

**Baseline** The synsets in WordNet are sorted according to frequency. The word “jaw”, for instance, occurs more often with its first synset than with its third. It is therefore a reasonable assumption for any kind of word sense disambiguation to pick always the first sense (see, for instance, [21]). We use this simple approach as baseline for our evaluation.

### 3.3 Evaluation

The system was evaluated with respect to precision, recall and f-score. Precision is the proportion of the meanings predicted by the system which are correct. Recall is the proportion of correct meanings which are predicted by the system. Finally, the f-score is the harmonic mean of precision and recall, and is the final measure to compare the performance of systems.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (first sense)</td>
<td>58.69</td>
<td>46.56</td>
<td>51.93</td>
</tr>
<tr>
<td>Only Synonyms</td>
<td>54.78</td>
<td>65.58</td>
<td>59.70</td>
</tr>
<tr>
<td>Hypernym</td>
<td>56.52</td>
<td>47.10</td>
<td>51.38</td>
</tr>
<tr>
<td>Hypernym (instance)</td>
<td>53.70</td>
<td>63.41</td>
<td>59.22</td>
</tr>
<tr>
<td>Hyponym</td>
<td>64.93</td>
<td>63.95</td>
<td>64.44</td>
</tr>
<tr>
<td>Topic</td>
<td>56.96</td>
<td>67.75</td>
<td>61.89</td>
</tr>
<tr>
<td>Holonym (part)</td>
<td>63.04</td>
<td>63.41</td>
<td>63.22</td>
</tr>
<tr>
<td>Holonym (substance)</td>
<td>55.87</td>
<td>65.58</td>
<td>60.34</td>
</tr>
<tr>
<td>Meronym (member)</td>
<td>52.61</td>
<td>63.41</td>
<td>57.51</td>
</tr>
<tr>
<td>Meronym (part)</td>
<td>58.05</td>
<td>68.12</td>
<td>62.68</td>
</tr>
<tr>
<td>Meronym (substance)</td>
<td>55.51</td>
<td>62.32</td>
<td>58.72</td>
</tr>
<tr>
<td>All other</td>
<td>54.78</td>
<td>65.58</td>
<td>59.70</td>
</tr>
<tr>
<td>Hyponym, Holonym (part), Meronym (part)</td>
<td>77.53</td>
<td>70.11</td>
<td>73.63</td>
</tr>
<tr>
<td>Topic, Holonym (substance), Meronym (part)</td>
<td>61.30</td>
<td>70.29</td>
<td>65.49</td>
</tr>
</tbody>
</table>
We calculated precision, recall and f-score separately for WordNet relations and test items. The test item “jaw”, for instance, was manually disambiguated to the first or second noun synset of the word. Our program, using the hyponymy relation, returned only the first noun synset. For this item, we calculate a precision of 100% and a recall of 50%.

Table 1 shows the results averaged over all 46 test items for the different relations. The relations not shown did not give other results than the algorithm without using any of the relations (results for “all other” are exactly the same as “only synonyms”).

The two lines at the bottom of the table are combinations of relations. In the first line, we use the three relations with the highest precision (and f-score, but that is a coincidence) together (hyponym, holonym (part) and meronym (part)). The last line shows the three relations with the highest recall taken together (topic, holonym (substance) and meronym (part)). Note, that the meronym (part) relation is the only relation that is among the top three in both cases.

3.4 Discussion

Our results show – in almost any configuration – a clear improvement compared to the baseline.

Using just the synonyms of WordNet and no additional relation(s), we observe an increase in recall (around 20%) and a relatively small decrease in precision (less than 5%). The increase in recall can easily be explained by the fact that our baseline takes only the first (and therefore: only one) synset – every term that is disambiguated to more than one synset already gets a recall of 50% or less. The decrease in precision can be explained by looking at the test samples. For some of the synsets, a synonym – especially when it comes to multi-word expressions – can not be found in the corpus. This leads to the same weight for a number of synsets and thus to more selected synsets, even if the evidence does not increase. The precision decreases because among the selected synsets, there are more inappropriate ones. Or, the other way around: if an appropriate synset has no synonyms (or only synonyms that do not appear in the corpus), the precision decreases.

![Fig. 3. Synonyms for the synsets of “alveolus”](image-url)
For the term “alveolus”, for instance, both noun synsets are annotated as appropriate in the gold standard. The baseline algorithm selects only the first synset and gets therefore a precision of 100% and a recall of 50%. Figure 3 shows the synonyms for the term “alveolus” graphically. In the configuration, where we just use the WordNet synonyms, both synsets get the same weight, because the synonym alveolus appears 137 times in the domain corpus, and all other synonyms do not appear at all (not a single occurrence of “air sac”, “air cell” and “tooth socket”).

This problem diminishes, if WordNet relations are taken into account. By using WordNet relations, we increase the number of synonyms that we search in the corpus and thus increase the number of actually appearing synonyms.

The relation that leads to the lowest recall is the hypernymy relation (47.1%). In general, one can speculate that this is due to the fact that a hypernym of a term does not necessarily lay in the same domain – and therefore receives a lower relevance ranking. Nevertheless, it may be a very general term that occurs very often, such that the low relevance score is compensated or even overruled by the high frequency.

The term “plasma”, for instance, has three synsets, from which the first one is the most appropriate in our domain. Based on the synonyms only, our program returns all three synsets. But if we add the hypernymy relation, the third synset gets selected by our program. This mistake is due to the fact that this synset has a synset of “state” {state, state of matter} as one of its synonyms, which has not a high domain relevance but occurs extremely often. The first synset of “plasma” has “extracellular fluid” as hypernym, which does not occur at all.

The relations hyponymy, holonymy and meronymy clearly stay in the same domain. A term like “lip” is partially disambiguated by looking at its holonyms: “vessel” or “mouth”. Since “mouth” lays in the domain of human anatomy, its relevance score is higher than vessel.

It is no surprise either that the topic relation, that assigns a category to synsets, is among the relations leading to high recall values. However, as many synsets do have a related topic, it does not contribute to precision.

There is a clear benefit of using several relations together. This combination increases the number of included synonyms further than by using a single relation.

4 Conclusions and Future Work

We presented a domain-specific corpus-based approach to the lexical enrichment of ontologies, i.e. enriching a given ontology with lexical entries derived from a semantic lexicon such as WordNet. Our approach was empirically tested by an experiment on combining the FMA with WordNet synsets that were disambiguated by use of a corpus of Wikipedia pages on human anatomy. The approach was evaluated on a benchmark of 50 ambiguous FMA terms with manually assigned
Lexical Enrichment of a Human Anatomy Ontology using WordNet synsets. Results show that the approach performs better than a most-frequent sense baseline. Further refinements of the algorithm that include the use of WordNet relations such as hyponym, hypernym, meronym, etc. showed a much improved performance, which was again improved upon drastically by combining the best of these relations. In summary, we achieved good performance on the defined task with relatively cheap methods. This will allow us to use our approach in large-scale automatic enrichment of ontologies with WordNet derived lexical information, i.e. in the context of the OntoSelect ontology library and search engine\(^7\) ([22]). In this context, lexically enriched ontologies will be represented by use of the LingInfo model for ontology-based lexicon representation ([15]).

5 Acknowledgements

We would like to thank Hans Hjelm of Stockholm University (Computational Linguistics Dept.) for making available the FMA term set and the Wikipedia anatomy corpus.

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors.

References


\(^7\) [http://olp.dfki.de/ontoselect/](http://olp.dfki.de/ontoselect/)
Abstract. We report on the current status of the Arabic WordNet project and in particular on the contents of the database, the lexicographer and user interfaces, the Arabic WordNet browser, linking to the SUMO ontology, the Arabic word spotter, and techniques for semi-automatically extending Arabic WordNet. The central focus of the presentation is on the semi-automatic extension of Arabic WordNet using lexical and morphological rules.

Keywords: Arabic NLP, Arabic WordNet, Ontology, Semi-automatic WordNet extension.

1 Introduction

Arabic WordNet (AWN – [1], [2], [3], inter alia) is currently under construction following a methodology developed for EuroWordNet [4]. The EuroWordNet approach maximizes compatibility across WordNets and focuses on the manual encoding of a set of base concepts, the most salient and important concepts as defined by various network-based and corpus-based criteria as reported in Rodríguez, et al [5]. Like EuroWordNet, there is a straightforward mapping from Arabic WordNet (AWN) onto Princeton WordNet 2.0 (PWN – [6]). In addition to constructing a WordNet for Arabic, the AWN project aims to extend a formal specification of the senses of its
synsets using the Suggested Upper Merged Ontology (SUMO), a language-independent ontology. This representation is essentially an interlingua between all WordNets ([7], [8]) and can serve as the basis for developing semantics-based computational tools for cross-linguistic NLP applications

The following discussion is divided into two main parts. We first present the current status of the Arabic WordNet and then we describe different techniques for semi-automatically extending AWN.

2 Current State of Arabic WordNet

2.1 Content of the Arabic WordNet Database

At the time of writing Arabic WordNet consists of 9228 synsets (6252 nominal, 2260 verbal, 606 adjectival, and 106 adverbial), containing 18,957 Arabic expressions. This number includes 1155 synsets that correspond to Named Entities which have been extracted automatically and are being checked by the lexicographers. Since these numbers are constantly changing, the interested reader can find the most up-to-date statistics at: http://www.lsi.upc.edu/~mbertran/arabic/awn/query/sug_statistics.php.

2.2 Interfaces

Two different web-based interfaces have been developed for the AWN project.

Lexicographer's Web Interface (Barcelona)
http://www.lsi.upc.edu/~mbertran/arabic/awn/update/synset_browse.php
The lexicographer’s interface has been designed to support the task of adding, modifying, moving or deleting WordNet synsets. Its functionalities include:
- listing the synsets assigned to each lexicographer (here, the lexicographer has many options to select from, including listing ‘completed synsets’ or ‘incomplete synsets’ or both),
- listing synsets by English word,
- listing synsets by synset offsets,
- listing synsets by date of creation,
- listing synsets without associated lexical items, or yet to be reviewed (to enhance validation, each lexicographer can review and comment on the others’ entries).

User's Web Interface (Barcelona)
http://www.lsi.upc.edu/~mbertran/arabic/awn/index.html

1 To our knowledge the only previous attempt to build a wordnet for the Arabic language consisted of a set of experiments by Mona Diab, [9] for attaching Arabic words to English synsets using only English WordNet and a parallel corpus Arabic English as knowledge sources.
This interface enables the user to consult AWN and search for Arabic words, Arabic roots, Arabic synsets, English words, synset offsets for English WordNet 2.0. Search can be refined by selecting the appropriate part of speech. A virtual keyboard is also available for users who do not have access to an Arabic keyboard.

### 2.3 WordNet to SUMO Mapping

SUMO ([7], [10]) and its domain ontologies form the largest publicly available formal ontology today. It is formally defined and not dependent on a particular application. SUMO contains 1000 terms, 4000 axioms, 750 rules and is the only formal ontology that has been mapped by hand to all of the PWN synsets as well as to EuroWordNet and BalkaNet. However, because WordNet is much larger than SUMO, many links are from general SUMO terms to more specific WordNet synsets. As of this writing, there are 3772 equivalence mappings, 100,477 subsuming mappings, and 10,930 mappings from a SUMO class to a WordNet instance. Most nouns map to SUMO classes, most verbs to subclasses of processes, most adjectives to subjective assessment attributes, and most adverbs to relations of and manners. While instance mappings are often from very specific SUMO classes, SUMO itself only includes a few sets of instances, such as the countries of the world. SUMO and its associated domain ontologies have a total of roughly 20,000 terms and 70,000 axioms.

SUMO synset definitions of the relevant synset can be viewed from the user’s web interface by using the SUMO Search Tool which relates PWN synsets to concepts in the SUMO ontology. To facilitate understanding of the ontology by Arabic speakers, the Sigma ontology management system [10] automatically generates Arabic paraphrases of its formal, logical axioms. SUMO has been extended with a number of concepts that correspond to words lexicalized in Arabic but not in English. They include concepts related to Arabic/Muslim cultural and religious practices and kinship relations. This is one way in which having a formal ontology provides an interlingua that is not limited by the lexicalization of any particular human language. For more information, see: http://sigmakee.cvs.sourceforge.net/*checkout*/sigmakee/KBs/ArabicCulture.kif

### 2.4 The AWN Browser

The Arabic WordNet Browser is a stand-alone application that can be run on any computer that has a Java virtual machine. In its current state, its main facilities include browsing AWN, searching for concepts in AWN, and updating AWN with latest data from the lexicographers.

Searching can be done using either English or Arabic. In Arabic, the search can be carried out using either Arabic script or Buckwalter transliteration [11] and can be for a word or root form, with the optional use of diacritics. For English, the browser supports a word-sense search alongside a graphical tree representation of PWN which allows a user to navigate via hyponym and hypernym relations between synsets. A combination of word-sense search and tree navigation enables a user to quickly and efficiently browse translations for English into Arabic.
Since users unfamiliar with Arabic cannot be expected to know how to convert an Arabic word they have copied from a Web page into an appropriate citation form, we have integrated Arabic morphological analysis into the search function, using a version of AraMorph [12]. A virtual Arabic keyboard is also accessible to enable Arabic script entry for the different search fields.

SUMO ontology navigation is currently being integrated into the browser, using a tree traversal procedure similar to that for PWN. Users will be able to search or browse AWN using SUMO as the interlingual index between English and Arabic. Also under construction are Arabic tree navigation and the automatic generation of Arabic glosses. These additions will be included in the next release version of the browser.

More detailed information and screen shots can be found at: http://www.globalwordnet.org/AWN/AWNBrowser.html

The browser is available for downloading from Sourceforge under the General Public License (GPL) at: http://sourceforge.net/projects/awnbrowser/

2.5 The Arabic Word Spotter

An Arabic Word Spotter has been developed to provide the user with a tool to test AWN’s coverage by identifying those words in an Arabic web page that can be found in AWN. The word spotter can be accessed at: http://www.lsi.upc.edu/~mbertran/arabic/wwwWn7/

Arabic words are searched for first in AWN and, failing that, in a few bilingual dictionaries. The procedure relies on the AraMorph stemmer and, once a match is found, a word level translation is provided. Translation of stop words is provided as well.

Help and HowTos are available from: http://www.lsi.upc.edu/~mbertran/arabic/wwwWn7/help/help.php?

3 Approaches to the Semi-automatic Extension of AWN

Although the construction of AWN has been manual, some efforts have been made to automate part of the process using available bilingual lexical resources. Using lexical resources for the semi-automatic building of WordNets for languages other than English is not new. In some cases a substantial part of the work has been performed automatically, using PWN as source ontology and bilingual resources for proposing correlates. An early effort along these lines was carried out during the development of Spanish WordNet within the framework of EuroWordNet project ([13], [5]). Later, the Catalan WordNet [14] and Basque WordNet [15] were developed following the same approach.

Within the BalkaNet project [16] and the Hungarian WordNet project [17], this same methodology was followed. In this case, the basic approach was complemented by methods that relied on monolingual dictionaries. As an experiment with the Romanian WordNet, [18] follow a similar approach, but use additional knowledge
sources including Magnini’s WordNet domains [19] and WordNet glosses. They use a set of metarules for combining the results of the individual heuristics and achieve 91% accuracy for the 9610 synsets covered. Finally, to build both a Chinese WordNet and a Chinese-English WordNet, [20] complement their bilingual resources with information extracted from a monolingual Chinese dictionary.

For AWN, we have investigated two different possible approaches. On the hand, we produce lists of suggested Arabic translations for the different words contained in the English synsets corresponding to the set of Base Concepts. In this case the input to the lexicographical task is the English synset, its set of synonyms and their Arabic translations. On the other hand, we derive new Arabic word forms from already existing, manually built, Arabic verbal synsets using inflectional and derivational rules and produce a list of suggested English synset associations for each form. In this case the input is the Arabic verb, the set of possible derivates and the set of English synsets which would be linked to corresponding Arabic synset. In both cases, the list of suggestions is manually validated by lexicographers.

3.1 Suggested Translations

For this approach, we start with a list of <English word, Arabic word, POS> tuples extracted from several publicly available English/Arabic resources. The first step was to clean and standardize the entries. The available resources differ in many details. Some contain POS for each entry while others do not. Arabic words were in some cases vocalized and in others not. In some cases certain diacritics are used, such as shadda (i.e., consonant reduplication), while in others no diacritics at all appear. Some dictionaries contain the perfect tense form for verbs while others use the imperfect form. After this standardization process, we merged all the sources (using both directions of translation) into one single bilingual lexicon and then took the intersection of this lexicon with the set of Base Concept word forms. This latter set was built merging the Base Concepts of EuroWordNet, 1024 synsets, with those of Balkanet, 8516 synsets.

Following 8 heuristic procedures used in building the Spanish WordNet [21] as part of EuroWordNet [4], the associations between Arabic words and PWN synsets in the Arabic-English bilingual lexicon were scored. The methodology assigned a score to each association, but since the Arabic WordNet has been manually constructed, no threshold was set and all associations were provided to the lexicographer for verification. Thus, when editing an Arabic synset, the lexicographer begins with a suggested association, rather than an empty synset with only the English data to go by. Some suggestions were correct or very similar to correct ones. Others were incorrect but served to trigger an Arabic word that might otherwise have been missed. The result has been a much richer set of Arabic synsets.

Initially 15,115 translations were suggested, of which only 9748 (64.5%) have been thus far checked by the lexicographers. The results show that of these, 392 candidates (4.0%) were accepted without any changes, 1246 (12.8%) were accepted with minor changes (such as adding diacritics), 877 (9.0%), while good candidates, were rejected because they were identical or very similar to translations that had already been chosen by the lexicographer, and 7233 (74.2%) were rejected because
they were incorrect given the gloss and examples. We will revise these results once all the Base concepts have been completed at the end of the project.

At first glance, these results are not especially impressive and, as a result, we turned to an alternative approach. At the same time, it is difficult to compare these figures with results obtained for other languages because we are interested exclusively in generating suggestions for Base Concepts which are to be confirmed by lexicographers while other approaches do not have this objective. Since the words belonging to Base Concept synsets are often highly polysemous, the accuracy of predicting translations is generally lower. In addition, since we are more interested in high coverage, no filters were applied with a corresponding drop in precision.

### 3.2 Semi-automatic Extension of AWN Using Lexical and Morphological Rules

In this section we explore an alternative methodology for the semi-automatic extension of Arabic WordNet using lexical rules as applied to existing AWN entries. This methodology takes advantage of one of a central characteristic of Arabic, namely that many words having a common root (i.e. a sequence of typically three consonants) have related meanings and can be derived from a base verbal form by means of a reduced set of lexical rules. Since AWN entries must be manually reviewed, our aim is once again not to automatically attach new synsets but rather to suggest new attachments and to evaluate whether these suggestions can help the lexicographer. As with previous approach, we are more interested in getting a broad coverage than high accuracy, although an appropriate balance between these two measures is nonetheless desirable.

#### 3.2.1 Setting

In the studies reported in this section, we deal only with a very limited but highly productive set of lexical rules which produce regular verbal derivative forms, regular nominal and adjectival derivative forms and, of course, inflected verbal forms.

From most of the basic Arabic triliteral verbal entries, up to 9 additional verbal forms can be regularly derived as shown in Table 1. We refer to the set of lexical rules that account for these forms as Rule Set 1. They have been implemented as regular expression patterns.

For instance, the basic form درس (DaRaSa, to study/to learn) has as its root درس (DRS). The first form pattern in Table 1 applied to this root produces the original basic forms (in this case simply adding diacritics). If we apply the second form pattern in Table 2 to the same root, the form درس (DaRRaSa, to teach) is obtained.
Table 1: Patterns of Arabic regular derived forms

<table>
<thead>
<tr>
<th>Class</th>
<th>Arabic Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Basic)</td>
<td>ﻓﻌﻞ</td>
</tr>
<tr>
<td>2</td>
<td>ﻓﻌّﻞ</td>
</tr>
<tr>
<td>3</td>
<td>ﻓﺎﻋﻞ</td>
</tr>
<tr>
<td>4</td>
<td>ﺗﻔﻌﻞ</td>
</tr>
<tr>
<td>5</td>
<td>ﺗﻔﺎﻋﻞ</td>
</tr>
<tr>
<td>6</td>
<td>ﺗﻔﺎﻋﻞ</td>
</tr>
<tr>
<td>7</td>
<td>ﺑﺎﻔﻊ</td>
</tr>
<tr>
<td>8</td>
<td>ﺑﺎﻔﻊ</td>
</tr>
<tr>
<td>9</td>
<td>ﺑﺎﻔﻊ</td>
</tr>
<tr>
<td>10</td>
<td>ﺑﺎﻔﻊ</td>
</tr>
</tbody>
</table>

From any verbal form (whether basic or derived by Rule Set 1), both nominal and adjectival forms can also be generated in a highly systematic way: the nominal verb (masdar) as well as masculine and feminine active and passive participles. We refer to this set of rules as Rule Set 2. Examples include the masdar درس (DaRSun, lesson, study) from درس (DaRaSa, to study/to learn) and مدرّس (MuDaRRiSun, male teacher) from درس (DaRRaSa, to teach).

Finally, a set of morphological rules for each basic or derived verb form is applied in order to produce the full set of inflected verb forms as exemplified in Table 2.

Table 2: Some inflected verbal forms (of 82 possible) for درس (DaRaSa, to learn)

<table>
<thead>
<tr>
<th>English form</th>
<th>Arabic form</th>
</tr>
</thead>
<tbody>
<tr>
<td>(he) learned</td>
<td>ﻳﺪرﺳﺖ</td>
</tr>
<tr>
<td>(I) learned</td>
<td>ﺑﺪرﺳﺖ</td>
</tr>
<tr>
<td>(I) learn</td>
<td>ﺑﺪاﺪرﺳ</td>
</tr>
<tr>
<td>(he) learns</td>
<td>ﻳﺪاﺪرﺳ</td>
</tr>
<tr>
<td>(we) learn</td>
<td>ﻳﺪاﺪرﺳ</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

As reported below, these forms are especially useful for searching a corpus as well as in various applications. The number of different forms depends on the class of the verb but it ranges from 44 to 84 forms. Class 1, for instance, has 82 forms and, thus, requires the application of 82 different morphological rules. We refer to this set of rules as Rule Set 3.

Beyond this, we aim to extend this basic approach to the derivation of additional forms including the feminine form from any nominal masculine form (for instance, مدرسة MuDaRRiSatun, female teacher, from مدرّس MuDaRRiSun, male teacher), or the regular plural forms from any nominal singular form. For instance, the regular nominative plural form is created by adding the suffix (Una) to the singular form (e.g., مدرّسون MuDaRRiSUna, male teachers, is derived from مدرّس MuDaRRiSun, male teacher).
3.2.2 Central Problems to Address
Implementing the ideas stated in the previous section is not straightforward. Several problems have to be addressed but perhaps the two most important are 1) filtering noise caused by over the generation of derivative verb forms and 2) mapping the newly created Arabic word forms to appropriate WordNet synsets, i.e., mapping words to their appropriate sense. Obviously not all the derivative forms generated by Rule Sets 1 and 2 are valid for any given basic verbal form in Arabic. For instance, for درس (DaRaSa, to learn) of the nine possible derivates generated by the application of Rule Set 1, shown in Table 1, only the six shown in Table 3 are valid according to [22]. Thus, some kind of filtering has to take place in order to reduce the noise wherever possible. That is to say, only the most promising candidates should be proposed to the lexicographer. In addition, once the set of candidate derivates has been built and the corresponding nominal and adjectival forms generated, we have to map all these forms to English translations and from these to the appropriate PWN synsets.

Table 3: Valid derivates from درس (DaRaSa, to learn)

<table>
<thead>
<tr>
<th>Class</th>
<th>English form</th>
<th>Arabic form</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (basic)</td>
<td>to learn, to study</td>
<td>درس</td>
</tr>
<tr>
<td>2</td>
<td>to teach</td>
<td>درّس</td>
</tr>
<tr>
<td>3</td>
<td>to study (together with someone)</td>
<td>دارس</td>
</tr>
<tr>
<td>4</td>
<td>to learn with</td>
<td>ادرس</td>
</tr>
<tr>
<td>6</td>
<td>to study (carefully together)</td>
<td>ﻣﺪارس</td>
</tr>
<tr>
<td>7</td>
<td>to vanish</td>
<td>ﻣﺪردد</td>
</tr>
</tbody>
</table>

3.2.3 Resources
The procedures described below make use of the following resources:
- Princeton’s English WordNet 2.0,
- Arabic WordNet (specifically the set of Arabic verbal synsets currently available),
- the LOGOS database of Arabic verbs which contains 944 fully conjugated Arabic verb (available at: http://www.logosconjugator.org/verbi_utf8/all_verbs_index_ar.html),
- the Arabic GigaWord Corpus (available through LDC: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T02).

3.2.4 Overview of the Approach
Broadly speaking, the procedure we follow in generating a set of likely <Arabic word, English synset> pairs is to:
1. produce an initial list of candidate word forms (as described in Section 3.2.6),
2. filter the less likely candidates from this list (as described in Section 3.2.7),
3. generate an initial list of candidate synsets attachments (as described in Section 3.2.8),
4. score the reliability of these candidates (as described in Section 3.2.9),
5. manually review the candidates and include the valid associations in AWN.

3.2.4.1 Building the initial set of word candidates
To build the initial set of candidate word forms, we first collect a set of basic (Class 1) verb forms, such as درس (DaRaSa, to learn), from the existing 2296 verbs in AWN and transliterate them using Buckwalter encoding [11]. We next apply Rule Set 1 to generate the 9 basic derivative verb forms (both valid or not). Then, for each of these new verb forms, we apply Rule Set 3 in order to derive the full set of possible inflected forms.

3.2.4.2 Learning filters on translations
In order to determine whether or not a particular possible word form is likely to turn out to be a valid word form, we build a decision tree classifier using machine learning for each of the 9 classes of derivation (i.e. Classes 2 through 10). The choice of decision trees is mainly motivated because their ease of interpretation, since otherwise they provided similar results to those of Adaboost, an alternative approach which we have tested. We used the C5.0 implementation within Weka toolbox [23]. The software can be obtained from: http://www.cs.waikato.ac.nz/~ml/weka/index.html.

The features used for learning included the following:
1. the relative frequency of each inflected form for a given class of derivatives in the GigaWord Corpus,
2. whether the base form appears in the NMSU dictionary or not,
3. the POS tag of the base form in NMSU dictionary,
4. the attribute TRUE (positive example) or FALSE (negative example).

In order to learn a decision tree, the algorithm must be presented with both positive and negative examples. For positive examples, we used the LOGOS database (946 examples), AWN (2296 examples) and the NMSU dictionary (15,654 examples). LOGOS and AWN are the most accurate but do not provide enough material. NMSU has broad coverage but is less accurate because the entries are not vocalized and lack diacritics (for some classes the lack of the shadda diacritic is a serious problem).

To build the training set, we matched each inflected form for each of the base forms (basic or derived) against the GigaWord Corpus and the NMSU dictionary in order to extract the relevant features for learning. Finally, we selected all the base forms corresponding to the word forms that occurred in the resources as positive examples, and used the remaining forms (i.e., those that do not occur in either the GigaWord Corpus or in the NMSU dictionary) as negative examples. All other forms are discarded. Table 4, for instance, shows the size of the training set used for learning the filter for Class 7.

---

2 In Arabic shadda is how consonant reduplication or germination is marked. Obviously if this diacritic is lost the correct orthographic form of a word is affected.
Following this general procedure, a decision tree classifier was learned for each class of derivation (in fact, only 8 filters were learned because there were too few examples for Class 9). We applied 10-fold cross-validation. The results for all the classifiers but one were over 99% of F1 value although in some cases the resultant decision tree consisted of only a single query on the occurrence of the base form in the NMSU dictionary (i.e. the form was accepted simply if it occurs in NMSU dictionary).

3.2.4.3 Building the list of candidate synsets attachments
To build a list of candidate synset attachments, we first generate a list of possible base verb forms by applying the filters described above. We then apply Rule Set 2 to each of the base verb forms to generate the set of related Arabic noun and adjective forms. Only those forms occurring in NMSU dictionary with English equivalents occurring in PWN are retained. For each of these word forms, all the English translations from NMSU dictionary and all their PWN synsets are collected as candidates. The result of this process is a candidate set of tuples of the form <Arabic word, English word, English synset>. The final step is to assign a reliability score to each <Arabic word, English synset> tuple.

3.2.4.4 Scoring the candidate synset attachments
Our scoring routine is based on the observation that in most cases the set of derivative forms have semantically related senses. For instance, درس (DaRaSa, to study) belongs to Class 1 and its masdar is درس (DaRSun, lesson). درّس (DaRRaSa, to teach) belongs to Class 2 and its masculine active participle is مدرّس (MuDaRRiSun, male teacher). Clearly these four words are semantically related. Therefore, if we map Arabic words to English translations and then to the corresponding PWN synsets, we can expect that the correct assignments will correspond to most semantically related synsets. In other words, the most likely <Arabic word, English synset> associations are those corresponding to the most semantically related items.

There are three levels of connections to be considered:
- relations between an Arabic word and its English translations,
- relations between an English word and its PWN synsets,
- relations between a PWN synset and other synsets in PWN.

---

The relations A_{base} -> A_{i} have not been considered explicitly because A_{base} comes from an existing AWN synset and thus its association has already been established manually.

### Table 4: Size of training set for learning the Class 7 filter

<table>
<thead>
<tr>
<th></th>
<th>Logos</th>
<th>AWN</th>
<th>NMSU</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>8</td>
<td>24</td>
<td>1718</td>
<td>1750</td>
</tr>
<tr>
<td>negative</td>
<td>70</td>
<td>0</td>
<td>4856</td>
<td>4926</td>
</tr>
<tr>
<td>Total</td>
<td>78</td>
<td>24</td>
<td>6574</td>
<td>6676</td>
</tr>
</tbody>
</table>
To identify the “most semantically related” associations between Arabic words and PWN synsets, we:

1. collect the set of <Arabic word, English word, English synset> tuples for a given Arabic base verb form and its derivatives,
2. extract the set of English synsets and identify all the existing semantic relations between these synsets in PWN4,
3. build a graph with three levels of nodes corresponding to Arabic words, English words, and English synsets respectively and edges corresponding to the translation relation between Arabic words and English words, the membership relation between English words and PWN synsets and finally, the recovered relations between PWN synsets.

These are represented in the graph in Figure 1.

Fig. 1. Example of Graph of dependencies

Two approaches to scoring are being examined. The first, described below, is based on a set of heuristics that use the graph structure directly while the second, more complex, maps the graph onto a Bayesian Network and applies a learning algorithm. The latter approach is the subject of ongoing research and will be described in a separate forthcoming paper.

Using the graph as input, the first approach to calculating the reliability of association between Arabic word and PWN synset consists of simply applying a set of five graph traversal heuristics. The heuristics are as follows (note that in what follows, “A_base”, “A1”, “A2”, etc., correspond to Arabic word forms, A_base being the initial verbal base form, “E”, “E1”, “E2”, etc. to English word forms, and “S”, “S1”, “S2”, etc. to PWN synsets):

1. If a unique path A-E-S exists (i.e., A is only translated as E), and E is monosemous (i.e., it is associated with a single synset), then the output tuple <A, S> is tagged as 1. See Figure 2.

---

4 As in the rest of experiments reported in this paper we have used the relations present in PWN2.0
2. If multiple paths $A-E_1-S$ and $A-E_2-S$ exist (i.e., $A$ is translated as $E_1$ or $E_2$ and both $E_1$ and $E_2$ are associated with $S$ among other possible associations) then the output tuple $<A, S>$ is tagged as 2. See Figure 3.

3. If $S$ in $A-E-S$ has a semantic relation to one or more synsets, $S_1, S_2, ...$ that have already been associated with an Arabic word on the basis of either heuristic 1 or heuristic 2, then the output tuple $<A, S>$ is tagged as 3. See Figure 4.

4. If $S$ in $A-E-S$ has some semantic relation with $S_1, S_2, ...$ where $S_1, S_2, ...$ belong to the set of synsets that have already been associated with related Arabic words, then the output tuple $<A-S>$ is tagged as 4. In this case there is only one translation $E$ of $A$ but more than one synset associated with $E$. This heuristic can be sub-classified by the number of input edges or supporting semantic relations (1, 2, 3, ...). See Figure 5.
5. Heuristic 5 is the same as heuristic 4 except that there are multiple translations $E_1, E_2, \ldots$ of $A$ and, for each translation $E_i$, there are possibly multiple associated synsets $S_{i1}, S_{i2}, \ldots$. In this case the output tuple <A-S> is tagged as 5 and again the heuristic can be sub-classified by the number of input edges or supporting semantic relations (1, 2, 3 ...). See Figure 6.

### 3.2.5 A Detailed Example
Consider once more the case of verb درس (DaRaSa, to learn). From the 9 forms obtained by applying Rule Set 1 to the basic form, the filter accepts the Classes 2, 4 and 7 (as shown in Table 3 on p.7 above). Here we look at the basic form and the Class 2 derivate. We begin by collecting the following tuples using the NMSU dictionary and PWN:
Between the synsets identified above, the following relations hold:

07776854 has as a member 07776854
07776854 is a member of 07776854
00801981 cause 00578275
00686199 is a part of 00831015
00831015 has as a part 00686199
00578275 is a type of 00587299
00587299 has as a type 00578275
00587299 is a type of 00584743
00584743 has as a type 00587299
00834401 is a type of 00836504
00836504 has as a type 00834401

Using these relations, we build an undirected graph where nodes correspond to synsets and edges to semantic relations between synsets. Table 5 shows the 12 candidate associations generated of which 9 are deemed correct by the lexicographers. Note that no candidates have been selected on the basis of the heuristic 1 or heuristic 4. Note also that subclasses of heuristic 5 (rows 9 to 12) are somewhat overvalued because nodes connected by relations with inverses are counted twice.
Table 5: Candidates for Class 1 and 2 derivates of درس (DaRaSa, to learn)

<table>
<thead>
<tr>
<th>Buckwalter</th>
<th>POS</th>
<th>Synset Off</th>
<th>Class</th>
<th>Arabic form</th>
<th>Lex Judge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>drs</td>
<td>580363</td>
<td>2</td>
<td>درس</td>
<td>ok</td>
</tr>
<tr>
<td>2</td>
<td>drs</td>
<td>801981</td>
<td>2</td>
<td>درس</td>
<td>ok</td>
</tr>
<tr>
<td>3</td>
<td>tdry</td>
<td>834401</td>
<td>2</td>
<td>تدريس</td>
<td>ok</td>
</tr>
<tr>
<td>4</td>
<td>tdry</td>
<td>831015</td>
<td>2</td>
<td>تدريس</td>
<td>ok</td>
</tr>
<tr>
<td>5</td>
<td>mdr</td>
<td>9997151</td>
<td>2</td>
<td>مدرس</td>
<td>ok</td>
</tr>
<tr>
<td>6</td>
<td>drs</td>
<td>836504</td>
<td>3</td>
<td>درس</td>
<td>ok</td>
</tr>
<tr>
<td>7</td>
<td>drs</td>
<td>686199</td>
<td>3</td>
<td>درس</td>
<td>ok</td>
</tr>
<tr>
<td>8</td>
<td>drs</td>
<td>578275</td>
<td>3</td>
<td>درس</td>
<td>ok</td>
</tr>
<tr>
<td>9</td>
<td>drs</td>
<td>587299</td>
<td>5,5</td>
<td>درس</td>
<td>ok</td>
</tr>
<tr>
<td>10</td>
<td>drs</td>
<td>584743</td>
<td>5,3</td>
<td>درس</td>
<td>no</td>
</tr>
<tr>
<td>11</td>
<td>mdr</td>
<td>7776854</td>
<td>5,3</td>
<td>مدرس</td>
<td>no</td>
</tr>
<tr>
<td>12</td>
<td>tdry</td>
<td>7787222</td>
<td>5,3</td>
<td>تدريس</td>
<td>no</td>
</tr>
</tbody>
</table>

The first row in Table 5 corresponds to the tuple <درس, 580363>. It has been selected on the basis of heuristic 2 because the synset 580363 occurs in:

درس: to learn: [.... '00580363', ...]
درس: to study: [.... '00580363', ...].

The sixth row of Table 5 corresponds to the tuple <درس, 836504>. In this case, heuristic 3 can be applied because in

درس: lesson: [.... '00836504' ....]

the synset 00836504 is related to the synset 00834401 by a hyponymy relation:

00836504 has as a type 00834401

which, in turn, has been suggested on the basis of heuristic 2 (see row 3 in Table 5). Finally, consider the tuple <درس, 587299> in row 9 of Table 5. This is an example of the application of heuristic 5. In

درس: to study: [.... '00587299', ...]

the synset 00587299 receives support from (among others):

00578275 is a type of 00587299
00584743 has as a type 00587299

where 00578275 and 00584743 have been associated with other derivative forms of درس (DaRaSa, to learn) as shown in rows 8 and 10 respectively of Table 5.

3.2.6 Evaluation
To perform an initial evaluation of this approach, we randomly selected 10 of the 2296 verbs currently in AWN that have a non null coverage and which satisfy all the requirements above. In addition, for the purpose of illustration, we added the verb درس (DaRaSa, to learn) as a known example. The process for building the candidate set of Arabic form-synset associations described in Section 3.2.4 was applied to each of the 11 basic verb forms resulting in 11 sets of candidate tuples. The size in words and synsets are presented in Table 6.
Table 6: Size of the candidate sets for testing

<table>
<thead>
<tr>
<th>Arabic form</th>
<th># of words</th>
<th># of synsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>ﻏَ져َلَ</td>
<td>107</td>
<td>190</td>
</tr>
<tr>
<td>ﺔَﻐَبَ</td>
<td>71</td>
<td>77</td>
</tr>
<tr>
<td>صَقَل</td>
<td>31</td>
<td>21</td>
</tr>
<tr>
<td>ﺮَﻴَب</td>
<td>62</td>
<td>102</td>
</tr>
<tr>
<td>أَذَر</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>أَذَر</td>
<td>80</td>
<td>105</td>
</tr>
<tr>
<td>ﺔَ 굉장히</td>
<td>40</td>
<td>22</td>
</tr>
<tr>
<td>ﺔَذَر</td>
<td>56</td>
<td>49</td>
</tr>
<tr>
<td>أَذَر</td>
<td>38</td>
<td>34</td>
</tr>
<tr>
<td>أَذَر</td>
<td>85</td>
<td>140</td>
</tr>
<tr>
<td>درَس</td>
<td>57</td>
<td>51</td>
</tr>
</tbody>
</table>

Each of the tuples was then scored following the procedure described in Section 3.2.4.4. We did not introduce a threshold and so the whole list of candidates, ordered by reliability score, was evaluated by a lexicographer. The results are presented in Table 7. Here, the first column indicates the scoring heuristic applied, the second the number of instances to which it applied, the third the number of instances judged acceptable by the lexicographer, the fourth the number of instances judged unacceptable, and the fifth the percentage correct.

These results are very encouraging especially when compared with the results of applying the EuroWordNet heuristics reported in Section 3.1. While the sample is clearly insufficient (for instance, there are no instances of the application of heuristic 1 and too few examples of heuristic 3), with few exceptions the expected trend for the reliability scores are as expected (heuristics 2 and 3 perform better than heuristic 4 and the latter better than heuristic 5). It is also worth noting that heuristic 3, the first that relies on semantic relations between synsets in PWN, outperforms heuristic 2. However, we have not attempted to establish statistical significance because of the small size of the test set. Otherwise, an initial manual analysis of the errors shows that several are due to the lack of diacritics in the resources.

Currently we are extending the coverage of the test set. We will then repeat the entire procedure using only dictionaries containing diacritics. We are also planning to refine the scoring procedure by assigning different weights to the different semantic relations between synsets. In addition, we expect to compare this approach with that based on Bayesian Networks mentioned earlier.
Table 7: Results of the evaluation of proposed Arabic word-PWN synset associations

<table>
<thead>
<tr>
<th>Heuristic</th>
<th># correct</th>
<th># ok</th>
<th># no</th>
<th>% correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>27</td>
<td>15</td>
<td>64</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>13</td>
<td>6</td>
<td>68</td>
</tr>
<tr>
<td>4,1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4,2</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>57</td>
</tr>
<tr>
<td>4,3</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>56</td>
</tr>
<tr>
<td>4,4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>4,5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>4,6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4,7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5,1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5,2</td>
<td>63</td>
<td>32</td>
<td>31</td>
<td>51</td>
</tr>
<tr>
<td>5,3</td>
<td>109</td>
<td>41</td>
<td>68</td>
<td>38</td>
</tr>
<tr>
<td>5,4</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5,5</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>60</td>
</tr>
<tr>
<td>5,6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>5,7</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5,13</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>272</td>
<td>135</td>
<td>137</td>
<td>50</td>
</tr>
</tbody>
</table>

4 Outlook and Conclusion

We have presented the current state of Arabic WordNet and described some procedures for semi-automatically extending AWN’s coverage. On the one hand, the procedure for suggesting translations on the basis of 8 heuristics used for EuroWordNet was presented and discussed. On the other, we described a set of procedures for the semi-automatic extension of AWN using lexical and morphological rules and provided the results of their initial evaluation.

We hope that work will continue on augmenting the AWN database by both manual and automatic means even after the current project ends. We welcome ideas, suggestions, and expressions of interest in contributing or collaborating on both further extension of the lexical database as well as on development of related software. Finally, we are looking forward to a wide range of NLP applications that make use of this valuable resource.
Acknowledgement

This work was supported by the United States Central Intelligence Agency.

References

Building a WordNet for Persian Verbs

Masoud Rouhizadeh, Mehrounoush Shamsfard, and Mahsa A. Yarmohammadi

Natural Language Processing Laboratory, Shahid Beheshti University, Tehran, Iran
m.rouhizadeh@mail.sbu.ac.ir, m-shams@sub.ac.ir, m_yarmohammadi@std.sbu.ac.ir

Abstract. This article is a report of an ongoing project to develop a WordNet for Persian Verbs. To build this WordNet we apply the expand approach used in EuroWordNet and BalkaNet. We are now building the core WordNet of Persian verbs by translating the verbs of BalkaNet Concept Sets 1, 2 and 3. The translation process includes automatic suggested equivalences of English synsets in our WordNet editor, and their manually refinement by a linguist using different dictionaries and corpora. We are also adding the frequent Persian verbs that are not included in the sets using an electronic Persian corpus. This core WordNet will be extended (semi)automatically. The most important fact about Persian verbs is that most of them are compound rather than simple. Compound verbs in Persian are formed in two major patterns: combination and incorporation. In many cases the compound verbs are semantically transparent, that is the meaning of compound verb is the function of the meaning of its verbal and non-verbal constituents. This suggests that many verbs in Persian WordNet can be directly connected to their non-verbal constituent in Persian WordNet and so inherit the existing relations among those words too.

1 Introduction

Persian is the official language of three countries and it is also spoken in more than six countries. There is no doubt in the necessity of basic NLP resources and tools such as a standard lexicon for this wide-spoken language. The WordNet of Persian verbs is an ongoing project to provide a part of the Persian WordNet, a powerful tool for Persian NLP applications.

Persian verbs WordNet goes closely in the lines and principles of Princeton WordNet, EuroWordNet and BalkaNet to maximize its compatibility to these WordNets and to be connected to the other WordNets in the world for crosslinguistic abilities such as MT and multilingual dictionaries and thesauri. It also aims to be merged to the other existing WordNets of Persian nouns [1] and Persian adjectives [2].

In this article, first we give an overview of the methodology we take, lexical resources we are using and the building process. Finally, we point to the most important characteristic of Persian verbs, that is they are mostly compound verbs and consist of a verbal and a non-verbal constituent. This feature leads to a highly cross part of speech connected WordNet.
2 Methodology

We are constructing the Persian Verbs WordNet according to the methods applied for EuroWordNet [3], [4] that is a widely used approach in many WordNets. This approach maximizes compatibility across WordNets and at the same time preserves the language specific structures of Persian. We follow the expand strategy in which a core WordNet should be developed manually and then extended (semi)automatically [5]. To develop the core WordNet of Persian verbs we are manually translating the verbs of BalkaNet Concept Sets 1, 2 and 3 (BCS1, BCS2 and BCS3) [6]. We are also adding the frequent Persian verbs that are not included in the sets using the electronic Persian corpus [7]. Adding hypernyms and the first level hyponyms to these verb Base Concepts will result to the core WordNet of Persian verbs. This core WordNet have to be extended (semi)automatically using specifically available resources, e.g. monolingual and bilingual dictionaries, lexicons, ontologies, thesauri, etc.

3 Building process and lexical resources

In this project we are making use of a machine readable dictionary to suggest the Persian equivalences of PWN synsets in our WordNet editor, VisDic. In the next step, the suggestions are refined, using our linguistic knowledge of English and Persian and English-Persian Millennium Dictionary [8], the most reliable English-to-Persian dictionary. Then we refer to Anvari [9] a Persian monolingual dictionary, to check out the consistency and correctness of our equivalences. We also use the Persian Linguistic Database (PLDB), [7]. This is an on-line database for the contemporary (Modern) Persian. The database contains more than 16.000.000 words of all varieties of the Modern Persian language in the form of running texts. Some of the texts are annotated with grammatical, pronunciation and lemmatization tags. A special and powerful software provides different types of search and statistical listing facilities through the whole database or any selective corpus made up of a group of texts. The database is constantly improved and expanded. It provides us a means of handling various types of texts to determine the frequency of verbs and helps us find and add the verbs that are not included in BalkaNet Concept Sets 1, 2 and 3. (At this time we have translated verbs from BCS1 and BCS2).

3.1 The editor

The editor we use to build our WordNet is BalkaNet multilingual viewer and editor, VisDic [6]. It is a graphical application for viewing and editing WordNet lexical databases stored in XML format. Most of the program behavior and the dictionary design can be configured.

Figure 1 shows the View tab of VisDic editor for the verb "درس دادن" "to teach". The POS, ID, synonyms, hypernyms and other WordNet relations of the selected word are shown in this tab.
Figure 1 shows the View tab of VisDic editor for the verb "درس دادن" ‘to teach’. This tab allows editing the actual entry. There are some other buttons in this tab: "New" button for creating a new entry with unique key, "Add" and "Update" buttons to add actual entry to the dictionary or update the actual entry.

The output file generated by VisDic is a human readable XML file. In this file, each synset is defined within a <SYNSET></SYNSET> including some other inner tags.

4 Compound verbs in Persian

Persian verbs can be divided into two major morphological categories: simple and compound verbs. As the names suggest, simple verbs have simple morphological structure, the verbal constituent. Compound verbs, on the other hand, consist of a non-verbal constituent, such as a noun, adjective, past participle, prepositional phrase, or adverb, and a verbal constituent.

As reported in Sadeghi [10] the maximum number of simple verbs in Persian today, is only 115 verbs. This and many other investigations such as existence of a great number of compound verbs formed based on various Arabic parts of speech, and using of all new borrowing verbs from western languages as compound verbs in Persian, reveal that compound-verb formation is highly productive in Persian today [11]. The number of registered compound verbs is around 2500-3000. Thus, most of the Persian verbs, containing the basic ones, are compound verbs.

We follow Dabir-Moghaddam’s account [11] as he suggests two major types of compound-verb formation in Persian: Combination and Incorporation.
**Fig. 2.** The Edit tab of VisDic editor for the verb "درس دادن" 'to teach'
4.1 Combination

In this type of compound-verb formation the non-verbal and the verbal constituent are combined in the following patterns:

4.1.1 Adjective + Auxiliary
delxor-shodan ‘to become annoyed’ ‘annoyed-become’

4.1.2 Noun + Verb
   bāzi-kardan ‘to play’ ‘play-do’
   pas-dādan ‘to return’ ‘back-give’
   dast-da tan ‘to be involved’ ‘hand-have’

4.1.3 Prepositional Phrase + Verb
   be donya ōmadan ‘to be born’ ‘to-world-come’

4.1.4 Adverb + Verb
   dar yāftan ‘to perceive’ ‘in-find’

4.1.5 Past Participle + Passive Auxiliary
   sāxte odan ‘to be built’ ‘built-ecome’

4.2 Incorporation:

In Persian, the **direct objects** (losing its grammatical endings) and can incorporate with the verb, to create an intransitive compound verb, which is a conceptual whole as shown in the following example:

4.2.1

a. mâ qazā-y-e-m-ân- râ xor-d-im
   we food-our-pl.-DO eat-past-we
   ‘We ate our food’

b. mâ qazā- xor-d-im
   ‘We did food eating’

Also, some prepositional phrases can incorporate with verbs. Here, the proposition disappears after incorporation:

4.2.2

a. ân-hâ be zamin xor-d-and
   that-pl. to ground eat-past-they
   ‘They fell to the ground.’

b. ân-hâ zamin xor-d-and
   ‘They fell down.’

As can be seen, the morphological structure of verbs in Persian is highly connected to the other parts of speech, especially nouns; and in many cases the compound verbs are semantically transparent, that is the meaning of the resulting compound verb is
the function of the meaning of its verbal and non-verbal constituents. This suggests
that many verbs in Persian WordNet can be directly connected to their non-verbal
constituent in Persian WordNet, i.e. the nouns, the adjectives and the adverbs; and so
inherit the existing relations among those words too; as in the verb qazâ- xordan ‘to
eat’ which is connected directly to the noun qazâ ‘food’ and to its hyponyms, for
instance, nâhâr ‘lunch’ in the verb nâhâr-xordan ‘to eat lunch’. So Persian WordNet
will be a strongly cross part of speech connected WordNet.

5 Conclusion

In this article we had a review of the on-going project on building a WordNet of
Persian verbs. Taking into consideration the fact that most of the verbs in Persian are
compounds and highly connected to the other parts of speech, the WordNets of
Persian verbs, noun, adjectives and adverbs have to be built closely to each other. The
inter-dependency of verbs and the other parts of speech in Persian is an interesting
fact that is not usually found in the other languages.

This WordNet can be evaluated in the following ways: first, we have to compare
the results with 3 reliable bilingual dictionaries; second, some human experts check
and evaluate the synsets, third, when completed, we have to use the WordNet in some
applications and evaluate the results and fourth, the WordNet has to be compared with
other lexicons built with another approach.

Acknowledgements

Special thanks to Dr. Ali Famian for his everlasting supports and ideas and Dr.
Mohammad Dabir-Moghaddam for his innovative ideas and findings on compound
verbs in Persian. We are also thankful to Mr. Alireza Mokhtaripour and Dr. Mohsen
Ebrahim Moghaddam for their efforts to provide us the electronic dictionary.

References

Coverage, Base Concepts and Top Ontology. J. Computers and the Humanities, Special
Developing FarsNet: A Lexical Ontology for Persian

Mehrnoush Shamsfard

NLP Research Laboratory, Faculty of Electrical & Computer Engineering,
Shahid Beheshti University, Tehran, Iran.
m-shams@sbu.ac.ir

Abstract. Semantic lexicons and lexical ontologies are important resources in natural language processing. They are used in various tasks and applications, especially where semantic processing is evolved such as question answering, machine translation, text understanding, information retrieval and extraction, content management, text summarization, knowledge acquisition and semantic search engines. Although there are a number of semantic lexicons for English and some other languages, Persian lacks such a complete resource to be used in NLP works. In this paper we introduce an ongoing project on developing a lexical ontology for Persian called FarsNet. It uses a combination of WordNet and FrameNet features to represent words meanings. We exploited a hybrid semi-automatic approach to acquire lexical and conceptual knowledge from resources such as WordNet, bilingual dictionaries, mono-lingual corpora and morpho-syntactic and semantic templates. FarsNet provides links between various types of words and also between words and their corresponding concepts in other ontologies.

Keywords: Lexical Ontology, Semantic Lexicon, Persian, WordNet, FrameNet.

1 Introduction

In recent years, there has been an increasing interest in semantic processing of natural languages. Some of the essential resources to make this kind of process possible are semantic lexicons and ontologies. Lexicon contains knowledge about words and phrases as the building blocks of language and ontology contains knowledge about concepts as the building blocks of human conceptualization (the world model)[1]. Lexical ontologies or NL-ontologies are ontologies whose nodes are lexical units of a language. Moving from lexicons toward ontologies by representing the meaning of words by their relations to other words, results in semantic lexicons and lexical ontologies.

One of the most popular lexical ontologies for English is WordNet. Princeton WordNet [2], is widely used in NLP research works. It covers English language and has been first developed by Miller in a hand-crafted way. Many other lexical ontologies (such as EuroWordNet, BalkaNet, …) have been created based on Princeton WordNet for other languages such as Dutch, Italian, Spanish, German, French, Czech and Estonian. Although there exist such semantic, lexical resources for English and some other languages, some languages such as Persian (Farsi) lack such a
semantic resource for use in NLP works. There have been some efforts to create a WordNet for Persian language too [3,4] but no available product have been announced yet. The only available lexical resources for Persian are some lexicons containing phonological and syntactic knowledge of words (such as [5]).

On the other hand, the major problems with WordNet are (1) its restricted relations (2) weak semantic knowledge on verbs. WordNet does not support cross-POS relationships and does not allow defining arbitrary relations. There are not coded information about verb arguments and their conceptual properties in WordNet.

In this paper we introduce an effort to develop a lexical ontology called FarsNet for Persian language which overcomes the above shortcomings. We exploit a semi automatic approach to acquire lexical and ontological knowledge from available resources and build the lexicon. FarsNet is a bilingual lexical ontology which not only represents the meaning of Persian words and phrases, but also links them to their corresponding concepts in other ontologies such as WordNet, Cyc, Sumo, etc. FarsNet aggregates the power of WordNet on nouns and the power of frameNet on verbs.

At the rest of the paper first I will discuss construction of lexicons based on WordNet, then discussing new features of FarsNet; I will explain our approach in brief.

2 Construction of Semantic Lexicons Based on Princeton WordNet

Semantic lexicons may be generated using automatic or manual methods. Manual approach requires direct interference of human and is estimated be a time-consuming task; therefore the use of automatic methods seems to be more desirable. One of the major resources for creating a semantic lexicon for a language (other than English) is Princeton WordNet that was constructed for English.

However it should be noted that although concepts are represented in the form of different words in different languages, the relation of these concepts are almost the same. Therefore we may take advantage of Princeton WordNet as a main resource for the development of WordNets for different languages.

The main challenges in this procedure are the lexical gaps that exist among different languages and the ambiguities produced during the translation procedures. Lexical gap results in words in a language that have no direct mate in the other language and can only be translated into a group of words that convey the same meaning instead of a single word. Ambiguities are results of translating polysemous words from one language to other polysemous words in other languages in creating a WordNet from other.

There are some proposed approaches to overcome these problems and build new WordNets for new languages based on Princeton WordNet. In this paper we use some of them to create some parts of FarsNet which is related to WordNet.
3 Introducing FarsNet

FarsNet consists of two main parts: a semantic lexicon and a lexical ontology. Each entry in the semantic lexicon contains natural language descriptions, phonological, morphological, syntactic and semantic knowledge about a lexeme. The lexemes can participate in relations with other lexemes in the same lexicon or to entries of other lexicons and ontologies, in the ontology part. Here, the semantic lexicon is serving as a lexical index to the ontology. The ontology part contains not only the standard relations defined in WordNet but also some additional conceptual ones. FarsNet is able to add new relations for its words or concepts. We have developed an interface for FarsNet from which one can add, remove or change the entries. From this interface the user can define new relations or use the existing ones and relate words by them. It can relate words from different syntactic types together (e.g. nouns to adjectives and verbs). It can also relate a word to its corresponding concept in an existing ontology. This makes the interoperability between various resources and various languages easier.

These general features are available for all types of words. In addition there are some specific features for specific POS tags too. For instance, adjectives may accept selectional restrictions. This way, in addition to features defined by WordNet, we have defined a new relation for adjectives which shows the category of nouns who can accept this word as a modifier. For example ‘khoshmazeh’ (delicious) usually is used for edibles while ‘dana’ (wise) is used for humans. This feature, showing the selectional restrictions of Persian adjectives, helps NLP systems to disambiguate syntactic parsing, chunking and understanding.

On the other hand FarsNet covers the relations introduced for verbs in WordNet and also adds the number, names and conceptual characteristics of the arguments of each verb in a similar way to FrameNet. We have defined the arguments for about 300 Persian verbs [6]. The next activity is to complete it for other verbs and define the feature set of each argument. For example now we know that ‘khordan’ (to eat) is a verb belonging to a verb class which needs an agent and a theme and can have an instrument, but we have not yet defined that the theme of this verb should be edible, its agent should be an animated thing, and the size of its instrument is small (usually smaller than a mouth) and it may be one of spoon, fork, knife, …. This feature helps NLP systems to extract thematic roles, represent the sentence meaning and acquire knowledge from texts.

The next section will discuss the building procedure of FarsNet.

4 Semi-Automatic Knowledge Acquisition for FarsNet

We use an incremental approach to build FarsNet; developing a kernel and extending it in a semi automatic way. The acquisition approach consists of the following main steps:
1- Providing initial resources
2- Developing an initial lexicon based on WordNet and performing word sense disambiguation
3- Extracting new knowledge (words and relations) from available resources
4- Evaluation and refinement

4.1 Initial Resources

We have the following resources available and use them to develop FarsNet.

- WordNet
  - a Lexicon [5] containing more than 50,000 entries with their POS tags,
  - a bilingual (English- Persian) dictionary.
- POS tagged corpora
  - a morphological analyzer for Persian [7]

4.2 Developing an Initial Lexicon Based on WordNet

To develop an initial lexicon we exploited three separate approaches in parallel: (a) automatic creation of a small kernel containing just the base concepts (2) automatic creation of an initial big lexicon containing almost anything covered by the bilingual dictionary (3) manually gathering a small lexicon.

For (a) we should start form English base concepts and translate them to Persian, but for (b) we move in two directions, from English to Persian and from Persian to English separately to compare their results.

Moving from Persian is simpler. Each Persian word will be assigned to an English synset using a Persian-English dictionary. To move from English, for each English synset, first we translate all the words in the synset using an electronic bilingual dictionary. Then we should arrange the Persian synsets by exploiting some heuristics and WSD (word sense disambiguation) methods. It is obvious that each synset has some English words and each word may have several senses and each sense may have several translations to Persian. So creating Persian synsets from English ones is not a straight forward task and each Persian word may be connected to a group of synsets in WordNet. For example the Persian word “dast” (hand) is connected to 14 synsets in WordNet. Some of them are listed below:

- Hand, Manus, mitt, paw -- the (prehensile) extremity of the superior limb;
- Hired hand, hand, hired man -- a hired laborer on a farm or ranch
- Handwriting, hand, script -- something written by hand

It can be seen that from all these 14 synsets, only the first one is a valid choice for the Persian word “dast”.

Therefore it is important to identify the right sense(s) of English word, the right translation of it and putting the right sense of translated word in the corresponding synset. We use some heuristics to find the corresponding synsets fast. For example, to
find the appropriate Persian synset for an English one, we consider word pairs in the English synset. For each word in this pair we list all synsets they appear in. If those two words appear together only in the current synset, their common Persian translations would be connected to that synset. The existence of a single common synset in fact implies the existence of a single common sense between the two words and therefore their Persian translations shall be connected to this synset.

On the other hand, if a word is known to be the English equivalent of a Persian word according to dictionary, the Persian word should at least be connected to one of the synsets that include the English word as a member. There will obviously be no ambiguities if the English word has only one sense and so appears at only one synset. In this case its translations will be added to that synset too.

We plan to use dictionary based WSD according to Lesk [8] too. In this approach we use other English translations of Persian word (PW) as context words.

### 4.3 Extracting New Knowledge

After creating the initial lexicon, extra words will be gathered from a tagged corpus, and assign to a synset as mentioned before.

Another part of ontology learning in FarsNet is dedicated to finding some relations from corpora exploiting lexico-syntactic patterns. The patterns we have tested so far are some adaptations of Hearst’s patterns for Persian. We are going to test other templates introduced at [9] too.

### 4.4 Evaluation

As it was mentioned before, we build each part of FarsNet using more than one approach. The evaluation procedure is done by tow methods too. The manual evaluation of the part of lexicon built so far shows an accuracy of about 70% in the resulting Persian lexicon.

In the second method we compare the results of various exploited methods on a common task to find the common built knowledge. For example to confirm the inclusion hierarchies, we extract hierarchical relations from text using templates in one hand and find this hierarchy according to the hyponym/hypernym relations between corresponding English synsets on the other hand. Comparing the results shows the most confident knowledge extracted by both two methods. However as it is an ongoing project, the evaluation procedures as well as some other parts are not complete yet.

### 5 Conclusion

FarsNet project is an ongoing project in NLP research laboratory of Shahid Beheshti University. The ontology development methodology we proposed for developing
FarsNet forces some series of tasks to be done in parallel and then combining or comparing the results.

We have done the following parallel activities:
- Manually developing a small lexicon as the kernel of FarsNet containing 2500 entries [7].
- Manually translating the base concepts of WordNet into Persian
- Automatic finding the corresponding WordNet synsets for each entry of the syntactic lexicon using the bilingual dictionary.
- Automatic making the preliminary list of potential synsets for Persian using WordNet and the above translations.
- Automatic learning of new words and relations from the tagged corpus.

Although a base ontology has been created with 32000 persian synsets, but there are many things to be added yet. The following activities are within our further works to continue the project:
- Exploiting (linking to) FrameNet as a basis for developing the verb knowledge base of FarsNet
- Completing the verbs knowledge base,
- Enhancing the sense disambiguation modules in the automatic translations
- Designing and using more templates to extract non-taxonomic relations from text.
- Working on some statistical approaches for lexical acquisition
- Exploiting other methods to learn ontological knowledge
- Finding a mapping between various ontologies.
- Integrating the works done.

References

KUI: Self-organizing Multi-lingual
WordNet Construction Tool

Virach Sornlertlamvanich¹, Thatsanee Charoenporn¹,
Kergrit Robkop¹, and Hitoshi Isahara²

¹ Thai Computational Linguistics Lab.,
NICT Asia Research Center, Pathumthani, Thailand
² National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gaun, Kyoto, Japan 619-0289
{virach, thatsanee, krergrit}@tcllab.org, isahara@nict.go.jp

Abstract. This paper describes a multi-lingual WordNet construction tool,
called KUI (Knowledge Unifying Initiator), which is a knowledge user
interface for online collaborative knowledge construction. KUI facilitates online
community in developing and discussing multi-lingual WordNet. KUI is a sort
of social networking system that unifies the various discussions following the
process of thinking model, i.e. initiating the topic of interest, collecting the
opinions to the selected topics, localizing the opinions through the translation or
customization and finally posting for public hearing to conceptualize the
knowledge. The process of thinking is done under the selectional preference
simulated by voting mechanism in the case that there are many alternatives. By
measuring the history of participation of each member, KUI adaptively
manages the reliability of each member’s opinion and vote according to the
estimated ExpertScore. As a result, the multi-lingual WordNet can be created
online and produce a reliable result.

Keywords: Multi-lingual WordNet, KUI, ExpertScore, social networking
system, information reliability.

1 Introduction

The constructions of the WordNet [1] for languages can be varied according to the
availability of the language resources. Some were developed from scratch, and some
were developed from the combination of various existing lexical resources. Spanish
and Catalan Wordnets¹, for instance, are automatically constructed using hyponym
relation, mono-lingual dictionary, bi-lingual dictionary and taxonomy [2]. Italian
WordNet [3] is semi-automatically constructed from definition in mono-lingual
dictionary, bi-lingual dictionary, and WordNet glosses. Hungarian WordNet uses bi-
lingual dictionary, mono-lingual explanatory dictionary, and Hungarian thesaurus in
the construction [4], etc.

¹ http://www.lsi.upc.edu/~nlp/
A tool to facilitate the construction is one of the important issues related to the WordNet construction. Some of the previous efforts were spent for developing the tools such as Polaris [5], the editing and browsing for EuroWordNet, and VisDic [6], the XML based Multi-lingual WordNet browsing and editing tool developed by Czech WordNet team. To facilitate an online collaborative development and annotate a reliability score to the proposed word entries, we, therefore, proposed KUI (Knowledge Unifying Initiator) to be a Knowledge User Interface (KUI) for online collaborative construction of multi-lingual WordNet. KUI facilitates online community in developing and discussing multi-lingual WordNet. KUI is a sort of social networking system that unifies the various discussions following the process of thinking model, i.e. initiating the topic of interest, collecting the opinions to the selected topics, localizing the opinions through the translation or customization and finally posting for public hearing to conceptualize the knowledge. The process of thinking is done under the selectional preference simulated by voting mechanism in the case that there are many alternatives.

This paper illustrates an online tool to facilitate the multi-lingual WordNet construction by using the existing resources having only English equivalents and the lexical synonyms. Since the system is opened for online contribution, we need a mechanism to inform the reliability of the result. We introduce ExpertScore which can be estimated from the history of the participation of each member. The weight of each vote and opinion will be determined by the ExpertScore. The result will then be ranked according to this score to show the reliability of the opinion.

The rest of this paper is organized as follows: Section 2 describes the process of managing the knowledge. Section 3 explains the design of KUI for Collaborative resource development. Section 4 provides some examples of KUI for WordNet construction. And, Section 5 concludes our work.

2 Process of Knowledge Development

A thought is dynamically formed up by a trigger which can be an interest from inside or a proposed topic from outside. However, knowledge can be formed up from the thought only when managed in an appropriate way. Since we are considering the knowledge of a community, we can consider the knowledge that is formed by a community in the following manner.

- Knowledge is managed by the knowledge users.
- Knowledge is dynamically changed.
- Knowledge is developed in an individual manner or a community manner.
- Knowledge is both explicit and tacit.

The environment of online community can successfully serve the requirement of knowledge management. Under the environment, the knowledge should be grouped up and narrowed down into a specific domain for each group. The domain specific group can then be managed to generate a concrete knowledge after receiving the consensus from the participants at any moment.

Open Source software development is a model for open collaboration in the domain of software development. The openness of the development process has
successfully established a largest software community that shares their development and using experience. The activities are dedicated to the domain of software knowledge development. SourceForge.net\(^2\) is a platform for project based Open Source software development. Open Source software developers deploy SourceForge.net to announce their initiation, to call for participation, to distribute their works and to receive feedbacks concerning their proposed software. Developers and users are actively using SourceForge.net to communicate with each other.

Adopting the concept of Open Source software development, we will possibly be able to develop a framework for domain specific knowledge development under the open community environment. Sharing and collaboration are the considerable features of the framework. The knowledge will be finally shared among the communities by receiving the consensus from the participants in each step. To facilitate the knowledge development, we deliberate the process into 4 steps.

1) **Topic of interest**
The topic will be posted to draw the intention from the participants. The selected topics will then be further discussed in the appropriate step.

2) **Opinion**
The selected topic is posted to call for opinions from the participants in this step. Opinion poll is conducted to get the population of each opinion. The result of the opinion poll provides the variety of opinions that reflects the current thought of the communities together with the consensus to the opinions.

3) **Localization**
Translation is the straightforward implementation of the localization. Collaborative translation helps producing the knowledge in multiple languages in the most efficient way.

4) **Public-Hearing**
The result of discussion will be revised and confirmed by gathering the opinions to the final draft of proposal.

Fig. 1 shows the process of how knowledge is developed within a community. Starting from posting 'Topic of Interest', participants express their supports by casting a vote. Upon a threshold the 'Topic of Interest' is selected for conducting a poll on 'Opinion', or introducing to the community by 'Localization', or posting a draft for 'Public-Hearing' to gather feedbacks from the community. The transition from 'Opinion' to either 'Localization' or 'Public-Hearing' occurs when the 'Opinion' has a concrete view for implementation. The discussion in 'Localization' and 'Public-Hearing' is however interchangeable due to purpose of implementation whether to adopt the knowledge to the local community or to get feedbacks from the community.

The knowledge creating is managed in 4 different categories corresponding to the stage of knowledge. Each individual in the community casts a vote to rank the appropriateness of solutions at each category. The community can then form the

---

community knowledge under the ‘Selectional Preference’ background. On the other hand, the under-threshold solutions become obsolete by nature of the ‘Selectional Preference’.

![Diagram of knowledge development](image.jpg)

**Fig. 1.** Process of knowledge development

### 3 Knowledge User Interface for Knowledge Unifying Initiative

#### 3.1 What is KUI?

KUI is a GUI for knowledge engineering, in other words Knowledge User Interface (KUI). It provides a web interface accessible for pre-registered members. An online registration is offered to manage an account by profiling the login participant in making contribution. A contributor can comfortably move around in the virtual space from desk to desk to participate in a particular task. A working desk can be a meeting place for collaborative work that needs discussion through the ‘Chat’, or allow a contributor to work individually by using the message slot to record each own comment. The working space can be expanded by closing the unnecessary frames so the contributor can concentrate on the task. All working topics can be statistically viewed through the provided tabs. These tabs help contributors to understand KUI in the aspects of the current status of contribution and the tasks. A knowledge community can be formed and can efficiently create the domain knowledge through the features provided by KUI. These KUI features fulfill the process of human thought to record the knowledge.

KUI also provides a ‘KUI look up’ function for viewing the composed knowledge. It is equipped with a powerful search and statistical browse in many aspects. Moreover, the ‘Chatlog’ is provided to learn about the intention of the knowledge composers. We frequently want to know about the background of the solution for better understanding or to remind us about the decision, but we cannot find one. To avoid the repetition of a mistake, we systematically provide the ‘Chatlog’ to keep the trace of discussion or the comments to show the intention of knowledge composers.
3.2 Feature of KUI

- **Poll-based Opinion or Public-Hearing**
  A contributor may choose to work individually by posting an opinion e.g. localization, suggestion etc., or join a discussion desk to conduct 'Public-Hearing' with others on the selected topic. The discussion can be conducted via the provided 'Chat' frame before concluding an opinion. Any opinions or suggestions are committed to voting. Opinions can be different but majority votes will cast the belief of the community. These features naturally realize the online collaborative works to create the knowledge.

- **Individual or Group works**
  Thought may be formed individually or though a concentrated discussion. KUI facilitates a window for submitting an opinion and another window for submitting a chat message. Each suggestion can be cast through the 'Opinion' window marked with a degree of its confidence. By working individually, comments to a suggestion can be posted to mark its background to make it better understanding. On the other hand, when working as a group, discussions among the group participants will be recorded. The discussion can be resumed at any points to avoid the iterating words.

- **Record of Intention**
  The intention of each opinion can be reminded by the recorded comments or the trace of discussions. Frequently, we have to discuss again and again on the result that we have already agreed. Misinterpretation of the previous decision is also frequently faced when we do not record the background of decision. Record of intention is therefore necessary in the process of knowledge creation. The knowledge interpretation also refers to the record of intention to obtain a better understanding.

- **Selectional Preference**
  Opinions can be differed from person to person depending on the aspects of the problem. It is not always necessary to say what is right and what is wrong. Each opinion should be treated as a result of intelligent activity. However, the majority accepted opinions are preferred at the moment. Experiences could tell the preference via vote casting. The dynamically vote ranking will tell the selectional preference of the community at the moment.

3.3 *ExpertScore*

KUI heavily depends on members’ voting score to produce a reliable result. Therefore, we introduce an adjustable voting score to realize a self-organizing system. Each member is initially provided a default value of voting score equals to one. The voting score is increased according to *ExpertScore* which is estimated by the value of *Expertise*, *Contribution*, and *Continuity* of the participation history of each member. *Expertise* is a composite score of the accuracy of opinion and vote, as shown in Equation 1. *Contribution* is a composite score of the ratio of opinion and vote posting
comparing to the total, as shown in Equation 2. *Continuity* is a regressive function based on the assumption that the absence of participation of a member will gradually decrease its *ExpertScore* to one after a year (365 days) of the absence, as shown in Equation 3.

\[
\text{Expertise} = \alpha \frac{\text{count}(\text{BestOpinion})}{\text{count}(\text{Opinion})} + \beta \frac{\text{count}(\text{BestVote})}{\text{count}(\text{Vote})}.
\]

\[
\text{Contribution} = \gamma \frac{\text{count}(\text{Opinion})}{\text{count}(\text{TotalOpinion})} + \rho \frac{\text{count}(\text{Vote})}{\text{count}(\text{TotalVote})}.
\]

\[
\text{Continuity} = 1 - \left( \frac{D}{365} \right)^4.
\]

where,

\[
\alpha + \beta + \gamma + \rho = 1
\]

D is number of recent absent date (0 ≤ D ≤ 365)

As a result, the *ExpertScore* can be estimated by Equation 4.

\[
\text{ExpertScore} = \left( 1 - \left( \frac{D}{365} \right)^4 \right) \times \left( \alpha \frac{\text{count}(\text{BestOpinion})}{\text{count}(\text{Opinion})} + \beta \frac{\text{count}(\text{BestVote})}{\text{count}(\text{Vote})} \right) + \gamma \frac{\text{count}(\text{Opinion})}{\text{count}(\text{TotalOpinion})} + \rho \frac{\text{count}(\text{Vote})}{\text{count}(\text{TotalVote})}
\]

If *ExpertScore* is less than one, it is reset to one. The value of *ExpertScore* is ranged between 1 to 365 according to the accuracy and the rate of contribution of each member. This means that reliable members are rewarded better score for each vote. However, the expertise of the member is decreased according to the continuity of the participation. By means of the *ExpertScore*, we can rank the opinions precisely and yield reliable results, especially for the results produced by an online community.

## 4 KUI for Multi-lingual WordNet Construction

There are some previous efforts in developing WordNets for Asian languages, e.g. Chinese, Japanese, Korean [7], [8], [9], [10], [11] and Hindi [12]. The number of languages that have been successfully developed their WordNets is still limited to some active research in this area. However, the extensive development of WordNet in other languages is important, not only to help in implementing NLP applications in each language, but also in inter-linking WordNets of different languages to develop multi-lingual applications to overcome the language barrier.
We adopt the proposed criteria for automatic synset assignment for Asian languages which has limited language resources. Based on the result from the above synset assignment algorithm, we provide KUI (Knowledge Unifying Initiator) [13], [14] to establish an online collaborative work in refining the WorNets. KUI allows registered members including language experts revise and vote for the synset assignment. The system manages the synset assignment according to the preferred score obtained from the revision process. The revision history reflects in the ExpertScore of each participant, and the reliability of the result is based on the summation of the ExpertScore of the contributions to each record. In case of multiple mapping, the record with the highest score is selected to report the mapping result.

As a result, the community WordNets will be accomplished and exported into the original form of WordNet database. Via the synset ID assigned in the WordNet, the system can generate a cross language WordNet result. Through this effort, an initial version of Asian WordNet can be established.

Table 1 shows a record of WordNet displayed for translation in KUI interface. English entry together with its part-of-speech, synset, and gloss are provided if exists. The members will examine the assigned lexical entry whether to vote for it or propose a new translation.

Table 1. A record of WordNet.

<table>
<thead>
<tr>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS : NOUN</td>
</tr>
<tr>
<td>Synset : auto, automobile, machine, motorcar</td>
</tr>
<tr>
<td>Gloss : a motor vehicle with four wheels; usually propelled by an internal combustion engine</td>
</tr>
</tbody>
</table>
Fig. 2 illustrates the translation page of KUI. In the working area, the login member can participate in proposing a new translation or vote for the preferred translation to revise the synset assignment. Statistics of the progress as well as many useful functions such as item search, record jump, chat, list of online participants are also provided. KUI is actively facilitating members in revising the Asian WordNet database.

Fig. 3 illustrates the lookup page of KUI. The returned result of a keyword lookup is sorted according to the best translated word of each language. The best translated word is determined by the highest vote score. As a result, the user can consult the WordNet to obtain a list of equivalent words of the same sense sorted by the languages. The ExpertScore provided in KUI will help selecting the best translation of each word.

5 Conclusion

KUI is a platform for composing knowledge in the Open Source style. A contributor can naturally follow the process of knowledge development that includes posting in 'Topic of interest', 'Opinion', 'Localization' and 'Public-Hearing'. The posted items are committed to voting to perform the selectional preference within the community. The results will be ranked according to the vote preference estimated by the ExpertScore for the purpose of managing the multiple results. 'Chatlog' is kept to indicate the record of intention of knowledge composers. A contributor may participate KUI individually or join a discussion group to compose the knowledge. We are expecting KUI to be a Knowledge User Interface for composing the knowledge in the Open Source style under the monitoring of the community. The statistical-base visualized

3 http://www.tclab.org/kui/
'KUI look up' is also provided for the efficient consultation of the knowledge. We introduce KUI for Asian WordNet development. The ExpertScore efficiently ranks the results especially in the case where there is more than one equivalent.

References

Extraction of Selectional Preferences for French using a Mapping from EuroWordNet to the Suggested Upper Merged Ontology

Dennis Spohr
Institut für Linguistik/Romanistik
Universität Stuttgart
Stuttgart, Germany
dennis.spohr@ling.uni-stuttgart.de

Abstract. This paper presents an approach to extracting selectional preferences of French verbal predicates with respect to the ontological types of their arguments. Selectional preference is calculated on the basis of Resnik’s measure of selectional association between a predicate and the class of its argument [1]. However, instead of using WordNet synsets to express sortal restrictions (cf. [2]), we employ conceptual classes of the Suggested Upper Merged Ontology (SUMO; [3]) that have been automatically mapped to synsets of the French EuroWordNet [4] in a generic way that is in principle applicable to all WordNets which are linked to the Inter-Lingual-Index.

1 Introduction

Lexical-semantic NLP and with it semantic lexicons have become increasingly important over the last decades, and the contribution of (Euro-)WordNet [5, 4] and FrameNet [6] within this field is of course so fundamental and well-known that it need not be discussed here. However, recent years have further seen a strong tendency towards interfacing such resources with knowledge bases or taxonomies of general knowledge, both commonly referred to as ontologies. Well-known examples of such efforts are e.g. [7], who linked EuroWordNet’s Inter-Lingual-Index to a number of base concepts and a top ontology as integral part of the EuroWordNet project, [8] who mapped Princeton WordNet to the Suggested Upper Merged Ontology (SUMO), and [9] who linked FrameNet and SUMO. Moreover, the recent Global WordNet Grid\textsuperscript{1} is pursuing such efforts on a considerable scale to create mappings from SUMO to all existing WordNets.

One of the main reasons why such approaches are so important is that while resources like (Euro-)WordNet and FrameNet attempt to model lexical-semantic knowledge, ontologies try to mediate common knowledge or knowledge of the world. Therefore, linking these two types of resources may be able to bridge the gap between language-dependent lexical knowledge and language-independent facts or statements about the world.

\textsuperscript{1}http://www.globalwordnet.org/gwa/gwa_grid.htm
Such statements appear to have a more universal character, and this is what makes combinations of ontological and lexical-semantic resources interesting for the formulation of selectional restrictions or preferences. We believe that a statement like “X prefers subjects of type Human or CognitiveAgent” is – from a meta-linguistic perspective – more informative than saying “X prefers the subjects \{human_1, individual_1, mortal_1, person_1, someone_1, soul_1\} or …”. In this paper, we present a general methodology for mapping EuroWordNets to the SUMO ontology by using both an existing mapping from Princeton WordNet 1.6 to SUMO [8] and the linking of the EuroWordNets to the Inter-Lingual-Index [7]. We apply our methodology to the French EuroWordNet and extract sortal selectional preferences that are calculated on the basis of an established measure of selectional association between a predicate and the classes of its argument [1]. Section 2 of this paper introduces some background on WN and SUMO, and Resnik’s approach to selectional preference extraction. In Section 3, we will present our methodology for mapping EWN to SUMO, and Section 4 gives details on how we extract selectional preferences based on this mapping. After an evaluation of the mapping methodology and selectional preference extraction, we conclude in Section 6 and briefly discuss ways to apply and further extend our approach.

2 Background

2.1 WordNet and the Suggested Upper Merged Ontology

In this section, we will briefly discuss work that has been done on linking WordNet to the Suggested Upper Merged Ontology. [8] have created such a mapping for version 1.6 of WordNet, and have in subsequent years released new mappings for each new version of WordNet, with the latest release in summer 2007 for WN3.0. In creating their linking, [8] have decided to use the following three mapping relations: synonymy (equivalence, indicated by ’=’; cf. Section 3.2), hypernymy (subclass-superclass relation, indicated by ‘+’), and instantiation (indicated by ’@’). In contrast to the approach presented in their paper, we try to create mappings automatically, though relying heavily on their manual preparatory work.

As was mentioned in the introduction, the Global WordNet Grid initiative, which was launched in early 2006, is trying to provide WordNet-SUMO mappings for all existing WordNets. The current state, as of late 2007, comprises mappings for 5,000 English base concepts, as well as for the Spanish and Catalan WordNets.

2.2 Selectional Preference Extraction

The measure we use for the calculation of selectional preference is that of [1]², who uses the notion of relative entropy known from information theory [11]. The strength of selectional preference $S_R(p)$ of a predicate $p$ with respect to a grammatical relation $R$ is

² See [10] for a recent survey of several other approaches to selectional preference acquisition.
defined as follows.

**Selectional preference strength:**

\[ S_R(p) = \sum_c P_r(c|p) \log \frac{P_r(c|p)}{P_r(c)} \]

The better \( P_r(c) \) approximates \( P_r(c|p) \), the closer \( \log \frac{P_r(c|p)}{P_r(c)} \) is to 0, i.e. the less influence \( p \) has on its argument, and therefore the less strong is its selectional preference. The selectional preference strength is on the one hand an indicator as to “how much information [...] predicate \( p \) provides about the conceptual class of its argument” ([1]: p. 53). On the other hand, it is used for normalising the selectional preference values of a predicate, in order to be able to compare the values of different predicates: a predicate that is generally weak in showing preferences will thus receive a higher value if it really shows a preference for a particular conceptual argument class. Selectional preference for a particular conceptual class is calculated in the form of selectional association \( A_R(p, c) \) between \( p \) and the class \( c \) of its argument. Its definition is given below.

**Selectional association:**

\[ A_R(p, c) = \frac{1}{S_R(p)} P_r(c|p) \log \frac{P_r(c|p)}{P_r(c)} \]

As Resnik points out, the fact that text corpora have usually not been annotated with explicit and unambiguous conceptual classes requires some sort of distribution of frequencies among the possible conceptual class of a word. The following formula calculates the frequency of predicate \( p \) and class \( c \).

**Frequency propagation:**

\[ freq_R(p, c) \approx \sum_{w \in c} \frac{count_R(p, w)}{classes(w)} \]

This means that the actual frequency count of a word \( w \), which stands in relation \( R \) (e.g. verb-object) to \( p \), is distributed equally among the classes \( c \) which \( w \) is a member of. In a hierarchical resource such as SUMO, this also has the effect of propagating the \( freq \) value up the hierarchy: if \( w \) is a member of class \( c \), it is, of course, also a member of the superclasses of \( c \), and thus \( freq(p, c) \) is also added to all the superclasses of \( c \).

## 3 Mapping EuroWordNet to SUMO

In this section, we will present how the French EuroWordNet has been mapped onto conceptual classes of the Suggested Upper Merged Ontology. The general methodology of creating the mapping to the French EuroWordNet is described in the following
subsection. Although a quite recent mapping to version 3.0 of WordNet exists, we
decided to use the very first mapping – namely that of WordNet version 1.6 – as a starting
point. The reasons for doing so mainly concern the sensemaps between different versions
of WordNet, and are explained in detail in Section 3.2 below.

3.1 General Methodology

As was just mentioned, we use the mapping of SUMO to version 1.6 of WordNet in order
to link the French EuroWordNet to SUMO. The French EWN itself – as is the case with
all EuroWordNets – is linked to the Inter-Lingual-Index, a set of concepts that is intended
to be largely language-independent (cf. [7]). A crucial prerequisite for our approach
to function is that the identifiers of entities in the Inter-Lingual-Index correspond to
synset identifiers in version 1.5 of WordNet. For example, entity 00058624-n of
the Inter-Lingual-Index, which is glossed by “the launching of a rocket under its own
power”, corresponds to synset \{décollage_1,lancement_d’une_fusée_1\} in
the French EWN and to \{blastoff_1,rocket_firing_1,rocket_launching_1,shoot_1\} in WN1.5. Starting from these observations, i.e. the mapping of
SUMO to WN1.6 and the linking of the French EWN to the Inter-Lingual-Index (≈
WN1.5), the only remaining task that is left is to move from WN1.5 to WN1.6. In order
to do this, we can avail ourselves of the sensemap files that came with the 1.6 release
of WordNet, which indicate the changes from WN1.5 to WN1.6. In order
to do this, we can avail ourselves of the sensemap files that came with the 1.6 release
of WordNet, which indicate the changes from WN1.5 to WN1.6. Ignoring particular
mapping issues for the moment (see Section 3.2 below), the resulting EuroWordNet
entries look like the one shown in Figure 1. The structure is based on the format suggested
by the Global WordNet Grid. The whole mapping process is summarised in Figure 2.

3.2 WordNet Sensemaps

Whenever updates of WordNet are released, the updated version comes with files
that, among others, indicate changes in the structure of the synsets. For example,
synset 00058624-n from above has been split in the step from WN1.5 to WN1.6:
{shoot_1} is now a member of synset 00078261-n, {blastoff_1} of synset
00065319-n, and {rocket_firing_1, rocket_launching_1} of synset
00065148-n. Therefore, version 1.6 contains new synsets that did not exist in WN1.5,
and further cases in which a synset is reorganised thus that some of its items belong to
different synsets in the updated version. The primary problem for the task of mapping
such instances comes from the fact that individual members of a synset do not have
unique identifiers themselves, but only the synset as a whole\(^3\). Therefore, when a synset
has been split, it is not possible to automatically determine the correct position at which
the synset has to be split in a different language, or even whether it has to be split at all.
Moreover, each update comes with a large number of such changes, and therefore using

\(^3\) This is, of course, not a problem of the WordNet approach, but rather of the fact that there is no
one-to-one mapping between languages.
the most recent mapping between SUMO and WN3.0, which is without a doubt desirable, would multiply the inaccuracies in the mapping right from the start. Just imagine a case where a synset has been split e.g. from WN1.5 to WN1.6, and the new synset is then split again when going to WN1.7, and so on.

Fig. 1. Sample EuroWordNet entry of synset 00002728-n (organisme, forme_de_vie, être, vie) after the mapping

The decision that was made for cases like these is to assign to the original synset two (or more if necessary) SUMO classes: first the one that has been mapped to this synset, and second the ones to which the new (or relevant existing) synsets have been mapped in WN1.6. The justification of this decision is based on the assumption that on a level as abstract as that of SUMO conceptual classes, a “slight” reorganisation of the synsets and some of their items should not lead to significant conceptual clashes, as this would imply that grave errors had been made when putting the respective senses into one synset in the first place. In Figure 3 below, which depicts the entry of synset 00058624-n after the mapping, we see that the two SUMO classes that have been assigned to this synset do at least remotely fit the senses: more specific than Impelling and Motion, and equivalent to Shooting. Of course, a qualitative evaluation is needed to determine the
4 Extraction of Selectional Preferences

4.1 Corpus extraction

The (potential) nominal arguments of the verbal predicates have been extracted from a portion of more than 350 million tokens from the French Agence France-Presse corpus licensed by the Linguistic Data Consortium\footnote{http://www.ldc.upenn.edu/}. The corpus has been part-of-speech tagged using the French TreeTagger parameter files\cite{12} and has been stored in the widely-used Corpus Workbench format\cite{13}. Figure 4 below shows the CQP query that extracted potential direct objects of ‘manger’.

We have decided to use a quite rigid syntactic structure, and therefore the query contains both the potential subject and direct object although only one of them is focussed on at a time. Lines 1-4 in Figure 4 represent the subject position – with the head of the subject noun phrase at the end of in line 1 –, and the direct object is described in lines 10-12. The verbal predicate, in this case ‘manger’, is shown in line 6. The results of this query, when grouped e.g. by object, look like the following (see Table 1).
Fig. 3. Part of the EuroWordNet entry of synset 00058624-n (décollage_1, lancement_d'une_fusée_1) after the mapping

Fig. 4. CQP query extracting direct objects of 'manger'
### Table 1. Results of the query in Figure 4 after grouping by object

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>pain ('bread')</td>
<td>16</td>
<td>revenu ('revenue')</td>
<td>2</td>
</tr>
<tr>
<td>enfant ('child')</td>
<td>8</td>
<td>chose ('thing')</td>
<td>2</td>
</tr>
<tr>
<td>plat ('dish')</td>
<td>4</td>
<td>nourriture ('nutrition')</td>
<td>2</td>
</tr>
<tr>
<td>glace ('ice')</td>
<td>4</td>
<td>partie ('part')</td>
<td>2</td>
</tr>
<tr>
<td>poisson ('fish')</td>
<td>3</td>
<td>méchoui (≈ 'Arabian dish')</td>
<td>1</td>
</tr>
<tr>
<td>chapeau ('hat')</td>
<td>3</td>
<td>victuaille ('comestible')</td>
<td>1</td>
</tr>
<tr>
<td>cœur ('heart')</td>
<td>3</td>
<td>pelican ('pelican')</td>
<td>1</td>
</tr>
<tr>
<td>steak ('steak')</td>
<td>2</td>
<td>vipere ('viper')</td>
<td>1</td>
</tr>
<tr>
<td>poussin ('poult')</td>
<td>2</td>
<td>raisin ('grape')</td>
<td>1</td>
</tr>
<tr>
<td>abat ('innards')</td>
<td>2</td>
<td>cervelle ('brains')</td>
<td>1</td>
</tr>
<tr>
<td>soupe ('soup')</td>
<td>2</td>
<td>hamburger ('hamburger')</td>
<td>1</td>
</tr>
<tr>
<td>feuille ('leaf')</td>
<td>2</td>
<td>christmas ('christmas')</td>
<td>1</td>
</tr>
</tbody>
</table>

### 4.2 Storage and Retrieval

Before we calculated the selectional preferences, we converted the file containing the SUMO-EuroWordNet mappings to OWL (Web Ontology Language; cf. [14]). We have further created a “class only” OWL version of SUMO based on the XML version of SUMO that is distributed with the KSMSA ontology browser\(^5\) (version 1.0.9.1.1). The reasons for not using the OWL version available from the SUMO project site\(^6\) are (i) that it is difficult to process by the “standard” ontology editing tool Protégé [15], which is mainly due to the fact that SUMO was originally written in the far more expressive Suggested Upper Ontology Knowledge Interchange Format (SUO-KIF\(^7\)) and contains, e.g., entities which are one-place predicates and two-place predicates at the same time and therefore occur in both class and property hierarchies, and (ii) that processing a class hierarchy for frequency propagation is far more straightforward and intuitive than processing a mixed hierarchy (see below). Therefore, if a synset had been mapped onto a SUMO concept in an instance relation (cf. ’?’ in Section 2.1 above), it was still created as an OWL class with a subclass relation to the SUMO concept. We believe that the cognitive differences between the instantiation and hypernymy relations (cf. [8]) can be neglected for this purpose. The two files (SUMO and the EWN-SUMO mapping) were then stored as an RDFS database in the Sesame Framework [16]. The main reason for doing all this is that we thus have the benefit of using OWL’s – and of course RDF’s – built-in subsumption and inheritance mechanism, which is very advantageous since the frequencies for the calculation of selectional preferences have to be propagated along the

---

\(^5\)http://virtual.cvut.cz/ksmsaWeb/browser/title/

\(^6\)http://www.ontologyportal.org/

\(^7\)http://suo.ieee.org/SUO/KIF/suo-kif.html
hierarchy (cf. Section 2.2). A further benefit is that we are thus able to use the Protégé
OWL API\(^8\) and the Sesame API\(^9\) in order to perform the propagation of frequencies.

4.3 Calculation of Selectional Preferences

In order to calculate the selectional association between the verbal predicate and its
arguments, it is necessary to first calculate prior values, i.e. to propagate the frequencies
of all arguments irrespective of the verbal predicate up the SUMO hierarchy. For each
word in the list (cf. Table 1), the synsets it belongs to are looked up in the database.
If it belongs to more than one synset, which is typically the case, then its frequency
is divided by the number of readings (cf. Section 2.2 above). After that, for each of
these synsets, first its equivalent SUMO classes are extracted, and then the frequency is
propagated up the hierarchy along the direct superclass relationship. In case of multiple
inheritance, i.e. one class having more than one direct superclass, the frequency is divided
by the number of direct superclasses, similar to what has already been explained for the
different readings of a word. The result is a structure in which every SUMO class has an
associated prior value. As was already mentioned in Section 2.2, the same is done in
order to determine the posterior values for the words occurring as arguments of a given
verbal predicate, and these are then compared with the prior values. Section 5.2 below
shows and discusses results for two French verbal predicates.

5 Evaluation

5.1 Evaluation of SUMO Mapping

Table 2 below shows the results of the mapping procedure. Lines 1-3 in the table display
the total number of synsets in the French EuroWordNet, as well as the numbers of those
which have or have not received a SUMO mapping. Of those 22,351 synsets which have
been assigned a SUMO class (cf. lines 4-6), 98.54% have been assigned exactly one class,
whereas 0.96% have been mapped to two and 0.50% to three or more SUMO classes\(^{10}\).
In line 8 we see that almost 55% of the synsets that have been assigned SUMO classes
occurred in multiple sensemaps, but were all mapped onto synsets belonging to the same
SUMO class, while only 1.46% where mapped onto two or more SUMO classes (cf. line
9). This means that only 1.46% are in principle able to cause “conceptual clashes” when
retaining the strategy presented in Section 3.2. Table 3 displays the 20 most frequent
SUMO classes that have been mapped to synsets in the French EuroWordNet.

---

8 http://protege.stanford.edu/plugins/owl/api/
9 http://www.openrdf.org/doc/sesame/users/ch07.html
10 One synset even received 16 SUMO classes. This was due to the fact that the English synset
contained the highly polysemous ‘cut’, which was split into 26 new synsets in the step from
WN1.5 to WN1.6. The fact that this number is reduced to 16 mappings shows that many of
them are still covered by the same conceptual class in SUMO.
Table 2. Number of SUMO mappings according to different types

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synsets in French EWN</td>
<td>22,745</td>
</tr>
<tr>
<td>with SUMO mapping</td>
<td>22,351</td>
</tr>
<tr>
<td>without SUMO mapping</td>
<td>394</td>
</tr>
<tr>
<td>Of those with SUMO mapping</td>
<td></td>
</tr>
<tr>
<td>with one mapping</td>
<td>22,026</td>
</tr>
<tr>
<td>with two mappings</td>
<td>214</td>
</tr>
<tr>
<td>with three or more mappings</td>
<td>111</td>
</tr>
<tr>
<td>with only one sensemap</td>
<td>9,739</td>
</tr>
<tr>
<td>with more than one sensemap</td>
<td>12,287</td>
</tr>
<tr>
<td>but only one SUMO class</td>
<td></td>
</tr>
<tr>
<td>with more than one sensemap and more than one SUMO class</td>
<td>325</td>
</tr>
</tbody>
</table>

Table 3. Distribution of the top 20 assigned SUMO classes

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubjectiveAssessmentAttribute</td>
<td>1,293</td>
</tr>
<tr>
<td>Device</td>
<td>1,088</td>
</tr>
<tr>
<td>Artifact</td>
<td>689</td>
</tr>
<tr>
<td>Motion</td>
<td>583</td>
</tr>
<tr>
<td>OccupationalRole</td>
<td>555</td>
</tr>
<tr>
<td>Communication</td>
<td>478</td>
</tr>
<tr>
<td>Human</td>
<td>460</td>
</tr>
<tr>
<td>Food</td>
<td>441</td>
</tr>
<tr>
<td>SocialRole</td>
<td>404</td>
</tr>
<tr>
<td>Process</td>
<td>379</td>
</tr>
<tr>
<td>IntentionalProcess</td>
<td>361</td>
</tr>
<tr>
<td>IntentionalPsychologicalProcess</td>
<td>276</td>
</tr>
<tr>
<td>Text</td>
<td>247</td>
</tr>
<tr>
<td>City</td>
<td>246</td>
</tr>
<tr>
<td>StationaryArtifact</td>
<td>243</td>
</tr>
<tr>
<td>NormativeAttribute</td>
<td>238</td>
</tr>
<tr>
<td>EmotionalState</td>
<td>227</td>
</tr>
<tr>
<td>Clothing</td>
<td>223</td>
</tr>
<tr>
<td>DiseaseOrSyndrome</td>
<td>220</td>
</tr>
<tr>
<td>FloweringPlant</td>
<td>205</td>
</tr>
</tbody>
</table>
5.2 Evaluation of Preference Extraction

In the following, we will discuss the results for the selectional preference extraction of direct objects of ‘lire’ (‘read’) and ‘manger’ (‘eat’). These words were chosen because we believe them to show strong selectional preferences as far as their direct objects are concerned. Thus they may serve as proof-of-concept cases for our approach.

The selectional preference strength $S_{obj}(lire)$, i.e. the preference strength of ‘lire’ wrt. to the direct object relation (cf. Section 2.2) is 1.37296, whereas $S_{obj}(manger)$ is 3.46397. This means that ‘manger’ generally shows a stronger preference wrt. to its direct object than ‘lire’. The effect of this is that if ‘lire’ shows a preference for a particular SUMO class, this preference will weigh more than a preference of ‘manger’, since it is generally weaker wrt. preferential behaviour. This is due to the fact that the value of the selectional association between a predicate $p$ and the class $c$ of its argument (cf. Section 2.2 above) is normalised by the selectional preference strength of $p$.

Table 4 shows that ‘lire’ shows a strong preference for objects of type “Text”, and further preferences for “ContentBearingPhysical” and “LinguisticExpression”. After these three items, the figures indicate a bigger gap to the next entity. As far as ‘manger’ is concerned, it shows a very strong preference for direct objects of type “Food”, with the second best class (“SelfConnectedObject”) reaching just over half of the score for “Food”. Looking at these results, it is fair to say that they do match our intuitions.

6 Conclusion

We have presented a generic method for mapping EuroWordNets to the Suggested Upper Merged Ontology and have shown its application to the French EuroWordNet. The mapping procedure builds on existing work on SUMO and version 1.6 of Princeton WordNet [8], EuroWordNet’s Inter-Lingual-Index [7] and WordNet’s sensemap files. The resulting mapping was used in the calculation of selectional preferences of French verbal predicates with respect to nominal arguments. Preference extraction within an experimental setup shows promising results for the French verbs ‘manger’ (‘eat’) and ‘lire’ (‘read’).

In the future, we intend to carry out a qualitative evaluation on a larger scale, both for the mapping procedure (cf. Section 3.2) and the extraction of selectional preferences. The ultimate goal is to use the extracted selectional preferences for word sense disambiguation of verbal predicates as well as their arguments, and we will work on this in the near future. Moreover, we plan to consider extracting pairs of subjects and objects in order to calculate preferences of a direct object given the subject and vice versa. Finally, it would be interesting to see how the mapping methodology performs when applied to

---

11 The frequency indicates the number of synsets which have been mapped directly onto the respective SUMO class, so no accumulation of frequency counts along the SUMO hierarchy was made, since that would, of course, leave the top 20 slots in the table to the top 20 nodes in the hierarchy. A synset such as 00058624-n (cf. examples above), which has been mapped onto three different SUMO classes, counts for each of these classes.
EuroWordNets other than French, provided that they are linked to the Inter-Lingual-Index as well. We do, however, expect our methodology to be generic enough to be applied to other languages without any major issues.

Table 4. Selectional preferences of 'lire' ('read') and 'manger' ('eat') wrt. direct objects

<table>
<thead>
<tr>
<th>SUMO concept c</th>
<th>$A_{obj}(\text{lire}, c)$</th>
<th>SUMO concept c</th>
<th>$A_{obj}(\text{manger}, c)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>0.3868</td>
<td>Food</td>
<td>0.2179</td>
</tr>
<tr>
<td>ContentBearingPhysical</td>
<td>0.2548</td>
<td>SelfConnectedObject</td>
<td>0.1253</td>
</tr>
<tr>
<td>LinguisticExpression</td>
<td>0.2431</td>
<td>NonFullyFormed</td>
<td>0.0779</td>
</tr>
<tr>
<td>Disseminating</td>
<td>0.1259</td>
<td>Object</td>
<td>0.0745</td>
</tr>
<tr>
<td>Communication</td>
<td>0.1083</td>
<td>DevelopmentalAttribute</td>
<td>0.0662</td>
</tr>
<tr>
<td>Stating</td>
<td>0.0840</td>
<td>Meat</td>
<td>0.0599</td>
</tr>
<tr>
<td>Word</td>
<td>0.0611</td>
<td>Animal</td>
<td>0.0453</td>
</tr>
<tr>
<td>Noun</td>
<td>0.0601</td>
<td>Organism</td>
<td>0.0403</td>
</tr>
<tr>
<td>Artifact</td>
<td>0.0582</td>
<td>OrganicObject</td>
<td>0.0402</td>
</tr>
<tr>
<td>ContentBearingProcess</td>
<td>0.0541</td>
<td>Vertebrate</td>
<td>0.0382</td>
</tr>
<tr>
<td>OccupationalRole</td>
<td>0.0498</td>
<td>FruitOrVegetable</td>
<td>0.0368</td>
</tr>
<tr>
<td>LinguisticCommunication</td>
<td>0.0409</td>
<td>WarmBloodedVertebrate</td>
<td>0.0295</td>
</tr>
<tr>
<td>CorpuscularObject</td>
<td>0.0377</td>
<td>Arachnid</td>
<td>0.0258</td>
</tr>
<tr>
<td>name</td>
<td>0.0368</td>
<td>Monkey</td>
<td>0.0239</td>
</tr>
<tr>
<td>Book</td>
<td>0.0311</td>
<td>BodySubstance</td>
<td>0.0222</td>
</tr>
<tr>
<td>Proposition</td>
<td>0.0295</td>
<td>Mammal</td>
<td>0.0216</td>
</tr>
<tr>
<td>SelfConnectedObject</td>
<td>0.0285</td>
<td>AnatomicalStructure</td>
<td>0.0201</td>
</tr>
<tr>
<td>Physical</td>
<td>0.0225</td>
<td>Fish</td>
<td>0.0190</td>
</tr>
<tr>
<td>FamilyGroup</td>
<td>0.0218</td>
<td>CorpuscularObject</td>
<td>0.0187</td>
</tr>
<tr>
<td>destination</td>
<td>0.0156</td>
<td>PlantAnatomicalStructure</td>
<td>0.0184</td>
</tr>
</tbody>
</table>

Acknowledgements

The research described in this work has been carried out as part of the project 'Polysemy in a Conceptual System' (project B5 of SFB 732) and was funded by grants from the German Research Foundation. I should like to thank Adam Pease, Achim Stein, Piek Vossen, Sabine Schulte im Walde and Christian Hying for their valuable comments and suggestions at the outset of this work, as well as the two anonymous reviewers for helping to improve the structure and content of the paper.

References

440 Dennis Spohr

ington, DC (1997) 52–57
Romanian WordNet: Current State, New Applications and Prospects

Dan Tufiș, Radu Ion, Luigi Bozianu, Alexandru Ceaușu, and Dan Ştefănescu

Romanian Academy Research Institute for Artificial Intelligence
13, Calea 13 Septembrie, 050711, Bucharest 5, Romania
{tufis, radu, bozi, aceausu, danstef}@racai.ro

1 Introduction

The development of the Romanian WordNet began in 2001 within the framework of the European project BalkaNet which aimed at building core WordNets for 5 new Balkan languages: Bulgarian, Greek, Romanian, Serbian and Turkish. The philosophy of the BalkaNet architecture was similar to EuroWordNet [1, 2]. As in EuroWordNet, in BalkaNet the concepts considered highly relevant for the Balkan languages (and not only) were identified and called BalkaNet Base Concepts. These are classified in three increasing size sets (BCS1, BCS2 and BCS3). Altogether BCS1, BCS2 and BCS3 contain 8516 concepts that were lexicalized in each of the BalkaNet WordNets.

The monolingual WordNets had to have their synsets aligned to the translation equivalent synsets of the Princeton WordNet (PWN). The BCS1, BCS2 and BCS3 were adopted as core WordNets for several other WordNet projects such as Hungarian [3], Slovene [4], Arabic [5, 6], and many others.

At the end of the BalkaNet project (August 2004) the Romanian WordNet, contained almost 18,000 synsets, conceptually aligned to Princeton WordNet 2.0 and through it to the synsets of all the BalkaNet WordNets. In [7], a detailed account on the status of the core Ro-WordNet is given as well as on the tools we used for its development.

After the BalkaNet project ended, as many other project partners did, we continued to update the Romanian WordNet and here we describe its latest developments and a few of the projects in which Ro-WordNet, Princeton WordNet or some of its BalkaNet companions were of crucial importance.

2 The Ongoing Ro-WordNet Project and its Current Status

The Ro-WordNet is a continuous effort going on for 6 years now and likely to continue several years from now on. However, due to the development methodology adopted in BalkaNet project, the intermediate WordNets could be used in various other projects (word sense disambiguation, word alignment, bilingual lexical knowledge acquisition, multilingual collocation extraction, cross-lingual question answering, machine translation etc.).
Recently we started the development of an English-Romanian MT system for the legalese language of the type contained in JRC-Acquis multilingual parallel corpus [8] and of a cross-lingual question answering system in open domains [9, 10]. For these projects, heavily relying on the aligned Ro-En WordNets, we extracted a series of high frequency Romanian nouns and verbs not present in Ro-WordNet but occurring in JRC-Acquis corpus and in the Romanian pages of Wikipedia and proceeded at their incorporation in Ro-WordNet. The methodology and tools were essentially the same as described in [11], except that the dictionaries embedded into the WNBuilder and WNCorrect were significantly enlarged.

The two basic development principles of the BalkaNet methodology, that is Hierarchy Preservation Principle (HPP) and Conceptual Density Principle (CDP), were strictly observed. For the sake of self-containment, we restate them here.

**Hierarchy Preservation Principle**

If in the hierarchy of the language L1 the synset \( M_2 \) is a hyponym of synset \( M_1 \) 
\((M_2 \text{ H}^m \text{ M}_1)\) and the translation equivalents in L2 for \( M_1 \) and \( M_2 \) are \( N_1 \) and \( N_2 \) respectively, then in the hierarchy of the language L2 \( N_2 \) should be a hyponym of synset \( N_1 \) \((N_2 \text{ H}^n \text{ N}_1)\). Here \( H^m \) and \( H^n \) represent a chain of \( m \) and \( n \) hierarchical relations between the respective synsets (hyperonymy relations composition).

**Conceptual Density Principle (noun and verb synsets)**

Once a nominal or verbal concept (i.e. an ILI concept that in PWN is realized as a synset of nouns or as a synset of verbs) was selected to be included in Ro-WordNet, all its direct and indirect ancestors (i.e. all ILI concepts corresponding to the PWN synsets, up to the top of the hierarchies) should be also included in Ro-WordNet.

By observing HPP, the lexicographers were relieved of the task to establish the semantic relations for the synsets of the Ro-WordNet. The hypernym relations as well as the other semantic relations were imported automatically from the PWN. The CDP compliance ensures that no dangling synsets, harmful in taxonomic reasoning, are created.

The tables below give a quantitative summary of the Romanian WordNet at the time of writing (September, 2007). As these statistics are changing every month, the updated information should be checked at http://nlp.racai.ro/Ro-wordnet.statistics. The Ro-wordnet is currently mapped on the various versions of Princeton WordNet: PWN1.7.1, PWN2.0 and PWN2.1. The mapping onto the last version PWN3.0 is also considered. However, all our current projects are based on the PWN2.0 mapping and in the following, if not stated otherwise, by PWN we will mean PWN2.0.
Table 1. POS distribution of the synsets in the Romanian WordNet.

<table>
<thead>
<tr>
<th>Noun synsets</th>
<th>Verb synsets</th>
<th>Adj. synsets</th>
<th>Adv. synsets</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>33151</td>
<td>8929</td>
<td>851</td>
<td>834</td>
<td>43765</td>
</tr>
</tbody>
</table>

Table 2. Internal relations used in the Romanian WordNet.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Count</th>
<th>Relation</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>hypernym</td>
<td>42794</td>
<td>category_domain</td>
<td>2668</td>
</tr>
<tr>
<td>near_anonym</td>
<td>2438</td>
<td>also_see</td>
<td>586</td>
</tr>
<tr>
<td>holo_part</td>
<td>3531</td>
<td>subevent</td>
<td>335</td>
</tr>
<tr>
<td>similar_to</td>
<td>899</td>
<td>holo_portion</td>
<td>327</td>
</tr>
<tr>
<td>verb_group</td>
<td>1404</td>
<td>causes</td>
<td>171</td>
</tr>
<tr>
<td>holo_member</td>
<td>1300</td>
<td>be_in_state</td>
<td>570</td>
</tr>
<tr>
<td>DOMAINS classes</td>
<td>165</td>
<td>SUMO&amp;MILO categories</td>
<td>1836</td>
</tr>
<tr>
<td>objective synsets</td>
<td>34164</td>
<td>subjective synsets</td>
<td>9601</td>
</tr>
</tbody>
</table>

As one can see from Table 2, the synsets in Ro-WordNet have attached, via PWN, DOMAINS-3.1 [12], SUMO&MILO [13, 14] and SentiWordNet [15] labels.

The DOMAINS labeling (http://wndomains.itc.it/) uses Dewey Decimal Classification codes and the 115425 PWN synsets are classified into 168 distinct classes (domains).

The SUMO&MILO upper and mid level ontology is the largest freely available (http://www.ontologyportal.org/) ontology today. It is accompanied by more than 20 domain ontologies and altogether they contain about 20,000 concepts and 60,000 axioms. They are formally defined and do not depend on a particular application. Its attractiveness for the NLP community comes from the fact that SUMO, MILO and associated domain ontologies were mapped onto Princeton WordNet. SUMO and MILO contain 1107 and respectively 1582 concepts. Out of these, 844 SUMO concepts and 1582 MILO concepts were used to label almost all the synsets in PWN. Additionally, 215 concepts from some specific domain ontology were used to label the rest of synsets in PWN (instances).

The SentiWordnet [15] adds subjectivity annotations to the PWN synsets. Their basic assumptions were that words have graded polarities along the Subjective-Objective (SO) & Positive-Negative (PN) orthogonal axes and that the SO and PN polarities depend on the various senses of a given word (context). The word senses in a synset are associated with a triple P (positive subjectivity), N (negative subjectivity) and O (objective) so that the values of these attributes sum up to 1. For instance, the sense 2 of the word nightmare (a terrifying or deeply upsetting dream) is marked-up by the values P:0.0, N:0.25 and O:0.75, signifying that the word denotes to a large extent an objective thing with a definite negative subjective polarity.

Due to the BalkaNet methodology adopted for the monolingual WordNets development, most of the DOMAINS, SUMO and MILO conceptual labels in PWN are represented in our Ro-WordNet (see Table 3).
Table 3. The ontological labeling (DOMAINS, SUMO, MILO, etc.) in RO-WordNet vs. PWN.

<table>
<thead>
<tr>
<th>LABELS</th>
<th>PWN</th>
<th>Ro-WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOMAINS-3.1</td>
<td>168</td>
<td>165</td>
</tr>
<tr>
<td>SUMO</td>
<td>844</td>
<td>781</td>
</tr>
<tr>
<td>MILO</td>
<td>949</td>
<td>882</td>
</tr>
<tr>
<td>Domain ontologies</td>
<td>215</td>
<td>173</td>
</tr>
</tbody>
</table>

The BalkaNet compliant XML encoding of a synset, including the new subjectivity annotations is exemplified in Figure 1.

```
<SYNSET>
  <ID>ENG20-05435872-n</ID>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>nightmare</LITERAL>
  </SYNONYM>
  <DEF>a terrifying or deeply upsetting dream</DEF>
  <SUMO>PsychologicalProcess</SUMO>
  <DOMAIN>factotum</DOMAIN>
  <SENTIWN><P>0.0</P><N>0.25</N><O>0.75</O></SENTIWN>
</SYNSET>
```

Fig. 1. Encoding of two EQ-synonyms synsets in PWN and Ro-WordNet.

The visualization of the synsets in Figure 1, by means of the VISDIC editor[^1][16], is shown in Figure 2.

[^1]: http://nlp.fi.muni.cz/projekty/visdic/
The Ro-WordNet can be browsed by a web interface implemented in our language web services platform (see Figure 3). Although currently only browsing is implemented, Ro-WordNet web service will, later on, include search facilities accessible via standard web services technologies (SOAP/WSDL/UDDI) such as distance between two word senses, translation equivalents for one or more senses, semantically related word-senses, etc.

### 3 Recent applications of the Ro-WordNet

In previous papers [17, 18] we demonstrated that difficult processes such as word sense disambiguation and word alignment of parallel corpora can reach very high accuracy if one has at his/her disposal aligned WordNets. Various other researchers showed the invaluable support of aligned WordNets in improving the quality of machine translation. In this section we will discuss some new applications of Ro-En pair of WordNets, the performances of which strongly argue for the need to keep-up the Ro-WordNet development endeavor.
3.1 WordNet as an important resource to monolingual WSD

The WordNet concept has practically revolutionized the way a WSD application is thought. The explicit semantic structure of WordNet enables WSD application writers to use the semantic relations between synsets as a way of primitive reasoning when establishing senses of the words in a text. The presence of a particular semantic relation called hypernymy has also provided the much-expected mechanism of generalization of words’ senses allowing for deploying of machine learning methods to WSD. It can be safely stated that WordNet has pushed forward the very nature of WSD algorithms in the direction of true semantic processing.

In [19] we presented an unsupervised WSD algorithm whose disambiguation philosophy is entirely based on the WordNet architecture. The idea of the algorithm is that of combining the paradigmatic information provided by the WordNet with the contextual information of the word in both the training and the disambiguation phases. The context of the word is given by its dependency relations with the neighboring words (which are not necessarily adjacent). In [20] we introduced the concept of meaning attraction model as theoretical basis for our monolingual WSD algorithm.

In the training phase, we estimate the measure of the meaning attraction between dependency-related words of a sentence. Given two dependency-related words, $W_i$
and $W_a$, each with its associated WordNet synset identifiers, the meaning attraction between synset id $i$ of word $W_a$ and synset id $j$ of word $W_b$ is a function of the frequency counts between pairs $<i,j>$, $<i,*>$ and $<*,j>$ collected from the entire training corpus. As meaning attraction functions, we chose DICE, Log Likelihood and Pointwise Mutual Information which can be all computed given the pair frequencies described above. Consider for instance the examples in Figure 4.

Both “recommended” and “suggested” are in the same synset with the id 00071572. Also “class” and “course” are in the same synset with the id 00831838. This means that the pair $<00071572,00831838>$ receives count 2 from these two examples as opposed to any other pair from the cartesian products which is seen only once. This fact translates to a preference for the meaning association “mentioned as worthy of acceptance” and “education imparted in a series of lessons or class meetings” which may not be correct in all contexts but it is a part of the natural learning bias of the training algorithm.

The synonymy lexical-semantic relation is just the first way of generalization in the learning phase. Another one, more powerful is given by the semantic relations graph encoded in WordNet. In Figure 4 we have simply used the synsets’ ids for computing frequencies. But their number is far too big to give us reliable counts. So, we are making use of the hypernym hierarchies for nouns and verbs to generalize the meanings that are learned but without introducing ambiguities. So, for a given synset id we select the uppermost hypernym that only subsumes one meaning of the word.

This WSD algorithm has recently participated in the 4th Semantic Evaluation Forum, Semeval 2007 on the English All-Words Coarse and Fine-Grained tasks where it attained the top performance among the unsupervised systems.

Because it is language independent, it also has been applied with encouraging results to Romanian using the Romanian WordNet. The test corpus was the Romanian SemCor, a controlled translation of the English version of the corpus. The test set comprised of 48392 meaning annotated content word occurrences and for different meaning attraction functions and combinations of results using them, the best F-measure was 59.269%.

---

2 By synset identifier, we understand the offset of the synset in the WordNet database. Knowing this ID and the word, we can extract the sense number of that word in the respective synset.
3.2 Romanian WordNet and Cross-Language QA

Romanian WordNet and its translation equivalence with the Princeton WordNet have been used as a general-purpose translation lexicon in the CLEF 2006 Romanian to English question answering track [9]. The task required asking questions in Romanian and finding the answer from an English text collection. For this task, the question analysis (focus/topic identification, answer type, keywords detection, query formulation, etc.) was made in Romanian and the rest of the process (text searching and answer extraction) was made in English.

Our approach was to generate the query for the text-searching engine in Romanian and then to translate every key element of the query (topic, focus, keywords) into English without modifying the query. Since we don’t have a Romanian to English translation system and because neither the question nor the text collection were word sense disambiguated, for every key element of the query, we selected all the synonyms in which it appeared from the Romanian WordNet. Then, for every synonym of the latter list, we extracted all English literals of the corresponding English synset making a list of all possible translation equivalents for the source Romanian word. Finally, we ordered this list by the frequency of its elements computed from the English text collection and selected the first 3 elements as translation equivalents of the Romanian word. While this translation method does not assure a correct translation of each source Romanian word of the initial question, it is good enough for the search engine to return a set of documents in which the correct answer would be eventually identified. The evaluation of the recall for the IR part of the QA system [10] was close to 80%, but its major drawback was not in the translation part but in the identification of the keywords subject to translation.

3.3 Machine Translation Development Kit

The aligned Ro-En WordNets have been incorporated into our MT development kit, which comprises tokenization, tagging, chunking, dependency linking, word alignment and WSD based on the word alignment and the respective WordNets. The interface of the MTKit platform allows for editing the word alignment, word sense disambiguation, importing annotations from one language to the other and a friendly visualization of all the preprocessing steps in both languages. Figure 5 illustrates a snapshot from the MTKit interface. One can see the word alignment of translation unit (no. 16) from a document (nou-jrc42002595) contained in the JRC-Acquis multilingual parallel corpus. The right-hand windows display the morpho-lexical information attached to selected word from the central window (journal). The upper-right window displays the POS-tag, the lemma, the orthographic form and the WordNet sense number. The windows below it display the WordNet relevant information as well as the SUMO&MILO label pertaining to the corresponding sense number. The lowest-right window displays the appropriate WordNet gloss and SUMO documentation.
3.4 Opinion analysis

One of the hottest research topics nowadays is related to subjectivity web mining with many applications in opinionated question answering, product review analysis, personal and institutional decision making, etc. Recent release of SentiWordNet [15] allowed for automatic import of the subjectivity annotations from PWN into any WordNet aligned with it. Thus, it became possible to develop subjectivity analysis programs for various languages equipped with a WordNet aligned to PWN.

We made some preliminary experiments with a naive opinion sentence classifier [21]. It simply sums up the O, P and N scores for each word in a sentence. For the words in the chunks immediately following a valence shifter, until the next valence shifter, the O, P and N scores are modified so that the new values are the following: $O_{new} = 1 - O_{old}$, $P_{new} = P_{old} \cdot O_{old}/(P_{old} + N_{old})$ and $N_{new} = N_{old} \cdot O_{old}/(P_{old} + N_{old})$. Taking advantage of the 1984 Romanian-English parallel corpus which is word aligned and word sense disambiguated in both languages, we applied our naive opinion sentence classifier on the English original sentences and their Romanian translations and OpinionFinder [22] on the English original sentences. Since the WordNet opinion annotations are the same in PWN and Ro-WordNet aligned synsets it was obvious that our opinion classifier would give similar results for the two languages. So, in the end we compared the classifications made by our opinion classifier and OpinionFinder for a few English sentences. From the total number of 6411 sentences in 1984 corpus there were selected 954 for which both internal classifiers of Opinion Finder agreed in
judging the respective sentences as being subjective (see for details [22]) as in the sentence below:

\[\textit{MPQASENT autoclass1="subj" autoclass2="subj" diff="30.8">The stuff was like nitric acid, and moreover, in swallowing it one had the sensation of being hit on the back of the head with a rubber club.</textit{MPQASENT}>\]

We manually analyzed the 20 top-certainty sentences from the 954 selected ones, extracted the valence shifters they contained and dry-run the naive classifier described above, using the subjectivity values from SentiWordNet. When the O value for a sentence was smaller than 0.5 we arbitrarily decided that it was subjective. All the 20 sentences were thus classified as subjective. For the same sentence\(^3\), chunked and WSDed as below, the naive opinion classifier computed the following scores:
P:0.063; N:0.563; O:0.375.

\[[\text{The stuff(1)} \text{ was(1) like (3) [nitric_acid(1)] , and moreover(1) , [in swallowing (1) it] one had(1) the sensation (1) of [being hit (4) [on the back_of_the_head (1)] [with a rubber(1) club(3)].}\]

Whether the threshold value or the final P, N and O values might be debatable, the main idea here is that one can use the SentiWordNet annotation of the synsets in a WordNet for language L, aligned to PWN, for dwelling on subjectivity mining in arbitrary texts in language L.

\section{Conclusions and further work}

The development of Ro-WordNet is a continuous project, trying to keep up with the new updates of the Princeton WordNet. The increase in its coverage is steady (approximately 10,000 synsets per year for the last three years) with the choice for the new synsets imposed by the applications built on the basis of Ro-WordNet. Since PWN was aimed to cover general language, it is very likely that specific domain applications would require terms not covered by Princeton WordNet. In such cases, if available, several multilingual thesauri (EUROVOC - http://europa.eu/eurovoc/, IATE - http://iate.europa.eu/iatediff/about_IATE.html, etc.) can complement the use of WordNets. Besides further augmenting the Ro-WordNet, we plan the development of an environment where various multilingual aligned lexical resources (WordNets, framenets, thesauri, parallel corpora) could be used in a consistent but transparent way for a multitude of multilingual applications.

\(^3\) The underlined words represent valence shifters, and the square parentheses delimit chunks as determined by our chunker; the numbers following the words represent their PWN sense number.
Acknowledgements

The work reported here was supported by the Romanian Academy program “Multilingual Acquisition and Use of Lexical Knowledge”, the ROTEI project (CEEX No. 29-E136-2005) and by the SIR-RESDEC project (PNCDI2, 4th Programme, No. D1.1-0.0.7), the last two granted by the National Authority for Scientific Research. We are grateful to many colleagues who contributed or continue to contribute the development of Ro-WordNet with special mentions for Cătălin Mihăilă, Magaretá Manu Magda and Verginica Mititelu.

References

Enriching WordNet with Folk Knowledge and Stereotypes

Tony Veale\(^1\) and Yanfen Hao\(^1\)

\(^1\) School of Computer Science and Informatics, University College Dublin, Dublin, Ireland
{Tony.Veale, Yanfen.Hao}@UCD.ie

Abstract. The knowledge that is needed to understand everyday language is not necessarily the knowledge one finds in an encyclopedia or dictionary. Much of this is “folk” knowledge, based on stereotypes and culturally-inherited associations that do not hold in all situations, or which may, strictly speaking, be false. We can open a linguistic window onto this knowledge through simile, since explicit similes make use of highly evocative and inference-rich concepts to ground comparisons and make the unfamiliar seem familiar. This paper describes a means of enriching WordNet with commonly ascribed cultural properties by mining explicit similes of the form “as ADJ as a NOUN” from the internet. We also show how these properties can be leveraged, through further web search, into rich frame structures for the most evocative WordNet concepts.

Keywords: simile, folk knowledge, frame representation.

1 Introduction

Many of the beliefs that one uses to reason about everyday entities and events are neither strictly true or even logically consistent. Rather, people appear to rely on a large body of folk knowledge in the form of stereotypes, clichés and other prototype-centric structures (e.g., see [1]). These prototypes comprise the landmarks of our conceptual space against which other, less familiar concepts can be compared and defined. For instance, people readily employ the animal concepts Snake, Bear, Bull, Wolf, Gorilla and Shark in everyday conversation without ever having had first-hand experience of these entities. Nonetheless, our culture equips us with enough folk knowledge of these highly evocative concepts to use them as dense short-hands for all manner of behaviours and property complexes. Snakes, for example, embody the notions of treachery, slipperiness, cunning and charm (as well as a host of other, related properties) in a single, visually-charged package. To compare someone to a snake is to suggest that many of these properties are present in that person, and thus, one would do well to treat that person as one would treat a real snake.

Descriptors like “snake”, “shark” and “wolf” find a great deal of traction in everyday conversation because they are “dense descriptors” – they convey a great deal of useful information in a simple and concise way. The information imparted is open-ended, so that a listener may take meaning X from the description when it is initially used (e.g., that a given person is treacherous) and meaning X+Y (e.g., that
this person is both treacherous and charming) in a later, more informed context. But the information imparted is rarely of the kind one finds in a dictionary or encyclopaedia, or in a resource like WordNet [2], because it is neither contributes to the definition of the given concept or is actually true of that concept. Insofar as WordNet is used to make sense of real texts by real, culturally-grounded speakers, it can be enriched considerably by the addition of such stereotypical knowledge. But where can this knowledge be found and exploited?

In “A Christmas Carol”, Dickens [3] notes that “the wisdom of our ancestors is in the simile; and my unhallowed hands shall not disturb it, or the Country’s done for” (chapter 1, page 1). In other words, folk knowledge is passed down through a culture via language, most often in specific linguistic forms. The simile, as noted by Dickens, is one common vehicle for folk wisdom, one that uses explicit syntactic means (unlike metaphor; see [4]) to mark out those concepts that are most useful as landmarks for linguistic description. Similes do not always convey truths that are universally true, or indeed, even literally true (e.g., bowling balls are not literally bald). Rather, similes hinge on properties that are possessed by prototypical or stereotypical members of a category (see [5]), even if most members of the category do not also possess them. As a source of knowledge, similes combine received wisdom, prejudice and oversimplifying idealism in equal measure. As such, similes reveal knowledge that is pragmatically useful but of a kind that one is unlikely to ever acquire from a dictionary (or, indeed, from WordNet). Although a simpler rhetorical device than metaphor, we have much to learn about language and its underlying conceptual structure by a comprehensive study of real similes in the wild (see [6]), not least about the recurring vehicle categories that signpost this space (see [7]).

In this paper we describe a means through which we can enrich WordNet with stereotypical folk-knowledge from similes that are mined from the text of the world-wide web. We describe the Google-based mining process in section 2, before describing how the acquired knowledge is sense-linked to WordNet in section 3. In section 4 we describe on-going work to elaborate this property-rich knowledge into more complex frame-representations, before providing an empirical evaluation of the basic properties in section 5. The paper concludes with thoughts on future work in section 6.

2 Acquiring Knowledge from Simile

As in the study reported in [6], we employ the Google search engine as a retrieval mechanism for accessing relevant web content. However, the scale of the current exploration requires that retrieval of similes be fully automated, and this automation is facilitated both by the Google API and its support for the wildcard term *. In essence, we consider here only partial explicit similes conforming to the pattern “as ADJ as a/an NOUN”, in an attempt to collect all of the salient values of ADJ for a given value of NOUN. We do not expect to identify and retrieve all similes mentioned on the world-wide-web, but to gather a large, representative sample of the most commonly used.
To do this, we first extract a list of antonymous adjectives, such as “hot” or “cold”, from WordNet [2], the intuition being that explicit similes will tend to exploit properties that occupy an exemplary point on a scale. For every adjective ADJ on this list, we send the query “as ADJ as *” to Google and scan the first 200 snippets returned for different noun values for the wildcard *. From each set of snippets we can ascertain the relative frequencies of different noun values for ADJ. The complete set of nouns extracted in this way is then used to drive a second phase of the search. In this phase, the query “as * as a NOUN” is used to collect similes that may have lain beyond the 200-snippet horizon of the original search, or that hinge on adjectives not included on the original list. Together, both phases collect a wide-ranging series of core samples (of 200 hits each) from across the web, yielding a set of 74,704 simile instances (of 42,618 unique types) relating 3769 different adjectives to 9286 different nouns.

2.1 Simile Annotation

Many of these similes are not sufficiently well-formed for our purposes. In some cases, the noun value forms part of a larger noun phrase: it may be the modifier of a compound noun (as in “bread lover”), or the head of complex noun phrase (such as “gang of thieves”). In the former case, the compound is used if it corresponds to a compound term in WordNet and thus constitutes a single lexical unit; if not, or if the latter case, the simile is rejected. Other similes are simply too contextual or under-specified to function well in a null context, so if one must read the original document to make sense of the simile, it is rejected. More surprisingly, perhaps, a substantial number of the retrieved similes are ironic, in which the literal meaning of the simile is contrary to the meaning dictated by common sense. For instance, “as hairy as a bowling ball” (found once) is an ironic way of saying “as hairless as a bowling ball” (also found just once). Many ironies can only be recognized using world (as opposed to word) knowledge, such as “as sober as a Kennedy” and “as tanned as an Irishman”. In addition, some similes hinge on a new, humorous sense of the adjective, as in “as fruitless as a butcher-shop” (since the latter contains no fruits) and “as pointless as a beach-ball” (since the latter has no points).

Given the creativity involved in these constructions, one cannot imagine a reliable automatic filter to safely identify bona-fide similes. For this reason, the filtering task was performed by human judges, who annotated 30,991 of these simile instances (for 12,259 unique adjective/noun pairings) as non-ironic and meaningful in a null context; these similes relate a set of 2635 adjectives to a set of 4061 different nouns. In addition, the judges also annotated 4685 simile instances (of 2798 types) as ironic; these similes relate 936 adjectives to a set of 1417 nouns. Perhaps surprisingly, ironic pairings account for over 13% of all annotated simile instances and over 20% of all annotated simile types.
3 Establishing Links to WordNet

It is important to know which sense of a noun is described by a simile if an accurate conceptual picture is to be constructed. For instance, “as stiff as a zombie” might refer either to a reanimated corpse or to an alcoholic cocktail (both are senses of “zombie” in WordNet, and drinks can be “stiff” too). Sense disambiguation is especially important if we hope to derive meaningful correlations from property co-occurrences; for instance, zombies are described in web similes as exemplars of not just stiffness, but of coldness, slowness and emotionlessness. If such co-occurrences are observed often enough, a cognitive agent might usefully infer a causal relationship among pairs of properties.

Disambiguation is trivial for nouns with just a single sense in WordNet. For nouns with two or more fine-grained senses that are all taxonomically close, such as “gladiator” (two senses: a boxer and a combatant), we consider each sense to be a suitable target. In some cases, the WordNet gloss for a particular sense will actually mention the adjective of the simile, and so this sense is chosen. In all other cases, we employ a strategy of mutual disambiguation to relate the noun vehicle in each simile to a specific sense in WordNet. Two similes “as ADJ as NOUN₁” and “as ADJ as NOUN₂” are mutually disambiguating if NOUN₁ and NOUN₂ are synonyms in WordNet, or if some sense of NOUN₁ is a hypernym or hyponym of some sense of NOUN₂ in WordNet. For instance, the adjective “scary” is used to describe both the noun “rattler” and the noun “rattlesnake” in bona-fide (non-ironic) similes; since these nouns share a sense, we can assume that the intended sense of “rattler” is that of a dangerous snake rather than a child’s toy. Similarly, the adjective “brittle” is used to describe both saltines and crackers, suggesting that it is the bread sense of “cracker” rather than the hacker, firework or hillbilly senses (all in WordNet) that is intended.

These heuristics allow us to automatically disambiguate 10,378 bona-fide simile types (85% of those annotated), yielding a mapping of 2124 adjectives to 3778 different WordNet senses. Likewise, 77% (or 2164) of the simile types annotated as ironic are disambiguated automatically. A remarkable stability is observed in the alignment of noun vehicles to WordNet senses: 100% of the ironic vehicles always denote the same sense, no matter the adjective involved, while 96% of bona-fide vehicles always denote the same sense. This stability suggests two conclusions: the disambiguation process is consistent and accurate; but more intriguingly, only one coarse-grained sense of any word is likely to be sufficiently exemplary of some property to be useful as a simile vehicle.

4 Acquiring Frame Representations

Each bona-fide simile contributes a different salient property to the representation of a vehicle concept. In our data, one half (49%) of all bona-fide vehicle nouns occur in two or more similes, while one third occur in three or more and one fifth occur in four or more. The most frequently used figurative vehicles can have many more; “snowflake”, for instance, is ascribed over 30 in our database, including: white, pure,
fresh, beautiful, natural, intricate, delicate, identifiable, fragile, light, dainty, frail, weak, sweet, precious, quiet, cold, soft, clean, detailed, fleeting, unique, singular, distinctive and lacy.

Because the same adjectival properties are associated with multiple vehicles, the resulting property graph allows different vehicles to be perceived as similar by virtue of these shared properties. For instance, Ninja and Mime are deemed similar by virtue of the shared property silent, while Artist and Surgeon are similar by virtue of the properties skilled, sensitive and delicate. Nonetheless, it can be claimed the property level is simply too shallow to allow for nuanced similarity judgements. For instance, are ninjas and mimes silent in the same way? Both surgeons and bloodhounds are prototypes of sensitivity, but the former has sensitive hands while the latter has a sensitive nose. To put these properties in context, we need to know the specific facet of each concept that is modified, so that sensible comparisons can be made. In effect, we need to move from a simple property-ascription representation to a richer, frame:slot:filler representation. In such a scheme, the property sensitive is a typical filler for the hands slot of Surgeon and the nose slot of Bloodhound, thereby disallowing any mis-matched comparisons.

This process of frame construction can also be largely automated via targeted web-search. For every bona-fide simile-type “as ADJ as a Noun_vehicle” (all 10,378 of them that have been WordNet-linked in section 3), we automatically generate the web-query “the ADJ * of a Noun_vehicle” and harvest the top 200 results from Google. From these snippets, we then extract all noun values of the wildcard *. In many cases, these noun values are precisely the conceptual facets we desire for a culturally-accurate and nuanced representation, ranging from hands for Surgeon to roar for Lion to eye for Hawk. The frequency of these values also allows us to create a textured representation for each concept, so that e.g., both hands and eye are notable facets for surgeon, but the latter is higher ranked. However, this web-pattern also yields a non-trivial amount of noise: while “the proud strut of a peacock” is very revealing about the concept Peacock, the snippet “the proud owner of a peacock” is not. Quite simply, we seek to fill intrinsic facets of a concept like hands, eye, gait and strut that contribute to the folk definition of the concept, while ignoring extrinsic and contingent facets such as owner, husband, brother and so on.

One can look to specific abstractions in WordNet – such as {trait} – to serve as a filter on the facet-nouns that are extracted, but such a simple filter would be unduly coarse. Instead, we consider all facet-nouns, but generalize the WordNet vehicle-senses to which they are attached, to create a high-level mapping of vehicle types (such as Person, Animal, Implement, Substance, etc.) to facets (such as hands, eye, sparkle, father, etc.). This high-level (and considerably more compressed) map is then human-edited, to remove any facets that are unrevealing or simply appropriate for the WordNet vehicle type. In this editing process (which requires about one man-day), contingent facets such as father, wife, etc. are quickly identified and removed.
Fig. 1. The acquired Frame:slot:filler representations for Peacock and Lion.

As can be seen in the examples of Lion and Peacock in Figure 1, the slot:filler pairs that are acquired for each concept do indeed reflect the most relevant cultural associations for these concepts. Moreover, there is a great deal of anthropomorphic rationalization of an almost poetic nature about these representations, of the kind that is instantly recognizable to native speakers of a language but which one would be hard pressed to find in a conventional dictionary (except insofar as some lexical concepts may give rise to additional word senses, such as “peacock” for a proud and flashily dressed person).

Overall, frame representations of this kind are acquired for 2218 different WordNet noun senses, yielding a combined total of 16,960 slot:filler pairings (or an average of 8 slot:filler pairs per frame). As the examples of Figure 1 demonstrate, these frames provide a level of representational finesse that greatly enriches the basic property descriptions yielded by similes alone. To answer an earlier question then, mimes and ninjas are now similar by virtue of each possessing the slot:filler Has_art: silent. But as this and other examples suggest, the introduction of finely discriminating frame structures can decrease a system’s ability to recognize similarity, if comparable slots or fillers are given different names. In Figure 1, for instance, a human can easily recognize that Has_strut:proud and Has_gait:majestic are similar properties, but to a computer they can appear very different ideas. WordNet can play a significant role in reconciling these superficial differences in structure (e.g., by recognizing the obvious relationship between strut and gait), while corpus-based co-occurrence models can reveal the comparable nature of proud and majestic. This work, however, is outside the scope of the current paper and is the subject of future development and research.

5 Empirical Evaluation

If similes are indeed a good place to mine the most salient properties of WordNet’s lexical concepts, we should expect the set of properties for each concept to accurately predict how that concept is perceived as a whole. For instance, humans – unlike
computers – do not generally adopt a dispassionate view of ideas, but rather tend to
associate certain positive or negative feelings, or affective values, with particular
ideas. Unsavoury activities, people and substances generally possess a negative affect,
while pleasant activities and people possess a positive affect. Whissell [8] uses
human-assigned ratings to reduce the notion of affect to a single numeric dimension,
to produce a dictionary of affect that associates a numeric value in the range 1.0 (most
unpleasant) to 3.0 (most pleasant) with over 8000 words across a range of syntactic
categories (including adjectives, verbs and nouns). So to the extent that the adjectival
properties yielded by processing similes paint an accurate picture of each noun
vehicle, we should be able to predict the affective rating of each vehicle via a
weighted average of the affective ratings of the adjectival properties ascribed to these
vehicles (i.e., where the affect of each adjective contributes to the estimated affect of
a noun in proportion to its frequency of co-occurrence with that noun in our web-
derived simile data). More specifically, we should expect ratings estimated via these
simile-derived properties to exhibit a strong correlation with the independent ratings
of Whissell’s dictionary.

To determine whether similes do offer the clearest perspective on a concept’s most
salient properties, we calculate and compare this correlation using the following data
sets:

a) Adjectives derived from annotated bona-fide (non-ironic) similes of section 2.1.
b) Adjectives derived from all annotated similes (both ironic and non-ironic).
c) Adjectives derived from ironic similes only.
d) All adjectives used to modify the given vehicle noun in a large corpus. We use
   over 2-gigabytes of text from the online encyclopaedia Wikipedia as our corpus.
e) All adjectives used to describe the given vehicle noun in any of the WordNet text
glosses for that noun. For instance, WordNet defines Espresso as “strong black
   coffee made …” so this gloss yields the properties strong and black for Espresso.

Predictions of affective rating were made from each of these data sources and then
correlated with the ratings reported in Whissell’s dictionary of affect using a two-
tailed Pearson test (p < 0.01). As expected, property sets derived from bona-fide
similes only (A) yielded the best correlation (+0.514) while properties derived from
ironic similes only (C) yielded the worst (-0.243); a middling correlation coefficient
of 0.347 was found for all similes together, demonstrating the fact that bona-fide
similes outnumber ironic similes by a ratio of 4 to 1. A weaker correlation of 0.15 was
found using the corpus-derived adjectival modifiers for each noun (D); while this data
provides far richer property sets for each noun vehicle (e.g., far richer than those
offered by the simile database), these properties merely reflect potential rather than
intrinsic properties of each noun and so do not reveal what is most salient about a
vehicle concept. More surprisingly, perhaps, property sets derived from WordNet
glosses (E) are also poorly predictive, yielding a correlation with Whissell’s affect
ratings of just 0.278.
While it is true that the WordNet-derived properties in (E) are not sense-specific, so that properties from all senses of a noun are conflated into a single property set for that noun, this should not have dramatic effects on predictions of affective rating. Instead, if one sense of a word acquires a negative connotation, then following what is often called “Gresham’s law of language”[9], the “bad meanings should drive out the good” so that the word as a whole becomes tainted. Rather, it may be that the adjectival properties used to form noun definitions in WordNet are simply not the most salient properties of those nouns. To test this hypothesis, we conducted a second experiment wherein we automatically generated similes for each of the 63,935 unique adjective-noun associations extracted from WordNet glosses, e.g., “as strong as espresso”, “as Swiss as Emmenthal” and “as lively as a Tarantella”, and counted how many of these manufactured similes can be found on the web, again using Google’s API.

We find that only 3.6% of these artificial similes have attested uses on the web. From this meagre result we can conclude that: a) few nouns are considered sufficiently exemplary of some property to serve as a meaningful vehicle in a figure of speech; b) the properties used to describe concepts in the glosses of general purpose resources like WordNet are not always the properties that best reflect how humans actually think about, and use, these concepts. Of course, the truth is most likely to lie somewhere between these two alternatives. The space of potential similes is doubtless much larger than that currently found on the web, and many of the similes generated from WordNet are probably quite meaningful and apt. However, even WordNet-based similes that can be found on the web are of a different character to those that populate our database of annotated web-similes, and only 9% of the web-attested WordNet similes (or 0.32% overall) also reside in this database. Thus, most (> 90%) of the web-attested WordNet similes must lie outside the 200-hit horizon of the acquisition process described in section 2, and so are less frequent (or used in less authoritative pages) than our acquired similes.

6 Conclusion

In this paper we have presented an approach to enriching WordNet with the cultural associations that pervade our everyday use of language yet which one rarely finds in authoritative linguistic resources like dictionaries and encyclopaedias. Our means of acquiring these associations – via explicit similes that are mined from the internet – has several important consequences for our enrichment scheme. First, we acquire associations that are neither necessarily true or necessarily consistent with each other, but which people happily assume to be true and consistent for purposes of habitual reasoning. Second, a large-scale mining effort allows us to identify the most frequently used vehicles of comparison, and thus, the landmarks of our shared conceptual space that are most deserving of enrichment in WordNet. Thirdly, we identify the most salient properties of these landmarks, also frequency weighted, as well as the most notable conceptual facets of these landmarks. Interestingly, these combinations of facets and properties (i.e., slot:filler pairings) have a poetic quality that can, in future work, be exploited in the automatic natural-language generation of
creative descriptions.

Despite these benefits, our continued reference to the notion of “culture” may seem misplaced given our focus on English-language similes and an English-language WordNet. Nonetheless, we see this work as a platform from which to explore the cultural diversity of ontological categorizations, and to this end, we are currently planning to replicate this approach for Chinese and Korean. In the case of Chinese, we intend the enrichment process to apply to the Chinese-English lexical ontology of HowNet [10]. To see how similes reflect different biases in different cultures, consider that of the 12,259 unique adjective/noun pairings judged as bona-fide (non-ironic) in section 2.1., only 2,440 (or 20%) have a Chinese translation that can also be found on the web (where translation is performed using the bilingual HowNet). The replication rate for the ironic similes of section 2.1. is even lower, at 5%, reflecting the fact that ironic comparisons are more creatively ad-hoc and less culturally entrenched than non-ironic similes. We can thus expect that the mining of Chinese texts on the web will yield a set of similes – and thus conceptual descriptions (both properties and frames) – that substantially differs from the English-language set described here, to enrich HowNet in an altogether different, culturally-specific way.

References

Comparing WordNet Relations to Lexical Functions

Veronika Vincze, Attila Almási, and Dóra Szauter

Hungarian Academy of Sciences (MTA) and University of Szeged (SZTE),
Research Group on Artificial Intelligence,
H-6720 Szeged, Aradi vértanúk tere 1., Hungary.
vinczev@inf.u-szeged.hu, vizipal@gmail.com, szauter.dora@freemail.hu

Abstract. In this paper, basic relations of WordNet and EuroWordNet are revisited and reconsidered from the viewpoint of lexical functions, that is, formalized semantic relations. Definitions of lexical functions and those of WordNet relations are contrasted and analyzed. The relation near_antonym is found to cover two different semantic relations. Thus, it is suggested that this relation should be divided into two new relations: conversive and antonym. The coding of derivational morphology can also be improved by introducing new relations that encode not only morphological but semantic derivations as well. Finally, some new semantic relations based on lexical functions are also proposed.

Keywords: semantic relations, lexical functions, synonymy, antonymy, holonymy, derivation

1 Introduction

WordNet is a lexical database in which words and lexical units are organized in terms of their meanings and these clusters of words are linked to each other through different semantic and lexical relations. Among the many possible lexical and semantic relations, it is synonymy, hypernymy and antonymy that are of special importance in the construction of WordNet, however, other relations are also encoded in the database. In this paper, basic relations of WordNet are revisited and reconsidered from the viewpoint of lexical functions, that is, formalized semantic relations [1]. We will contrast the definitions of lexical functions and those of WordNet relations and we will show how this comparison can be fruitfully applied in the lexicographical practice of WordNet. We will pay special attention to antonymy, however, other relations such as synonymy, holonymy, hypernymy and different derivations are also analysed in detail. Finally, we will give some hints for some new relations that are listed among lexical functions but are not yet applied in WordNet.
Comparing WordNet Relations to Lexical Functions

2 Relations between Words in WordNet

Dictionaries are usually structured on the basis of word forms: words are alphabetically listed in the dictionary, and their meanings are given one after the other. However, the most innovative aspect of WordNet is that lexical information is organized in terms of meaning, that is, a synset (the basic unit of WordNet) contains words which have approximately the same meaning. Thus, it is synonymy that functions as the essential principle in the construction of WordNet [2].

There are two types of relations among words in WordNet. On the one hand, semantic relations can be found between concepts, in other words, it is not the form of the word that counts but its meaning. Such relations include hyponymy and meronymy. On the other hand, lexical relations are related to word forms, for instance, synonymy, antonymy and different morphological relations belong to this group [2]. Thus, the inner structure of WordNet is based on a specific lexical relation, namely, synonymy.

3 Lexical Functions

The theory of lexical functions was born within the framework of Meaning Text Theory (the model is described in detail in e.g. [1, 3, 4, 5, 6, 7, 8, 9, 10]). The most important theoretical innovation of this model is the theory of lexical functions, which is universal: with the help of lexical functions, all relations between lexemes of a given language can be described – a lexeme is a word in one of its meanings. This theory has been thoroughly applied to different languages such as Russian [7], French [8], English [9, 10] or German [9, 10] and, occasionally to other languages such as Hungarian [11, 12].

Lexical functions have the form $f(x) = y$, where $f$ is the lexical function itself, $x$ stands for the argument of the function and $y$ is the value of the function. The argument of the lexical function is a lexeme, while its value is another lexeme or a set of lexemes. A given lexical function always expresses the same semanto-syntactic relation, that is, the relation between an argument and the value of the lexical function is the same as the relation between another argument and value of the same lexical function. Thus, lexical functions express semantic relations between lexemes [1].

4 Relations Used in WordNet Compared to Lexical Functions

In the following, lexical functions and WordNet relations are contrasted. First, semantic relations such as hyponymy and meronymy are discussed, then lexical and semanto-lexical relations (synonymy and antonymy) are analysed. Finally, derivations are also presented.
4.1 Hypernymy and Hyponymy

In Meaning – Text Theory, it is the lexical function $\text{Gener}$ (generic) that expresses hypernymic relations between words [1]. Some illustrative examples:

\[
\text{Gener}(\text{gas}) = \text{substance} \\
\text{Gener}(\text{wardrobe}) = \text{furniture}
\]

This relation completely corresponds to hypernym used in WordNet since both relations give a more generic term for the word. Thus, it is possible to encode taxonomic relations (such as hypernymy) with the help of lexical functions [13]. The abovementioned examples are present in WordNet, too:

\[
\{\text{substance:1, matter:1}\} \text{ is hypernym of } \{\text{fluid:2}\}, \text{ which is hypernym of } \{\text{gas:2}\} \\
\{\text{furniture:1, piece of furniture:1, article of furniture:1}\} \text{ is hypernym of } \{\text{wardrobe:1, closet:3, press:6}\}
\]

That is, in the first case, the application of lexical functions provided a hypernym situated one level higher than the one present in WordNet, but it is still true that it is a hypernym of the original word. In the second case, however, the two versions are in complete accordance: they provide the same (direct) hypernym for the same word.

Hyponymy can be grabbed by the lexical function $\text{Spec}$ (specific) (proposed in [14]). By definition, $\text{Spec}$ is the inverse function of $\text{Gener}$, thus, it yields less general, i.e. more specific terms (that is, hyponyms) for the word:

\[
\text{Spec}(\text{furniture}) = \text{wardrobe}, \text{ table, chair, desk, bed etc.}
\]

In WordNet, the synset $\{\text{furniture:1, piece of furniture:1, article of furniture:1}\}$ has got no less than 24 hyponyms including $\{\text{table:3}\}$, $\{\text{chest of drawers:1, chest:3, bureau:2, dresser:1}\}$ and $\{\text{bed:1}\}$. Thus, similarly to the case of hypernymy, we can state that the lexical function $\text{Spec}$ and the WordNet relation $\text{hyponym}$ are equivalent.

4.2 Meronymy and Holonymy

In WordNet, holonymy is encoded by three different relations, and in EuroWordNet there are two more relations besides those. First, $\text{holo\_part}$ indicates that a thing is a component part of another thing, second, $\text{holo\_member}$ indicates that a thing or person is a member of a group, and third, $\text{holo\_portion}$ refers to the stuff that a thing is made from [15], however, this relation links a whole and a portion of the whole in EuroWordNet [16]. Fourth, $\text{holo\_madeof}$ encodes the stuff a thing is made from in EuroWordNet, and fifth, $\text{holo\_location}$ indicates a thing that is situated within another place [16].

As for the first relation, its reversed form is $\text{mero\_part}$, which corresponds to the lexical function $\text{Part}$ proposed in [17]:

\[
\text{Part}(\text{label}) = \text{bar code}
\]
The lexical function **Mult** (collective) yields either the group which the word is a member of or a bigger quantity of the entity referred to by the word [1]:

\[
\text{Mult}\text{(ship)} = \text{fleets} \\
\text{Mult}\text{(sheep)} = \text{flocks}
\]

The inverse function of **Mult** is **Sing**, that is, it expresses a member of a group or a minimal unit of a thing:

\[
\text{Sing}\text{(fleets)} = \text{ships} \\
\text{Sing}\text{(crews)} = \text{seamen} \\
\text{Sing}\text{(bread)} = \text{slices}
\]

In WordNet, these relations are shown in the following way:

\{
\text{fleets}:3 \text{ is holo\_member of } \text{ships}:1 \\
\text{sheep}:1 \text{ is mero\_member of } \text{flocks}:5
\}

In EuroWordNet, the last one is encoded by the relation **holo\_portion**:

\{
\text{bread}:1 \text{ is holo\_portion of } \text{pieces}:8, \text{slices}:2
\}

Thus, the functions **Mult** and **Sing** overlap with the WordNet relations **holo\_member** and **mero\_member**, respectively. **Sing** can also encode **mero\_portion** in EuroWordNet. However, other relations such as **holo\_portion** in WordNet and **holo\_madeof** and **holo\_location** in EuroWordNet are not (yet) encoded in the system of lexical functions.

### 4.3 Synonymy

The most basic relation in WordNet is synonymy. Among lexical functions, it is **Syn** that expresses this relation. In the case of the lexical function **Syn**, emphasis is put on quasi-synonyms, that is, the value of this function can be – besides total synonyms – a partial (or quasi-)synonym as well. For instance:

\[
\text{Syn}\text{(bicycle)} = \text{bikes, cycles, wheels}
\]

The relation encoded by this lexical function appears in two forms in WordNet, and there is an additional form in EuroWordNet [16]. First, synonymy is the organizing principle behind the structure of WordNet, thus, between the literals of a synset the relation of synonymy holds by definition (that is, without any explicit reference to this relation). The above-mentioned example is shown in WordNet in the following format:

\{
\text{bicycle}:1, \text{bike}:2, \text{wheel}:6, \text{cycle}:6
\}
Second, in the case of adjectives, the relation similar_to expresses that the meaning of two adjectives are similar though they do not belong to the same synset [17]. An example:

{heavy:2} is similar_to {harsh:4}
{heavy:2} is similar_to {thick:8}

Third, EuroWordNet makes use of the relation near_synonym, which stands between synsets whose meanings are similar but not similar enough to be included in the same synset [16], for instance:

{device:1} is near_synonym of {tool:1}

4.4 Antonymy

Antonymy proved to be hard to define since “[t]he antonym of a word \(x\) is sometimes not-\(x\) but not always” [2]. Conceptually, people can easily find the antonym of a word: they usually give the antonym of a word as a response in word association tests. However, words with similar meanings (especially adjectives and adverbs) do not always have the same antonym. For instance, heavy and light are considered to be antonyms, furthermore, weighty and weightless are also antonyms, heavy and weightless are hardly seen as antonyms, nevertheless [18].

Thus, antonymy seems to behave in two different ways: on the one hand, it is a semantic relation between word meanings, and, on the other hand, it is a lexical relation between word forms, since the antonym of an adjective or an adverb is mostly produced morphologically (with the addition of a negative prefix). The organization of adverbs and adjectives in WordNet reflects this ambiguity: there is an antonymy relation stated only between “real” antonyms (that is, conceptual opposites that are lexical pairs) by the means of the relation near_antonym, and antonymy between indirect antonyms is expressed in WordNet only through inheritance [18].

In the theory of lexical functions, however, conceptual antonyms are not distinguished from lexical antonyms, that is, antonymy is considered to be a semantic relation rather than a lexical one. Antonymy (that is, the lexical function \text{Anti}) is defined in the following way [1]: "la lexie \(L_1\) est un antonyme de la lexie \(L_2\) si et seulement si leurs signifiés sont identiques sauf pour la négation se trouve « au sein » d’un des deux signifiés” [lexeme \(L_1\) is an antonym of lexeme \(L_2\) if and only if their meanings are identical except that negation is present « within » one of the meanings – translation is ours]. Some examples are given here:

\[1\] In the case of adjectives of Italian WordNet, antonymy is further divided into two subrelations: complementary_antonymy (if a word holds, then its opposite is excluded, e.g. dead and alive) and gradable_antonymy (words referring to gradable properties, e.g. big and small). Besides, the underspecified relation antonymy also survives for cases when the nature of the opposition is unclear [19]. In the present discussion, however, only the general relation antonymy is examined thoroughly.
Comparing WordNet Relations to Lexical Functions 467

\[ \text{Anti} (\text{despair}) = \text{hope} \]
\[ \text{Anti} (\text{construct}) = \text{destroy} \]
\[ \text{Anti} (\text{respect}) = \text{disrespect} \]

There is another lexical function, \textit{Conv} (conversive), that expresses a relation similar to antonymy: the semantic content of the argument and the value of \textit{Conv} are identical, however, the actants, that is, the participants of situation described are reversed, which is indicated by index numbers [1] such as in the following examples:

\[ \text{Conv}_{21}(\text{frighten}) = \text{fear} \ (\text{Death frightens me vs. I fear death}) \]
\[ \text{Conv}_{3214}(\text{buy}) = \text{sell} \ (\text{John bought a pair of shoes from Mary for $25 vs. Mary sold a pair of shoes to John for $25}) \]

It is important to emphasize that the lexical functions \textit{Anti} and \textit{Conv} differ from each other: \textit{Conv} often yields a lexeme which seems to be a quasi-antonym of the original word, however, it is not the case since the antonym of the word is provided by the application of \textit{Anti}. The following examples nicely illustrate the difference between \textit{Anti} and \textit{Conv}:

\[ \text{Conv}_{21}(\text{send}) = \text{receive} \ (\text{Peter sent a letter to John vs. John received a letter from Peter}) \]
\[ \text{Anti}(\text{send}) = \text{intercept} \ (\text{to cause that the letter does not arrive}) \]
\[ \text{Conv}_{21}(\text{equal}) = \text{equal} \ (\text{1000 metres are equal to 1 kilometre vs. 1 kilometre is equal to 1000 metres}) \]
\[ \text{Anti}(\text{equal}) = \text{unequal} \]

It is quite clear from the examples above that the semantic content of the two lexical functions are different for their application to the same word provides different values. This is especially striking in the second case: \textit{equal} is its own conversive, however, it cannot be its own antonym.

Nevertheless, in WordNet, these two relations are not differentiated: they are covered by the relation \textit{near_antonym} (or, in Italian WordNet, \textit{antonymy}, \textit{grad_antonymy} and \textit{comp_antonymy}). Some of “real” antonyms are not listed at all or some conversives are given as (near-)antonyms. Here we provide a selection of synset pairs that are connected through the relation \textit{near_antonym} in WordNet:

\{ \text{give}:3 \} \text{ is near_antonym of } \{ \text{take}:8 \}
\{ \text{sell}:1 \} \text{ is near_antonym of } \{ \text{buy}:1, \text{purchase}:1 \}
\{ \text{rise}:16, \text{come up}:10, \text{uprise}:5, \text{ascend}:6 \} \text{ is near_antonym of } \{ \text{set}:10, \text{go down}:7, \text{go under}:2 \}
\{ \text{hire}:1, \text{engage}:3, \text{employ}:2 \} \text{ is near_antonym of } \{ \text{fire}:4, \text{give notice}:1, \text{can}:2, \text{dismiss}:4, \text{give the axe}:1, \text{send away}:2, \text{sack}:2, \text{force out}:2, \text{give the sack}:1, \text{terminate}:4 \}
It can be seen on the basis of the examples that there is no unique well-defined relation that holds between all members of pairs connected by the relation near_antonym. Basically, these antonym pairs can be divided into two groups: in one group we can find pairs that are opposites of each other in the sense that the words of the pair cannot be applied in the same situation, that is, the members of the pair are mutually exclusive – for instance, if you get on a bus, you cannot get off the bus at the same time). However, in the other group, there are pairs whose members necessarily coexist – as an example, if there exists a wife, then there must be a husband, too.

Thus, we suggest that the relation near_antonym (and antonymy) should be divided into two new relations: antonym and conversive. On the one hand, antonym would function similar to the lexical function Anti, that is, it would form a link between synsets whose meanings differs from each other only with respect to an inner negation. On the other hand, synsets connected to each other through the relation conversive would describe the same situation or refer to the same action but from a different perspective: another participant of the situation becomes more important, thus, a new aspect is emphasized – just like the lexical function Conv does.

The above-mentioned examples can be categorized into an Anti- and a Conv-group. This is shown here:

\[
\begin{align*}
\text{Conv}_{31}(sell) &= \text{buy, purchase} \\
\text{Conv}_{34}(give) &= \text{take} \\
\text{Conv}_{21}(parent) &= \text{child, kid} \\
\text{Conv}_{21}(wife) &= \text{husband} \\
\text{Anti}(get on) &= \text{get off} \\
\text{Anti}(man) &= \text{woman} \\
\text{Anti}(rise) &= \text{set} \\
\text{Anti}(enemy) &= \text{friend} \\
\text{Anti}(employ) &= \text{fire}
\end{align*}
\]

We also mention that there are some words having both conversive and antonym pairs. These are the most illustrative examples of the necessity of splitting the original near_antonym relation into two relations: conversive and antonym. Besides the examples of equal and receive (given above), another case is provided here:

\[
\text{Conv}_{21}(spouse) = \text{spouse} \quad \text{(that is, spouse is its own conversive)}
\]
**Anti** (*spouse*) = *lover* (that is, someone who acts similarly to a spouse towards someone who is not his or her spouse)

However, in WordNet, {spouse:1, partner:1, married person:1, mate:4, better half:1} and {lover:3} are not connected to each other in any way. They share their hypernym but no antonymy relation holds between them. According to our proposed relations, these synsets should be represented in the following way:

{spouse:1, partner:1, married person:1, mate:4, better half:1} is **antonym** of {lover:3}

{spouse:1, partner:1, married person:1, mate:4, better half:1} is **conversive** of {spouse:1, partner:1, married person:1, mate:4, better half:1}

In the same way, the synsets {wife:1, married woman:1} and {husband:1, hubby:1, married man:1} should also be linked to {mistress:1, kept woman:1, fancy woman:2} and {fancy man:2, paramour:1} with the relation **antonym**, respectively. However, they are each other’s conversive.

### 4.5 Derivation

Certain morphological relations between word forms – namely, instances of derivational morphology – are also encoded in WordNet by the means of the relations eng_derivative and derived. These relations differ from the previously mentioned ones in an important aspect: they do not hold between all members of the two synsets linked to each other. It is usually one literal in the synset that serves as the basic form for the derivation, and, on the other hand, the derived form can also have some synonyms within its own synset. Thus, morphological relations hold rather between word forms than between synsets.

In WordNet, we can find examples of nominal, verbal, adjectival and adverbial derivations as well although the nature of the derivation (that is, nominal, verbal etc.) is not explicitly stated:

{quantify:2, measure:2} —>> {quantification:2}

{energy:1} —>> {excite:3, energize:2, energise:1}

{membrane:2, tissue layer:1} —>> {membranous:1}

{real:1, existent:2} —>> {actually:1, really:1}

However, these derivations are not differentiated with respect to the semantic connection that exists between the original word and the derived one. To put it in another way, it is indicated that a certain morphological relation holds between the words but the semantic nature of this relation is left underspecified. Obviously, the definitions of the words contain pieces of information from which the semantic relation can be calculated but when looking for a special type of word or words having specific grammatical features (for instance, agents or patients), it is time-

consuming to look up every single synset that is connected to the original one through
the relation eng\_derivative or derived in order to find the necessary ones.

As it can be expected, there are lexical functions which – instead of changing the
semantic content of the word – change the syntactic features of the word, in other
words, they preserve the semantic content but change the part-of-speech of the word
by derivation. Lexical functions $S_0$, $V_0$, $A_0$ and $Adv_0$ nominalize, verbalize, adjectivize and adverbialize the original word, respectively [1]:

\[
S_0(\text{present}) = \text{presentation} \\
V_0(\text{verbal}) = \text{verbalize} \\
A_0(\text{beauty}) = \text{beautiful} \\
Adv_0(\text{hardy}) = \text{hard}
\]

Other lexical functions specify a participant in the situation described by a verb. For instance, the lexical function $S_1$ gives the agent while the patient is provided by $S_2$
and $S_3$ generates another participant who is involved in the situation:

\[
S_1(\text{write}) = \text{author} \\
S_2(\text{speak}) = \text{speech} \\
S_3(\text{speak}) = \text{addressee}
\]

Lexical functions belonging to the latter type usually represent derivations that are
not considered to be productive or systematic, however, the syntacto-semantic
relation between the two lexical units is evident [1].

As the comparison of lexical functions and WordNet relations concerning
derivational morphology reveals, lexical functions offer a more detailed analysis of
derivational relations than WordNet relations in their present form do. Thus, we
propose that WordNet relations encoding derivational morphology should be
enhanced in order to provide a more precise and accurate network of words from a
derivational point of view as well. With the introduction of relations $S_0$, $V_0$, $A_0$ and
$Adv_0$, instances of derivational morphology would be easier to be detected in
WordNet. On the other hand, the application of the relations $S_1$, $S_2$ and $S_3$ would make
it possible to search for semantic derivations in WordNet. Finally, both innovations
would be of use in second language acquisition (for WordNet used in second
language teaching, see e.g. [20]).

To conclude the comparison of lexical functions and WordNet relations, we
provide a table summarizing the parallels between the two systems:
Comparing WordNet Relations to Lexical Functions

Table 1. Parallels between WordNet relations and lexical functions.

<table>
<thead>
<tr>
<th>Relation</th>
<th>WordNet, EuroWordNet</th>
<th>Lexical function</th>
</tr>
</thead>
<tbody>
<tr>
<td>synonymy</td>
<td>synset</td>
<td>Syn</td>
</tr>
<tr>
<td></td>
<td>similar_to</td>
<td></td>
</tr>
<tr>
<td></td>
<td>near_synonym</td>
<td></td>
</tr>
<tr>
<td>antonymy</td>
<td>near_antonym</td>
<td>Anti, Conv</td>
</tr>
<tr>
<td></td>
<td>antonymy</td>
<td></td>
</tr>
<tr>
<td>hypernymy</td>
<td>hypernym</td>
<td>Gener</td>
</tr>
<tr>
<td>hyponymy</td>
<td>hyponym</td>
<td>Spec</td>
</tr>
<tr>
<td>holonymy</td>
<td>holo_member</td>
<td>Mult</td>
</tr>
<tr>
<td>meronymy</td>
<td>mero_member, mero_portion</td>
<td>Sing. Part</td>
</tr>
<tr>
<td></td>
<td>(in EWN), mero_part</td>
<td></td>
</tr>
<tr>
<td>derivational</td>
<td>eng_derivative, derived</td>
<td>S0, V0, A0, Adv0</td>
</tr>
<tr>
<td>morphology</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.6 New Relations

In this section we discuss some semantic relations encoded by lexical functions that have no equivalent in WordNet. However, we think that their addition would contribute to the future development of WordNet in a useful way.

A semantic relation that is not yet present in WordNet is formalized with the help of the lexical function Cap: it signals the leader or boss of something [1]. For instance:

Cap(university) = chancellor
Cap(ship) = captain
Cap(school) = headmaster

In WordNet, this semantic relation could be marked in the following way:

{master:7, captain:4, sea captain:1, skipper:2} is leader of {ship:1}

Another lexical function that can be applied in WordNet, too is Equip, which refers to the staff of something [1]. Some illustrations:

Equip(theatre) = company
Equip(ship) = crew

This relation can be encoded in WordNet in this way:

{crew:1} is staff of {ship:1}
A third relation that we propose links nouns and verbs, namely, it is the verb having the sense of producing the typical sound of the noun. This lexical function is called Son [1]:

Son(\textit{dog}) = \textit{bark}
Son(\textit{pig}) = \textit{grunt}

In WordNet, the newly proposed relation sound can stand for this relation:

\{\textit{grunt}:1\} sound of \{\textit{hog}:3, \textit{pig}:1, \textit{grunter}:2, \textit{squealer}:2, \textit{Sus scrofa}:2\}

In sections 4.4 and 4.5 we already proposed the splitting of the relation near\_antonym into conversive and antonym and the introduction of relations encoding derivational connections. Thus, we refrain from repeating our argumentation here. Instead, we summarize the newly proposed semantic relations in the following table:

<table>
<thead>
<tr>
<th>Proposed relation</th>
<th>WordNet</th>
<th>Lexical Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>conversion</td>
<td>conversive</td>
<td>Conv</td>
</tr>
<tr>
<td>antonymy</td>
<td>antonym</td>
<td>Anti</td>
</tr>
<tr>
<td>leadership</td>
<td>leader</td>
<td>Cap</td>
</tr>
<tr>
<td>staff</td>
<td>staff</td>
<td>Equip</td>
</tr>
<tr>
<td>typical sound</td>
<td>sound</td>
<td>Son</td>
</tr>
</tbody>
</table>

5 Conclusion

In this paper, we compared some semantic and lexical relations used in WordNet and EuroWordNet and the corresponding lexical functions in Meaning-Text Theory. We found that in the case of synonymy, hypernymy and holonymy, lexical functions and WordNet relations are equivalent. However, we found that the relation near\_antonym, that is, the one encoding antonymy in WordNet covers two different semantic relations. Thus, we suggested that this relation should be divided into two new relations: conversive and antonym. The coding of derivational morphology can also be improved by introducing new relations that encode not only morphological but semantic derivations as well. Finally, we proposed some new semantic relations that may contribute to the further development of the complex but colourful network of words found in WordNet.

Acknowledgements

Thanks are due to our two reviewers, Antonietta Alonge and Kadri Vider for their useful comments and remarks, which helped us improve the quality of this article.
References

KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures

Piek Vossen1, 2, Eneko Agirre3, Nicoletta Calzolari4, Christiane Fellbaum5, 6, Shu-Kai Hsieh7, Chu-Ren Huang8, Hitoshi Isahara9, Kyoko Kanzaki9, Andrea Marchetti10, Monica Monachini4, Federico Neri11, Remo Raffaelli11, German Rigau3, Maurizio Tesconi10, and Joop VanGent2

1 Faculteit der Letteren, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081HV Amsterdam, Netherlands
p.vossen@let.vu.nl
2 Irion technologies, Delftechpark 26, 2628XH Delft, Netherlands {piek.vossen, gent}@irion.nl
3 IXA NLP group, University of the Basque Country, Manuel Lardizabal 1, Donostia, Basque Country {e.agirre, g.rigau@ehu.es}
4 Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche, Via Moruzzi 1, 56124 Pisa, Italy
nicoletta.calzolari@ilc.cnr.it, monica.monachini@ilc.cnr.it
5 Berlin-Brandenburg Academy of Sciences, Berlin, Germany
6 Princeton University, Princeton, USA
7 National Taiwan Normal University, Republic of China
8 Academia Sinica, Taipei, Republic of China
9 NICT, Kyoto, Japan
10 Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Via Moruzzi 1, 56124 Pisa, Italy
andrea.marchetti@iit.cnr.it, maurizio.tesconi@iit.cnr.it
11 Synthema, Pisa, Italy

Abstract. We outline work to be carried out within the framework of an impending EC project. The goal is to construct a language-independent information system for a specific domain (environment/ecology) anchored in a language-independent ontology that is linked to WordNets in several languages. For each language, information extraction and identification of lexicalized concepts with ontological entries will be done by text miners (“Kybots”). The mapping of language-specific lexemes to the ontology allows for crosslinguistic identification and translation of equivalent terms. The infrastructure developed within this project will enable long-range knowledge sharing and transfer to many languages and cultures, addressing the need for global and uniform transition of knowledge beyond the domain of ecology and environment addressed here.

Keywords: Global WordNet Grid, ontologies and WordNets, multilinguality, semantic indexing and search, text mining.
1 Introduction

Economic globalization brings challenges and the need for new solutions that can serve all countries. Timely examples are environmental issues related to rapid growth and economic developments such as global warming. The universality of these problems and the search for solutions require that information and communication be supported across a wide range of languages and cultures. Specifically, a system is needed that can gather and represent in a uniform way distributed information that is structured and expressed differently across languages. Such a system should furthermore allow both experts and laymen to access information in their own language and without recourse to cultural background knowledge.

Addressing sudden and unpredictable environmental disasters (fires, floods, epidemics, etc.) requires immediate decisions and actions relying on information that may not be available locally. Moreover, the sharing and transfer of knowledge are essential for sustainable growth and long-term development. In both cases, it is important that information and experience are not only distributed to assist with local emergencies but are universally re-usable. In these settings, natural language is the most ubiquitous and flexible interface between users – especially non-experts – and information systems.

The goal of "Knowledge-Yielding Ontologies for Transition-Based Organization" (KYOTO) is, first, to develop a content enabling system that provides deep semantic search. KYOTO will cover access to a broad range of multimedia data from a large number of sources in a variety of culturally diverse languages. The data will be accessible to both experts and the general public on a global scale.

KYOTO is funded under project number 211423 in the 7th Frame work project in the area of Digital Libraries: FP7-ICT-2007-1, Objective ICT-2007.4.2: Intelligent Content and Semantics. It will start early 2008 and last 3 years. The consortium consists of research institutes, companies and environmental organizations: Vrije Universiteit Amsterdam (Amsterdam, The Netherlands), Consiglio Nazionale delle Ricerche (Pisa, Italy), Berlin-Brandenburg Academy of Sciences and Humanities (Berlin, Germany), University of the Basque Country (Donostia, Basque Country), Academia Sinica (Taipé, Taiwan), National Institute of Information and Communications Technology (Kyoto, Japan), Irion Technologies (Delft, The Netherlands), Synthema (Pisa, Italy), European Centre for Nature Conservation (Tilburg, The Netherlands), World Wide Fund for Nature (Zeist, The Netherlands), Masaryk University (Brno, Czech). In total 364 person months of work are involved. The partners from Taiwan and Japan are funded by national grants.

2 The KYOTO System: Overview

KYOTO is a generic system offering knowledge transition and information across different target groups, transgressing linguistic, cultural and geographic boundaries. Initially developed for the environmental domain, KYOTO will be usable in any knowledge domain for mining, organizing, and distributing information on a global scale in both European and non-European languages.
KYOTO's principle components are an ontology linked to WordNets in a broad range of languages (Basque, Chinese, Dutch, English, Italian, Japanese, Spanish), linguistic text miners, a Wiki environment for supporting and maintaining the system, and a portal for the environment domain that allows for deep semantic searches.

Concept extraction and data mining are applied through a chain of semantic processors ("Kybots") that share a common ground and knowledge base and re-use the knowledge for different languages and for particular domains.

Information access is provided through a cross-lingual user-friendly interface that allows for high-precision search and information dialogues for a variety of data from wide-spread sources in a range of different languages. This is made possible through a customizable, shared ontology that is linked to various WordNets and that guarantees a uniform interpretation for diverse types of information from different sources and languages.

The system can be maintained and kept up to date by specialists in the field using an open Wiki platform for ontology maintenance and WordNet extension.

Figure 1 gives an overview of the complete system. In this schema, information stored in various media and languages, distributed over different locations, is collected through a Capture module and stored in a uniform XML representation. For each language, concept miners are applied to derive concepts that occur in the textual data and compare these with the given WordNets for the different languages. The WordNets provide a mapping to a single shared ontology. Both the WordNets and the ontology can be modified and edited in a special Wiki environment by the people in a community; in the present project, these will be specialists in the environment domain. Encoding of knowledge and WordNets for a domain will result in more
precise and effective mining of information and data through fact mining by the so-called Kybots. Kybots will be able to detect specific patterns and relations in text because of the concepts and constraints coded by the experts. These relations are added to the XML representation of the captured text. An indexing module then creates the indexes for different databases and data types that can be accessed by the users through a text search interface or possibly dialogue systems. The users can be the same environmental organisations, and/or governments and citizens.

In the next sections, we will discuss KYOTO’s major components in more detail.

3 The Ontology

The ontology, where knowledge of concepts is formally encoded, consists of three layers. The top layer is based on existing top level ontologies, among them SUMO [10, 11], DOLCE [8] and the MEANING Top Concept Ontology [3]. We will investigate what ontology will be the best basis for our purpose and can also be shared across the diverse languages and cultures. If necessary, ontology fragments or elements can be shared or a selection will be made. We do not expect major differences in the fundamental semantic organisation of the different languages. Recent studies, for example, show that the Chinese radical system and character compounding tend to be based on the same qualia distinctions as in the Generative Lexicon [4, 5].

The middle layer will be derived from existing WordNets, where concepts are mapped to lexical units. The ontology's mid-level must be developed such that it connects domain terms and concepts to the top-level. We define all the high-level and mid-level concepts that are needed to accommodate the information in the environmental domain. Knowledge is implemented at the most generic level to maximize re-usability yet precisely enough to yield useful constraints in detecting relations. Within the domain, we extend the ontologies to cover all necessary concepts and applicable, sharable relations.

The domain terms are extracted semi-automatically from the source documents or manually created through a Domain Wiki. The Domain Wiki allows experts to modify and extend the domain level of the ontology and extend the WordNets accordingly. It enables community-based resource building, which will lead to increased, shared understanding of the domain and at the same time result in the formalization of this knowledge, so that it can be used by an automatic system.

This resource will build on the Multilingual Central Repository (MCR) knowledge base [1] developed in the MEANING project [12]. Currently, the MCR consistently integrates more than 1.6 million semantics links among concepts. Moreover, the current MCR has been enriched with about 460,000 semantic and ontological properties [2]: Base Concepts and Top Concept Ontology [3], WordNet Domains [7], Suggested Upper Merged Ontology (SUMO) [10], providing ontological coherence to all the uploaded WordNets.

Extensions to WordNets and the ontology will be propagated through appropriate sharing protocols, developed exploiting LeXFlow, a framework for rapid prototyping of cooperative applications for managing lexical resources (XFlow [9] and LeXFlow
The shared ontology guarantees a uniform interpretation layer for the diverse information from different sources and languages. At the lowest level of the ontology, we expect that abstract constraints and structures can be hidden for the users but can still be used to prevent fundamental errors, e.g. creating a concrete concept for an adjective. The Wiki users should focus on formulating conditions and specifications that they understand without having to worry about the linguistic and knowledge engineering aspects. They can discuss these specifications within their community to reach consensus and provide proper labels in each language.

4 Kybots

Once the ontological anchoring is established, it will be possible to build text mining software that is able to detect semantic relations and propositions. Data miners, so-called Kybots (Knowledge-yielding robots), can be defined using constraints among relations at a generic ontological level. These logical expressions need to be implemented in each language by mapping the conceptual constraint onto linguistic patterns. A collection of Kybots created in this way can be used to extract the relevant knowledge from textual sources represented in a variety of media and genres and across different languages and cultures. Kybots will represent such knowledge in a uniform and standardized XML format, compatible with WWW specifications for knowledge representation such as RDF and OWL.

Kybots will be developed to cover users' questions and answers as well as generic concepts and relations occurring in any domain, such as named-entities, locations, time-points, etc. Kybots are primarily defined at a generic level to maximize re-usability and inter-operability. We develop the Kybots that are necessary for the selected domain but the system can easily be extended and ported to other domains.

The Kybots will operate on a morpho-syntactic and semantic encoding level that will be the same across all the languages. Every group will use existing linguistic processors or develop additional ones when needed to foresee in a basic linguistic analysis, which involves: tokenization, segmentation, morpho-syntactic tagging, lemmatization and basic syntactic parsing. Each of these processes can be different but the XML encoding of the output will be the same. This will guarantee that Kybots can be applied to the output of text in different languages in a uniform way. We will use as much as possible existing and available free software for this process. Note that the linguistic expression rules of ontological patterns in a specific Kybot are to be defined on the basis of the common output encoding of the linguistic processors. Likewise, they can share specifications of linguistic expression in so far the relations are expressed in the same way in these languages.

5 Indexing, Searching, and Interfacing

The extracted knowledge and information is indexed by an existing search system that can handle fast semantic search across languages. It uses so-called contextual conceptual indexes, which means that occurrences of concepts in text are interpreted
by their co-occurrence with other concepts within a linguistically defined context, such as a noun phrase or sentence. The co-occurrence patterns of concepts can be specified in various ways, possibly based on semantic relations that are defined in the logical expressions. Thus, the system yields different results for searches for *polluting substance* and *polluted substance*, because these involve different semantic relations between the same concepts. By mapping a query to concepts and relations, very precise matches can be generated, without the loss of scalability and robustness found in regular search engines that rely on string matching and context windows.

Reasoning over facts and ontological structures will make it possible to handle diverse and more complex types of questions. Crosslinguistic and crosscultural understanding is vouchsafed through the ontological anchoring of language via WordNets and text miners.

6 The Wiki Environment

The Wiki environment enables domain experts to easily extend and manage the ontology and the WordNets in a distributed context, to constantly reflect the continuous growth and changes of the data they describe. It owns the characteristics typical of a generic Wiki engine:

- Web based highly-interactive interface, tailored to domain experts who don't know the underlying complex data model (ontology plus WordNet of different languages);
- tools to support collaborative editing and consensus achievement such as discussion forums, and list of last updates;
- automatic acquisition of information from external Web resources (e.g. Wikipedia);
- rollback mechanism: each change to the content is versioned;
- search functions providing the possibility to define different search patterns (synset search, textual search and so on);
- role-based user management.

In addition, the Wiki engine manages the underlying complex data model of the ontology and the WordNets so as to keep it consistent: this is achieved through the definition of appropriate sharing protocols. For instance, when a new domain term such as *water pollution* is inserted into a language-specific WordNet by a domain expert, a new entry, referred to as dummy entry because of the incompleteness of the information represented, will be automatically created and added to the ontology and in the remaining WordNets. The Wiki environment will list all dummy entries to be filled in, in order to notify them to domain experts allowing for their complete definition and integration into KYOTO ontological and lexical resources. In this context, English can be used as the common ground language in order to support the extension process and the propagation of changes among the different WordNets and the ontology.
7 Sharing

Knowledge sharing is a central aspect of the KYOTO system and occurs on multiple levels.

7.1 Sharing and Re-Use of Generic Knowledge

Sharing of generic ontological knowledge in the domain takes place mainly through subclass relations. We collect all the relevant terms in each language for the domain and add them to the general ontology. Possibly, these concepts can be imported from a specific WordNet and "ontologized." It will be important to specify exactly the ontological status of the terms. Only disjunct types need to be added [6]. For example, CO2 is a type of substance, whereas greenhouse gases do not represent a different type of gas or substance but refers to substances that play a specific role in specific circumstances. In so far as new definitions and axioms need to be specified, they can be added for the specific subtypes in the domain. However, this is only necessary if the related information also needs to be mined from the text and is not already covered by the generic miners. Next, the generic and domain knowledge is shared among all participating languages through the mapping of the different WordNets to the ontology.

Extension to different domains is possible though not within the scope of the current project.

7.2 Sharing and Re-Use of Generic Kybots

The sharing of Kybots is more subtle. For example, concentrations of substances, causal relations between processes or conditional states for processes can be stated as general conceptual patterns using a simple logical expression. Within a specific domain, any of these relations and conditions could be detected in the textual data by just using these general patterns. Thus, people usually do not use special words in a language to refer to the causal relation itself but they use general words such as "cause" or "factor". Since any causal relation may hold among processes and or states, they can also hold in the environmental domain. Certain valid conditions can be specified in addition to the general ones, as they are relevant for the users. For example, CO2 emissions can be derived from a certain process involving certain amounts of the substance CO2 but critical levels can be defined in the text miner as a conceptual constraint. Furthermore, we may want to limit the ambiguity of interpretation that arises at the generic levels to only one interpretation at the domain level; it is currently an open question to what extent generic patterns can be used or need to be tuned.

Each language group can build a Kybot, capturing a particular relation. A given logical expression that underlies the Kybot of another language can be re-used, or a new pattern can be formulated for a language and a generic universal pattern derived from it. We foresee a system where the text miners can load any set of Kybots in combination with the ontology, a set of WordNets and expression rules in each
language. Each Kybot, a textual XML file, contains a logical expression with constraints from the ontology (either the general ontology or a domain instantiation). Through the WordNets and the expression rules, the text miner knows how to detect a pattern in running text for each specific language. In this way, logical patterns can be shared across languages and across domains.

A Kybot can likewise be developed by a group in one language and taken up by another group to apply it to another language. Consider the case where a generic linguistic text miner is formulated for Dutch, based on Dutch words and expressions. This Kybot is projected to the ontology via the Dutch WordNet, becoming a generic ontological expression which relates two ontological classes: a Substance to a Process. This expression may be extended to a domain, where it is applied to CO2 and CO2 emissions. Next, the Spanish group can load the domain specific expression and transform it into a Spanish Kybot that can be applied to a domain text in Spanish. To turn an ontological expression into a Kybot, language expressions rules and functions need to be provided. This process can be applied to all the participating languages, where the basic knowledge is shared.

7.3 Cross-Linguistic Sharing of Ontologies

KYOTO will thus generate Kybots in each language that go back to a shared ontology and shared logical expressions. Thus, KYOTO can be seen as a sophisticated platform for anchoring and grounding meaning within a social community, where meaning is expressed and conceived differently across many languages and cultures. It also immediately makes this shared knowledge operational so that factual knowledge can be mined from unstructured text in domains. KYOTO supports interoperability and sharing across these communities since much knowledge can be re-used in any other domain, and the ontologies support both generic and domain-specific knowledge.

8 Evaluation

The KYOTO system is evaluated in various ways:

1. WordNets and ontologies are evaluated across linguistic partners;
2. Language and ontology experts will use the Wiki system to build the basic ontology and WordNet layers needed for the extension to the domain;
3. Domain experts will use the top layer and middle layer of WordNets and ontologies plus the Wiki system to encode the knowledge in their domains and reach consensus;
4. The system is tested by integration in a retrieval system;

Cross-linguistic re-use and agreement on the semantic organization is the prime evaluation of the architecture and the system. Proposals for concepts are verified by other WordNet builders and need to be agreed across the languages and cultures. The same happens by domain experts in their domain, except that they do not need to
discuss the technical conceptual issues. Both groups will extensively use the Wiki environment to reach agreements and consensus.

The application driven evaluation will use a baseline evaluation that uses the current indexing and retrieval system and the multilingual WordNet database. The knowledge in KYOTO will lead to more advanced indexes in those cases that Kybots have been able to detect the relations in the text. These will lead to more precision in the indexes and also make it possible to detect complex queries for these relations. The performance if the system will be evaluated with respect to the baseline systems. This will be done in two ways:

1. using an overall benchmark system that runs a fixed set of queries on the different indexes and compares the results;
2. using end-user scenarios and interviews carried out on different indexes by test persons;

The questions and queries are selected to show the capabilities of deep semantic processing. They will be harvested from current portals in the environmental domain.

Finally, we plan to give public access to the databases (ontologies and WordNets) and to the retrieval system through the project website. Visitors are invited to try the system and give feedback.

9 Summary and Outlook

KYOTO will represent a unique platform for knowledge sharing across languages and cultures that can represent a strong content based standardisation for the future that enables world wide communication.

KYOTO will advance the state-of-the-art in semantic processing because it is a unique collaboration that bridges technologies across semantic web technologies, WordNet development and acquisition, data and knowledge mining and information retrieval.

On top of the systems and data described earlier, we will build a Wiki environment that will allow communities to maintain the knowledge and information, without expert knowledge of ontologies, knowledge engineering and language technology. The system can be used by other groups and for other domains. Through simple and clear interfaces that exploit the generic knowledge and check the underlying structures, users can reach semantic agreement on the definition and interpretation of crucial notion in their domain. The agreed knowledge can be taken up by generic Kybots that can then detect possible relations on the basis of this knowledge in text that will be indexed and made searchable. All knowledge resources in KYOTO will be public and open source (GPL). This applies to the ontology and the WordNets mapped to the ontology. The GPL condition also applies to the data miners in each language, the DEB servers, the LeXFlow API and the Wiki environments. Any research group should be able to further develop the system, to integrate their own language and/or to apply it to any other domain.
Acknowledgement

The work described here is funded by the European Community, 7th Framework

References


**Relevant URLs**

The Cornetto Database: Architecture and Alignment
Issues of Combining Lexical Units, Synsets and an Ontology

Piek Vossen1,2, Isa Maks1, Roxane Segers1, Hennie van der Vliet1, and Hetty van Zutphen2

1 Faculteit der Letteren, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
{p.vossen, e.maks, r.segers, hd.vandervliet}@let.vu.nl
2 Irion Technologies, Delftechpark 26, 2628 XH, Delft, The Netherlands
hetty.van.zuthphen@irion.nl

Abstract. Cornetto is a two-year Stevin project (project number STE05039) in which a lexical semantic database is built that combines WordNet with Framenet-like information for Dutch. The combination of the two lexical resources (the Dutch WordNet and the Referentie Bestand Nederlands) will result in a much richer relational database that may improve natural language processing (NLP) technologies, such as word sense-disambiguation, and language-generation systems. In addition to merging the Dutch lexicons, the database is also mapped to a formal ontology to provide a more solid semantic backbone. Since the database represents different traditions and perspectives of semantic organization, a key issue in the project is the alignment of concepts across the resources. This paper discusses our methodology to first automatically align the word meanings and secondly to manually revise the most critical cases.

Keywords: WordNet, synsets, lexical units, frames, ontologies, automatic alignment

1 Introduction

Cornetto is a two-year Stevin project (project number STE05039) in which a lexical semantic database is built that combines WordNet with Framenet-like information for Dutch. In addition, the database is also mapped to a formal ontology to provide a more solid semantic backbone. The combination of the lexical resources will result in a much richer relational database that may improve natural language processing (NLP) technologies, such as word sense-disambiguation, and language-generation systems. The database will be filled with data from the Dutch WordNet [18] and the Referentie Bestand Nederlands [10]. The Dutch WordNet (DWN) is similar to the Princeton WordNet for English, and the Referentie Bestand (RBN) includes frame-like information as in FrameNet plus other information on the combinatoric behaviour of word meanings. RBN has corpus-based examples and rich morpho-syntactic
structures with complementation information. It furthermore contains many multi
word expressions, both free, partly fixed and frozen expressions.

An important aspect of combining the resources is the alignment of the semantic
structures. In the case of RBN these are lexical units (LUs) and in the case of DWN
these are synsets. Various heuristics have been developed to do an automatic
alignment. Following automatic alignment of RBN and DWN, this initial version of
the Cornetto database will be further extended both automatically and manually. The
resulting data structure is stored in a database that keeps separate collections for
lexical units (mainly derived from RBN), for the synsets (derived from DWN) and for
a formal ontology: SUMO/MILO plus extensions [15]. These 3 semantic resources
represent different viewpoints and layers of linguistic, conceptual information. The
alignment of the viewpoints is stored in a separate mapping table. The database is
itself set up so that the formal semantic definition of meaning can be tightened for
lexical units and synsets by exploiting the semantic framework of the ontology. At the
same time, we want to maintain the flexibility to have a wide coverage for a complete
lexicon and to encode additional linguistic information. The resulting resource will be
made freely available for research in the form of an XML database.

Combining two lexical semantic databases with different organizational principles
offers the possibility to study the relations between these perspectives on a large
scale. However, it also makes it more difficult to align the two databases and to come
to a unified view on the lexical semantic organization and the sense distinctions of the
Dutch vocabulary. In this paper, we discuss the alignment issues. In section 2, we first
give an overview of the structure of the database. Section 3 describes the approach
and results of the automatic alignment. Section 4, discusses the manual work of
checking and improving the automatic process. This work mainly involves comparing
the LUs from RBN with the synset structure of DWN. Finally, in section 5, we
discuss the relation between synsets and the ontology.

2 Architecture of the Database

The Cornetto database (CDB) consists of 3 main data collections:

- Collection of Lexical Units, mainly derived from the RBN
- Collection of Synsets, mainly derived from DWN
- Collection of Terms and axioms, mainly derived from SUMO and MILO

Both DWN and RBN are semantically based lexical resources. RBN uses a
traditional structure of form-meaning pairs, so-called Lexical Units [3]. Lexical Units
are word senses in the lexical semantic tradition. They contain all the necessary
linguistic knowledge that is needed to properly use the word in a language. Word
meanings that are synonyms are separate structures (records) in RBN. They have their
own specification of information, including morpho-syntax and semantics. DWN is
organized around the notion of Synsets. Synsets are concepts as defined by Miller and
Fellbaum [4, 12, 13] in a relational model of meaning. They are mainly conceptual
units strictly related to the lexicalization pattern of a language. Concepts are defined
by lexical semantic relations. Typically in WordNet, information is provided for the synset as a whole and not for the individual word meanings. For example, in WordNet the synset has a single gloss but the different lexical units in RBN each have their own definition. From a WordNet point of view, the definitions of lexical units that belong to the same synset should thus semantically be compatible or synonymous.

Outside the lexicon, an ontology will provide a third layer of meaning. The Terms in an ontology represent the distinct types in a formal representation of knowledge. Terms can be combined in a knowledge representation language to form expressions of axioms. In principle, meaning is defined in the ontology independently of language but according to the principles of logic. In Cornetto, the ontology represents an independent anchoring of the relational meaning in WordNet. The ontology is a formal framework that can be used to constrain and validate the implicit semantic statements of the lexical semantic structures, both the lexical units and the synsets. In addition, the ontology provides a mapping of a vocabulary to a formal representation that can be used to develop semantic web applications.

In addition to the 3 data collections, a separate table of so-called Cornetto Identifiers (CIDs) is provided. These identifiers contain the relations between the lexical units and the synsets in the CDB but also to the original word senses and synsets in the RBN and DWN. In Figure 1, a single CID record is shown that contains the following records:

- C_form = form of the word in Cornetto
- C_seq = the sequence of sense number in Cornetto
- C_lu_id = the identifier of the lexical unit in Cornetto
- C_syn_id = the identifier of the synset in Cornetto
- R_lu_id = the identifier of the lexical unit in RBN from which it was derived
- R_seq_nr = the original sequence number or sense number in RBN
- D_lu_id = the identifier of the synonym in DWN
- D_syn_id = the identifier of the of the synset in DWN from which it was derived
- D_seq_nr = the original sequence number or sense number in DWN

Figure 1 shows an overview of the different data structures and their relations. The different data can be divided into 3 layers of resources, from top to bottom:

- The RBN and DWN (at the top): the original database from which the data are derived;
- The Cornetto database (CDB): the ultimate database that will be built;
- External resources: any other resource to which the CDB will be linked, such as the Princeton WordNet, WordNets through the Global WordNet Association, WordNet domains, ontologies, corpora, etc.

The center of the CDB is formed by the table of CIDs. The CIDs tie together the separate collections of LUs and Synsets but also represent the pointers to the word meaning and synsets in the original databases: RBN and DWN and their mapping relation. As you can see in this example, the identifiers of the record match the

1 For Cornetto, the semantic relations from EuroWordNet are taken as a starting point [18].
original identifiers of synsets and lexical units in the original databases. The CIDs are just administrative records. The Cornetto data itself are stored in the collection of LUs and the collection of Synsets.

The LUs will contain semantic frame representations. The frame elements may have co-indexes with Synsets from the WordNet and/or with Terms from the ontology. This means that any semantic constraints in the frame representation can directly be related to the semantics in the other collections. Any explicit semantic relation that is expressed through a frame structure in the LU can also be represented as a conceptual semantic relation between Synsets in the WordNet database. The Synsets in the WordNet are represented as a collection of synonyms, where each synonym is directly related to a specific LU. The conceptual relations between Synsets are backed-up by a mapping to the ontology. This can be in the form of an equivalence relation or a subsumption relation to a Term or an expression in a knowledge representation language. Finally, a separate equivalence relation is provided to one ore more synsets in the Princeton WordNet.

The Cornetto database provides unique opportunities for innovative NLP applications. The LUs contain combinatoric information and the synsets place these words within a semantic network. Figure 2 shows an example of this combination for several meanings of the word band: with meanings as musical band, as a tube or tire filled with air, a magnetic band, and a relationship. The semantic network position of
the word is depicted in separate WordNet fragments, relating the meanings to hypernyms, hyponyms and other related concepts. Above each fragment, we list the framelike combinatoric information that is given in RBN for these different meanings. A musical band is started, performs, a tube or tire is inflated, can leak, can blow, or you can fix it, etc. Each of these examples not only illustrates a typical conceptual usage or interaction but also the particular wording of it in Dutch. From these combinations, Dutch speakers immediately know what meaning of the word band applies. These typical examples can be used for the disambiguation of occurrences in text. Moreover, the same contexts can also be used for other words related to these meanings. We can easily extend the examples of band as a tire/tube to the hyponyms fietsband (bike tire) and autoband (car tire) and the examples of band as a relationship to the hypernym verhouding (affair) and relatie (relation).

<table>
<thead>
<tr>
<th>Combinatorics</th>
<th>Combinatorics</th>
<th>Combinatorics</th>
<th>Combinatorics</th>
</tr>
</thead>
<tbody>
<tr>
<td>in een band spelen (to play in a band)</td>
<td>de band oppompen (to pump air in a tire)</td>
<td>de band starten (to start a tape)</td>
<td>een goed sterk band (a good strong band)</td>
</tr>
<tr>
<td>een band oprichten (to start a band)</td>
<td>een band plakken (to fix a whole in a tire)</td>
<td>op de band opnemen (to record on a tape)</td>
<td>de banden verbreken (to break all bonds)</td>
</tr>
<tr>
<td>de band speelt (the band plays)</td>
<td>een lekke band (flat tire)</td>
<td>de band afspelen (to play from a tape)</td>
<td>een band hebben met iemand (to have a bond with s.o.)</td>
</tr>
</tbody>
</table>

**Fig. 2.** Combinatorics and semantics combined.

Another example, where combinatorics and semantic network relations are combined, relates to drinks. In Dutch, the preparation of drinks is usually referred to by the general verb maken (to prepare). However, in the case of koffie (coffee) and thee (tea), another specific verb is used: zetten. So, you typically use the phrases koffie zetten and thee zetten (to make coffee or tea) but you use the standard phrase limonade maken (to make lemonade) in Dutch. This example illustrates that conceptual combinations and constraints that are encoded in the WordNet or the ontology, do not explain the proper and most intuitive way of phrasing relations. The benefits of combining resources in this way are however only possible if the word meanings, representing concepts are properly aligned in the database. This is discussed in the next sections.
To create the initial database, the word meanings in the Referentie Bestand Nederlands (RBN) and the Dutch part of EuroWordNet (DWN) have been automatically aligned. The word *koffie* for example has 2 word meanings in RBN (*drink* and *beans*) and 4 word meanings in DWN (*drink*, *bush*, *powder* and *beans*). This can result in 4, 5, or 6 distinct meanings in the Cornetto database depending on the degree of matching across these meanings. This alignment is different from aligning WordNet synsets because RBN is not structured in synsets. For measuring the match, we used all the semantic information that was available. Since DWN originates from the Van Dale database VLIS, we could use the definitions and domain labels from that database. The domain labels from RBN and VLIS have been aligned separately by first cleaning up the labels manually (e.g., *pol* and *politiek* can be merged) and then measuring the overlap in vocabulary associated with each domain. The overlap was expressed using a correlation figure for each domain in the matrix with each other domain. Domain labels across DWN and RBN do not require an exact match. Instead, the scores of the correlation matrix can be used for associating them. Overlap of definitions was based on the overlapping normalized content words relative to the total number of content words. For other features, such as part-of-speech, we manually defined the relations across the resources.

We only consider a possible match between words with the same orthographic form and the same part-of-speech. The strategies used to determine which word meanings can be aligned are:

1. The word has one meaning and no synonyms in both RBN and DWN
2. The word has one meaning in both RBN and DWN
3. The word has one meaning in RBN and more than one meaning in DWN
4. The word has one meaning in DWN and more in RBN
5. If the broader term (BT) of a set of words is linked, all words which are under that BT in the semantic hierarchy and which have the same form are linked
6. If some narrow term (NT) in the semantic hierarchy is related, siblings of that NT that have the same form are also linked.
7. Word meanings that have a linked domain, are linked
8. Word meanings with definitions in which one in every three content words is the same (there must be more than one match) are linked.

Each of these heuristics will result in a score for all possible mappings between word meanings. In the case of *koffie*, we thus will have 8 possible matches. The number of links found per strategy is shown in Table 1. To weigh the heuristics, we manually evaluated each heuristics. Of the results of each strategy, a sample was made of 100 records. Each sample was checked by 8 persons (6 staff and 2 students). For each record, the word form, part-of-speech and the definition was shown for both RBN and DWN (taken from VLIS). The testers had to determine whether the definitions described the same meaning of the word or not. The results of the tests were averaged, resulting in a percentage of items which were considered good links. The averages per strategy are shown in Table 1.
Table 1. Results for aligning strategies

<table>
<thead>
<tr>
<th></th>
<th>Conf</th>
<th>Dev.</th>
<th>Factor</th>
<th>LINKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: 1 RBN &amp; 1 DWN meaning, no synonyms</td>
<td>97.1</td>
<td>4.9</td>
<td>3</td>
<td>9936</td>
</tr>
<tr>
<td>2: 1 RBN &amp; 1 DWN meaning</td>
<td>88.5</td>
<td>8.6</td>
<td>3</td>
<td>25366</td>
</tr>
<tr>
<td>3: 1 RBN &amp; &gt;1 DWN meaning</td>
<td>53.9</td>
<td>8.1</td>
<td>1</td>
<td>22892</td>
</tr>
<tr>
<td>4: &gt;1 RBN &amp; 1 DWN meaning</td>
<td>68.2</td>
<td>17.2</td>
<td>1</td>
<td>1357</td>
</tr>
<tr>
<td>5: overlapping hyperonym word</td>
<td>85.3</td>
<td>23.3</td>
<td>2</td>
<td>7305</td>
</tr>
<tr>
<td>6: overlapping hyponyms</td>
<td>74.6</td>
<td>22.1</td>
<td>2</td>
<td>21691</td>
</tr>
<tr>
<td>7: overlapping domain-clusters</td>
<td>70.2</td>
<td>15.5</td>
<td>2</td>
<td>11008</td>
</tr>
<tr>
<td>8: overlapping definition words</td>
<td>91.6</td>
<td>7.8</td>
<td>3</td>
<td>22664</td>
</tr>
</tbody>
</table>

The minimal precision is 53.9 and the highest precision is 97.1. Fortunately, the low precision heuristics also have a low recall. On the basis of these results, the strategies were ranked: some were considered very good, some were considered average, and some were considered relatively poor. The ranking factors per strategy are:

- Strategies 1, 2 and 8 get factor 3
- Strategies 5, 6 and 7 get factor 2
- Strategies 3 and 4 get factor 1

A factor 3 means that it counts 3 times as strong as factor 1. It is thus considered to be a better indication of a link than factor 2 and factor 1, where factor 1 is the weakest score. The ranking factor is used to determine the score of a link. The score of the link is determined by the number of strategies that apply and the ranking factor of the strategies. In total, 136K linking records are stored in the Cornetto database. Within the database, only the highest scoring links are used to connect WordNet meanings to synsets. There are 58K top-scoring links, representing 41K word meanings. In total 47K different RBN word meanings were linked, and 48K different VLIS/DWN word meanings. 19K word meanings from RBN were not linked, as well as 59K word meanings from VLIS/DWN. Note that we considered here the complete VLIS database instead of DWN. The original DWN database represented about 60% of the total VLIS database. VLIS synsets that are not part of DWN can still be useful for RBN, as long as they ultimately get connected to the synset hierarchy of DWN.
4 Aligning Manually RBN with DWN

The next alignment step is a manual process that consists of the editing of low-scoring and non-existing links between lexical units and synsets. We identified four major groups of problematic cases and defined editing guidelines for them, which will be presented in the following sections. Many of the low-scoring links turned out to be, not unexpectedly, links between lexical units and synsets of very frequent and highly polysemous words (section 4.1 and 4.2). Many of the non-links, i.e. a link between a synset and an automatically created and therefore empty lexical unit or vice versa, turned out to be between adjective synsets and lexical units (section 4.3). The fourth group, the multiword expressions, is different from the others, since for these automatic alignment could only be performed for few cases (section 4.4).

4.1 Frequent polysemous verbs and nouns

The low-scoring links within the group of verb synsets and lexical units and within the group of noun synsets and lexical units are in great deal due to the difference regarding the underlying principles of meaning discrimination which plays an important role in the alignment of synsets and lexical units. We defined a set of 1000 most frequent verbs in Dutch as a set to manually verify. For nouns, we defined a similar set of 1800 words that are most polysemous (4 or more word meanings). The matching of nouns is relatively straightforward and the manual process consists mainly of correcting the choices or cases where different meanings are given in the two resources. In the latter case, we either create a new synset or add the word to an existing synset as a synonym or we provide the information in the lexical unit that is lacking. Mappings for verbs are more complicated as will be explained below.

Characteristic for the verbal LUs is that they contain detailed information on verbal complementation, event structure and combinatoric properties. For the verb *behandelen* (to treat), the complementation patterns are:

- np: *iedem behandelen* (to treat someone)
- np, pp: *iedem aan/ voor/ tegen/met iets behandelen* (to treat someone for/with/… something)

In the representation of complementation patterns, all possible patterns are encoded. This may lead to a lot of patterns, but the result is a very explicit description of the syntactic behavior of the LU. As a rule, each pattern is worked out as an example in the combinatoric information. The corresponding event structure of *behandelen* contains the information that:

- this meaning of *behandelen* is an action verb.
- the subject np is the agent
- the object-np is the patient
- an optional pp-complement with *met* (with) is the instrument
- an optional pp-complement with *aan/voor/tegen* (for/with/against) is the theme
In the Dutch WordNet, these complements and roles are reflected in semantic relations:

- [causes] \[v\] genezen:2, beteren:1, herstellen:1 (*to recover*)
- [involved_agent] [n] arts:1; dokter:1 (*doctor*)
- [involved_patient] [n] zieke:1; patiënt:1 (*patient*)
- [involved_instrument] [n] hart-longmachine:1 (*heart-long machine*)
- [involved_instrument] [n] mitella:1, draagdoek:1 (*sling*)
- [involved_instrument] [n] geneesmiddel:1; medicijn:1 (*medicine*)
  etc.

As long as there is a one-to-one mapping from LUs and synsets, the features of the two resources will probably match. However, difficulties arise when the mapping is not one-to-one. Frequent verbs are often very polysemous. The RBN, as the source of the LUs, tries to deal with polysyem in a systematic and efficient way. The synsets are however much more detailed on different readings. As a result, in many cases there are more synsets than LUs. In combination with the detailed information on complementation, event structure and lexical relation, this results in interesting (and time consuming!) editing problems.

A typical example of an economically created LU in combination with a detailed synset is *aflopen* (*to come to an end, to go off (an alarm bell), to flow down, to run down, to slope down, etc.*). Input to the alignment are seven LUs and 13 synsets. Much of the asymmetry was caused by the fact that one of the LUs represents one basic and comprehensive meaning: *to walk to, to walk from, to walk alongside something or someone*. In DWN these are all different meanings, with different synsets. This is the result of describing lexical meaning by synsets; these three readings of *aflopen* obviously have a lot in common, but they match with different synonyms. Aligning the LUs and synsets leads to splitting the LUs and may lead to subtle changes in the complementation patterns, event structure and certainly to adapting and extending the combinatory information. Sometimes the LUs are more detailed. In that case a synset must be split, which of course gives rise to changes in all related synsets and to new sets of lexical relations.

In every day editing of frequent verbs it is often a problem to find out the exact meaning of a verb in a synset. This is certainly the case for isolated meanings without synonyms, forming a synset on their own, but also for frequent verbs with other frequent verb meanings in the synset. It does not help to know that *afspelen* (*to take place*) is in a synset with *passeren, spelen and geschieden* (*to happen, take place, occur*), all being ambiguous in the same way. These puzzles can often be solved by keeping a close watch on the lexical relations; especially instrument-relations are often of great help in disambiguating. However, it will be clear that alignment in the case of frequent verbs is hardly ever a matter of just confirming a suggestion for a mapping.
4.2 Nouns and semantic shifts

As is mentioned above, there are some differences in the lexicographical approach between the DWN and RBN resource for Cornetto. One important aspect is the economical distribution of LUs in the RBN, compared to the more extensive distribution of synsets. With regard to the nouns, this dissimilarity is mainly caused by the use of semantic shifts in the RBN.

A semantic shift can be defined as an aspect of a meaning that is closely connected to the central meaning. A shift can thus be seen as an extension of a meaning. Like in the RBN, the extension is not explicitly given but indicated, whereas DWN follows another approach to explicitly list these meanings. The RBN uses the semantic shift for groups of words that show the same semantic behavior. In the case of artikel (article) we find a LU with a shift that predicts that besides ‘text’, an artikel can also be an Artifact. This shift from Non-Dynamic to Artifact is also found consequently in LUs like reprint and script. There are about 30 different defined types of shifts that can occur in verbs, adjectives and nouns, like Process → Action in verbs and Dynamic → Non-dynamic in nouns. Due to the difference in approach, we expect that the matching of LUs from RBN to synonyms in DWN is more likely to be incorrect for all words labeled with a shift in RBN. We therefore decided to manually verify all the mappings for shifts. The vast majority of 4500 LUs with a semantic shift is found in nouns, on which we have decided to concentrate the manual work.

Because of the difference in approach, the DWN resource will have an extra synset for the meaning that is implied with a shift in the LU. If not, the presence of a shift might be a reason to create a new synset. This makes editing the LUs with a semantic shift a successful strategy to improve and extend the Cornetto database.

Editing an LU with a shift however, does not only mean splitting it and align it with the corresponding synset. Both resources show sometimes subtle differences in their description of a meaning, or a meaning happens to be missing in one of the resources. This means that if we want to edit the shift cases properly, we need to edit entries that contain an LU with a shift, and not just only the shift cases. This approach means that we aim at editing about 15,000 LUs and synsets, since most of the entries with a semantic shift are polysemous or will be so after editing. For these and some other edit related issues and decisions, we keep an edit log that will result in a final editing guideline.

All of this can be demonstrated by the word bekendmaking (announcement) that has one LU with a shift in RBN from Dynamic to Non-dynamic. This means that (in Dutch) an announcement can be a process and the result of this process. In DWN, we find a synset for each of these aspects, stating that the first one is a subclass of the SUMO term ‘Communicating’, and the second one is equivalent to ‘Statement’. We can see this as a good argument to split the LU and define the difference in terms of the definition and the semantic relations. In almost all of the dynamic and non-dynamic cases we use the following scheme to specify the relation and differences between both synsets and LUs (fig. 3 and 4):
The main advantage of editing shifts is the expansion and enrichment of the database. By creating a new LU for a synset we can add essential combinatory information and example sentences. When we add a new synset for a LU, we create new semantic relations, thus enriching the existing semantic structure of DWN. By editing clusters of the same shift type as e.g. dynamic → non-dynamic, we can ensure consistency at the same time. Note that the label shift will be kept in both LUs: in the original LU from the RBN and in the new LU which is the explicit meaning of the shift. In this way, we can always reconstruct the original RBN approach to store a single condensed meaning, or use the fact that there is a metonymic relation between these LUs. Furthermore, we express that there is a tight relation between these synsets.
4.3 Adjectives and fuzzy synsets

A considerable part of the adjectives is not successfully aligned by the automatic alignment procedures. This is especially due to the fact that adjective synsets have few semantic relations lacking hypernyms and hyponyms. By consequence, the automatic alignment strategies which involve broader and narrower terms, are in these cases not applicable.

Another problematic aspect of the adjective synsets is the fact that the automatically formed DWN adjective synsets are not – unlike the noun and verb synsets – edited and corrected manually. As a result, DWN adjective synsets have the following two characteristics:

- They are rather large and fuzzy often including words which are semantically related but not really synonymous, eg. Synset A: [dol, gek, dwaas, gaga (mad, crazy, foolish) achterlijk, gestoord (retarded, disturbed)]

  The synset needs to be splitted up in at least two new synsets: A1 [dol, gek, dwaas, gaga] ‘behaving irrational’ and A2 [gestoord , achterlijk] ‘affected with insanity’.

- They are often quite similar to each other, e.g. Synset B [dol, dwaas, maf, (mad, crazy, foolish) idioot (idiotic), krankzinnig (mad, insane)...]

  Although synset A includes other synonyms than synset B, they are both quite similar with respect to their meanings. They need to be partly merged into a new synset C [dol, dwaas, maf, gaga] ‘behaving irrational’ as is illustrated below (example 1).

  Of course, RBN’s lexical units - with numerous corpus based examples - can be helpful in solving these problems. However, it is already mentioned that the systematic and efficient way of word sense discrimination is often not consistent with the WordNet approach. For example, the following lexical unit kort (short) shows that the RBN does not always take into consideration possible synonym or hypernym relations.

  **Ex. 1. RBN kort (short).**

<table>
<thead>
<tr>
<th>LU</th>
<th>Resume</th>
<th>Syntax</th>
<th>Combinatorics</th>
</tr>
</thead>
</table>
| Kort (short) | of time and length | attr/pred | een korte dag (a short day),
|       |         |          | een korte vakantie (a short holiday),
|       |         |          | een korte broek (short trousers)
|       |         |          | kort haar (short hair) |

In this case the LU need to be split up in two LUs, distinguishing one temporal (with the combinations (1) and (2)) and one spatial sense (with the combinations (3) and (4)). Thus DWN’s semantic relations can be aligned correctly to the LUs (example 2):
Ex. 2. DWN *kort* (short).

<table>
<thead>
<tr>
<th>Synset</th>
<th>Synonyms</th>
<th>Semantic relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Of time</td>
<td>kort, kortdurend, kortstondig</td>
<td>Antonym: lang [1], langdurig (for a long period of time)</td>
</tr>
</tbody>
</table>

To be able to deal in a systematic way with these problems, we introduced the use of a semantic classification system for adjectives (Hundschnurser & Splett, Germanet). The classification regards the relation between the adjective and the modified noun. Adjectives are split up in 70 semantic classes which are organized in 15 main classes. In addition to this class, we also encode the ‘semantic orientation’ indicating a positive (+), negative (-) or neutral ( ) connotation of the involved adjectives. Since the semantic class and the semantic orientation hold for all synonyms within the synset, it is encoded at the level of the synset.

The following example presents the aligned version – after editing both LUs and synsets – of the word *dol* (crazy, fond). We distinguished three LUs and aligned them to synsets A, B and C respectively (example 3).

Ex. 3. *dol* (LUs and Synsets).

<table>
<thead>
<tr>
<th>LU</th>
<th>LU Resume</th>
<th>Syntax</th>
<th>Combinatorics</th>
<th>to Synset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>With a strong liking for</td>
<td>Predicative Fixed preposition: ‘op’ (on)</td>
<td>Dol op kinderen (fond of children)/ dol op chocola (fond of chocolate)</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>Offering fun and gaiety</td>
<td>Attr/pred</td>
<td>Een dolle avond (a merry evening)</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>Behaving irrational</td>
<td>Attr/pred</td>
<td>Het is genoeg om dol van te worden (it is enough to drive you crazy)</td>
<td>C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Synset</th>
<th>Synonyms</th>
<th>Semantic classification</th>
<th>Semantic orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>dol, verzot, gek, verrukt</td>
<td>CHARACTER/BEHAVIOUR</td>
<td>+</td>
</tr>
<tr>
<td>B</td>
<td>dol, uitgelaten, jolig <em>(crazy, jolly)</em></td>
<td>MOOD</td>
<td>+</td>
</tr>
<tr>
<td>C</td>
<td>dol, gek ,maf dwaas, gaga, geflipt <em>(crazy, foolish)</em></td>
<td>CHARACTER/BEHAVIOUR</td>
<td>-</td>
</tr>
</tbody>
</table>
4.4 Multimword units

Special attention is paid to the encoding and alignment of multimword units. The combinatoric information in the Cornetto Database is classified into the following types: (1) free illustrative examples, (2) grammatical collocations (3) pragmatic formula (4) transparent lexical collocations (5) semi-transparent lexical collocations (6) idioms and (7) proverbs. In RBN these combinations were not included in the macrostructure, but given within the microstructure of the meaning of one particular word contained in the expression. One of the objectives of Cornetto is to introduce part of them, i.e. the fixed combinations with a reduced semantic (and often syntactic) transparency - into the macrostructure thus making it possible to align them with a synset and via the synset with the ontology. We focus on those combinations which have a reduced semantic (and often syntactic) transparency and a reduced or lack of compositionality. The following 3 types meet the criterium set for this new group:

- Idioms: expressions with a reduced or lack of semantic transparency (e.g. stoken in een goed huwelijk (drive a wedge between two people), een rare snijboon (an odd person)).
- Proverbs: completely frozen sentences.
- Semi-transparent lexical collocations: these are lexical collocations of which one of the combination words has got a more specific meaning or less literal meaning than its basic meaning. Therefore the whole combination has a reduced semantic transparency. (systematische catalogus (systematic catalogue), open break (compound fracture), enkelvoudige breuk (simple fracture)).

The alignment of the idioms and proverbs multimword units with the synsets will be done exclusively by hand. The alignment of the semi-transparent lexical collocations with synset hierarchy will be performed in a semi-automatic way: in most cases the synset which includes the head of the NP (systematische catalogus) will be the hypernym synset of the multimword unit.

With regard to their semantic description, multimword units are regarded as a sequence of words that act as a single unit. Examples (2) and (3) illustrate the encoding of a lexical collocation and an idiom respectively. The description focuses on the semantics of the whole expression: each entry consists of a canonical form, its syntactic category, a textual form (if this applies), a lexicographic definition, information regarding its use if needed, one or more examples of the construction in context. The link to the synset is realised by a pointer to a cid-entry (c_cid_id) and links to the individual words of the combination are realised by pointers to single word lexical units (c_lu_id). Morpho-syntactic information relative to the individual words is included in the description of those particular words. The pointers to the individual words are pointers to lexical units. This seems contradictory - and sometimes is- with the uncompositionality of the multimword units. However, many multimword units are only semi-transparent and their syntactic and semantic behaviour is often related to their individual parts.
Ex. 4. Multiword unit *blinde muur* (blank wall).

<table>
<thead>
<tr>
<th>CanonicalForm</th>
<th>blinde muur (NP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sy-subtype</td>
<td>lexical collocation</td>
</tr>
<tr>
<td>meaningdescription</td>
<td>muur zonder ramen of deuren (a wall unbroken by windows or other openings)</td>
</tr>
<tr>
<td>C_LU_ID</td>
<td>muur (N) (wall)</td>
</tr>
<tr>
<td>C_LU_ID</td>
<td>blind (A) (blind)</td>
</tr>
<tr>
<td>Synset</td>
<td>[blinde muur] blank wall</td>
</tr>
<tr>
<td>Hyponym</td>
<td>[muur] (wall)</td>
</tr>
<tr>
<td>OntologicalType</td>
<td>StationaryArtifact (an artifact that has a fixed spatial location)</td>
</tr>
</tbody>
</table>

Ex. 5. Multiword unit *roomser dan de paus* (more Catholic than the Pope).

<table>
<thead>
<tr>
<th>CanonicalForm</th>
<th>roomser dan de paus (AdjP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sy-subtype</td>
<td>idiom</td>
</tr>
<tr>
<td>Sem-meaningdescription</td>
<td>overdreven principieel (extremely principled)</td>
</tr>
<tr>
<td>Prag-Connotation</td>
<td>pejorative</td>
</tr>
<tr>
<td>C_LU_ID</td>
<td>rooms (A) (catholic)</td>
</tr>
<tr>
<td>C_LU_ID</td>
<td>paus (N) (pope)</td>
</tr>
<tr>
<td>Synset</td>
<td>roomser dan de paus</td>
</tr>
<tr>
<td>Hyponym</td>
<td>principieel (principled), beginselvast (consistent)</td>
</tr>
<tr>
<td>OntologicalType</td>
<td>TraitAttribute</td>
</tr>
</tbody>
</table>

5 Alining synsets with ontology terms

A new relation is the mapping from the synset to the ontology. The ontology is seen as an independent anchoring of concepts to some formal representation that can be used for reasoning. Within the ontology, Terms are defined as disjoint Types, organized in a Type hierarchy where:

- a Type represents a class of entities that share the same essential properties.
- Instances of a Type belong to only a single Type: => disjoint (you cannot be both a cat & a dog)

Terms can further be combined in a knowledge representation language to form expressions of axioms (*you can be a watch dog & a bull dog*), i.e. the Knowledge
Interchange Format, KIF, based on first order predicate calculus and primitive elements.

Following the OntoClean method [6, 7], identity criteria can be used to determine the set of disjunct Types. These identity criteria determine the essential properties of entities that are instances of these concepts:

- **Rigidity**: to what extent are properties of an entity true in all or most worlds? E.g., a *man* is always a *person* but may bear a Role like *student* only temporarily. Thus *manhood* is a rigid property while *studenthood* is anti-rigid.

- **Essence**: which properties of entities are essential? For example, *shape* is an essential property of *vase* but not an essential property of the clay it is made of.

- **Unicity**: which entities represent a whole and which entities are parts of these wholes? An *ocean* or *river* represents a whole but the *water* it contains does not.

The identity criteria are based on certain fundamental requirements. These include that the ontology is descriptive and reflects human cognition, perception, cultural imprints and social conventions [21].

The work of Guarino and Welty [6, 7] has demonstrated that the WordNet hierarchy, when viewed as an ontology, can be improved and reduced. For example, roles such as AGENTS of processes are anti-rigid. They do not represent disjunct types in the ontology and complicate the hierarchy. As an example, consider the hyponyms of dog in WordNet, which include both types (races) like *poodle*, *Newfoundland*, and *German shepherd*, but also roles like *lapdog*, *watchdog* and *herding dog*. “Germanshepherdhood” is a rigid property, and a German shepherd will never be a Newfoundland or a poodle. But German shepherds may be herding dogs. The ontology would only list the rigid types of dogs (dog races): Canine $\Rightarrow$ PoodleDog; NewfoundlandDog; GermanShepherdDog, etc.

The lexicon of a language then may contain words that are simply names for these types and other words that do not represent new types but represent roles (and other conceptualizations of types). For example, English *poodle*, Dutch *poedel* and Japanese *pudoru* will become simple names for the ontology type: $\Leftrightarrow (\text{instance x PoodleDog})$. On the other hand, English *watchdog*, the Dutch word *waakhond* and the Japanese word *banken* will be related through a KIF expression that does not involve new ontological types: $\Leftrightarrow ((\text{instance x Canine}) \text{ and } (\text{role x GuardingProcess}))$, where we assume that GuardingProcess is defined as a process in the hierarchy as well. The fact that the same expression can be used for all the three words indicates equivalence across the three languages.

In a similar way, we can use the notions of Essence and Unicity to determine which concepts are justifiably included in the type hierarchy and which ones are dependent on such types. If a language has a word to denote a lump of clay (e.g. in Dutch *kleibrok* denotes an irregularly shaped chunk of clay), this word will not be represented by a type in the ontology because the concept it expresses does not satisfy the Essence criterion. Similarly, a word like *river water* (Dutch *rivierwater*) is not represented by a type in the ontology as it does not satisfy Unicity; such words are dependent on valid types. Satisfying the rigidity criterion, for example, is a condition for type status.
From this basic starting point, we can derive two types of mappings from synsets to the ontology [5, 19]:

- Synsets represent disjunct types of concepts, where they are defined as:
  - a. names of Terms;
  - b. subclasses of Terms, in case the equivalent class is not provided by the ontology
- Synsets represent non-rigid conceptualizations, which are defined through a KIF expression;

When we look at the different dogs in the Dutch WordNet then we see 3 types of hyponyms:

- bokser; corgi; loboor; mopshond; pekinees; pointer; spaniel (all dog races)
- pup (puppy); reu (male dog); teef (bitch)
- bastaard (bastard); straathond (street dog); blindengeleidehond (dog for blind people); bullebijter (nasty dog); diensthond (police dog); gashond (dog for detecting gas leaks); jachthond (hunting dog); lawinehond (aveline dog); schoothondje (lap dog);waakhond (watch dog)

The first group are names for dog races that are clearly rigid and disjunct. They represent names for Terms. The second group are words for male/female and baby dogs. They can be encoded in the same way as man, woman and child for humans. The third group refers to dogs with certain non-rigid attributes. They will thus not represent names for types but are related to the ontology by a mapping to the term Canine and the attribute that applies.

The KIF expressions are currently restricted to triplets consisting of the relation name, a first argument and a second argument. The default operator of the triplets is AND, and we assume default existential quantification of any of the variables, specified as a value of the arguments. Furthermore, we follow the convention to use a zero symbol as the variable that corresponds to the denotation of the synset being defined and any other integer for other denotations. Finally, we use the symbol \( \iff \) for full equivalence (bidirectional subsumption). In the case of partial subsumption, we use the symbol \( \Rightarrow \), meaning that the KIF expression is more general than the meaning of the synset. If no symbol is specified, we assume an exhaustive definition by the KIF expression. The symbol \( \iff \) applies by default.

The following simplified expression can then be found in the Cornetto database for the non-rigid synset \{waakhond\} (watchdog): (instance, 0 , Canine) (instance, 1 , GuardingProcess) (role, 0 , 1). This should then be read as follows:

\( \iff \)
(\( \text{The expression exhaustively defines the synset (instance, 0 ,Canine)} \)
(\( \text{Any referent of an expression with this synset as the head is also an instance of the type Canine (the special status of the zero variable),} \))
(\( \text{AND} \))
(\( \text{There exists an instance of the type Canine 0,} \))
(\( \text{AND} \))
(\( \text{(instance, 1 ,GuardingProcess)} \))
There exists an instance of the type GuardingProcess 0+1, AND
(role, 0, 1)
  The entity 0 has a role relation with the entity 1.

Other expressions that we use are:

Bokser (+, 0, Canine)
  The synset \{bokser\} is a rigid concept which is a subclass of the type Canine

hond (=, 0, Canine)
  The synset \{hond\} is a Dutch name for the rigid type Canine

The latter two relations are mainly imported from the SUMO mappings to the English WordNet. In the case of \{bokser\} it is manually added because it is dog race that is not in the English WordNet.

Another case of mixed hyponyms are words for water. In the Dutch WordNet there are over 40 words that can be used to refer to water in specific circumstances or with specific attributes. Water is in SUMO a CompoundSubstance just as other molecules. We can thus expect that the synset of water in Dutch matches directly to Water in SUMO, just as sand matches to Sand. However, water has 3 major meanings in the Dutch WordNet: water as liquid, water as a chemical element and a water area, while there are only two concepts in SUMO: Water as the CompoundSubstance and a WaterArea. In SUMO there is no concept for water in its liquid form, even though this is the most common concept for most people. Most of the hyponyms of water in the Dutch WordNet are linked to the liquid. To properly map them to the ontology, we thus first must map water as a liquid. This can be done by assigning the Attribute Liquid to the concept of Water as a CompoundSubstance:

\[(\mathrm{and})\]
\[(\mathrm{exists} \ ?L \ ?W)\]
\[(\mathrm{instance}, \ ?W, \ \mathrm{Water})\ ,\]
\[(\mathrm{instance}, \ ?L, \ \mathrm{LiquidState})\]
\[(\mathrm{hasAttributeinstance}, \ ?W, \ ?L) \ ]\]

In the Cornetto database, this complex KIF expression is represented by the slightly simpler relation triplets:

\[(\mathrm{instance}, \ 0, \ \mathrm{Water})\]
\[(\mathrm{instance}, \ 1, \ \mathrm{LiquidState})\]
\[(\mathrm{hasAttributeinstance}, \ 0, \ 1) \]

The hyponyms of water in the Dutch WordNet can further be divided into 3 groups:
• Water used for a purpose: theewater (for making tea), koffiewater (for making coffee), bluswater (for extinguishing fire), scheerwater (for shaving), afwaswater (for cleaning dishes), waswater (for washing), badwater (for bading), koelwater (for cooling), spoelwater (for flushing), drinkwater (for drinking)
• Water occurring somewhere or originating from: putwater (in a well), slootwater (in a ditch), welwater (out of a spring), leidingwater, gemeentepils, kraanwater (out of the tap), gootwater (in the kitchen sink or gutter), grachtwater (in a canal), kwelwater (coming from underneath a dike), grondwater, grondwater (in the ground), buiswater (on a ship)
• Being the result of a process: pompwater (being pumped away), smeltwater, dooiwater (melting snow and ice), afvalwater (waste water), condens, condensatiewater, condenswater (from condensation), lekwater (leaking water), regenwater (rain water), spuiwater (being drained for water maintenance)

In Figure 6, you find some of the mapping expressions that are used to relate these synsets to the ontology:

<table>
<thead>
<tr>
<th>theewater (tea water)</th>
<th>putwater (water at the bottom of well)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(instance, 0, Water)</td>
<td>(instance, 0, Water)</td>
</tr>
<tr>
<td>(instance, 1, Human)</td>
<td>(instance, 1, MineOrWell)</td>
</tr>
<tr>
<td>(instance, 2, Making)</td>
<td>(located, 0, 1)</td>
</tr>
<tr>
<td>(instance, 3, Tea)</td>
<td></td>
</tr>
<tr>
<td>(agent, 1, 2)</td>
<td></td>
</tr>
<tr>
<td>(resource, 0, 2)</td>
<td></td>
</tr>
<tr>
<td>(result, 3, 2)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>bluswater (water for extinguishing fire)</th>
<th>slootwater (in a ditch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(instance, 0, Water)</td>
<td>(instance, 0, Water)</td>
</tr>
<tr>
<td>(instance, 1, Human)</td>
<td>(instance, 1, SmallStaticWaterArea)</td>
</tr>
<tr>
<td>(instance, 2, Extinguishing)</td>
<td>(part, 0, 1)</td>
</tr>
<tr>
<td>(instrument, 0, 2)</td>
<td></td>
</tr>
<tr>
<td>(agent, 1, 2)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>drinkwater (drinking water)</th>
<th>leidingwater, gemeentepils, kraanwater (out of the tap)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(instance, 0, Water)</td>
<td>(instance, 0, Water)</td>
</tr>
<tr>
<td>(instance, 1, Drinking)</td>
<td>(instance, 1, Faucet)</td>
</tr>
<tr>
<td>(resource, 0, 1)</td>
<td>(instance, 2, Removing)</td>
</tr>
<tr>
<td>(capability, 0, 1)</td>
<td>(origin, 1, 2)</td>
</tr>
<tr>
<td></td>
<td>(patient, 0, 2)</td>
</tr>
</tbody>
</table>

Fig. 6. KIF-like mapping expressions for some hyponyms of the Dutch water.

Through the complex mappings of non-rigid synsets to the ontology, the latter can remain compact and strict. Note that the distinction between Rigid and non-Rigid
does not down-grade the relevance or value of the non-rigid concepts. To the contrary, the non-rigid concepts are often more common and relevant in many situations. In the Cornetto database, we want to make the distinction between the ontology and the lexicon clearer. This means that rigid properties are defined in the ontology and non-rigid properties in the lexicon. The value of their semantics is however equal and can formally be used by combining the ontology and the lexicon.

The work on the ontology is mainly carried out manually. The mappings of the synsets to SUMO/MILO are primarily imported through the equivalence relation to the English WordNet. We used the SUMO-WordNet mapping provided on: http://www.ontologyportal.org/, dated on April 2006. If there are more than one equivalence mappings with English WordNet, this may result in many to one mappings from SUMO to the synset. The mappings are manually revised traversing the Dutch WordNet hierarchy top-down so that we give priority to the most essential synsets. Furthermore, we will revise all synsets with a large number of equivalence relations or low-scoring equivalence relations. Finally, we also plan to clarify the synset-type relations for large sets of co-hyponyms as shown above for water. This work is still in progress. We do not expect this to be completed for all the synsets in this 2-year project with limited funding but we hope that a discussion on this topic can be started by working out the specification for a number of synsets and concepts.

6 Conclusion

In this paper, we presented the Cornetto project that combines three different semantic resources in a single database. Such a database presents unique opportunities to study different perspectives of meaning on a large scale and to define the relations between the different ways of defining meaning in a more strict way. We discussed the methodology of automatic and manual aligning the resources and some of the differences in encoding word-concept relations that we came across. The work on Cornetto is still ongoing and will be completed in the summer of 2008. The database and more information can be found on:

http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm

Acknowledgments

This research has been funded by the Netherlands Organisation for Scientific Research (NWO) via the STEVIN programme for stimulating language and speech technology in Flanders and The Netherlands.

References

CWN-Viz : Semantic Relation Visualization in Chinese WordNet

Ming-Wei Xu1, Jia-Fei Hong2, Shu-Kai Hsieh3, and Chu-Ren Huang1

1 Institute of Linguistics Academia Sinica
No. 128, Section 2, Academia Road 115, Taipei, Taiwan R.O.C
2 National Taiwan University, Graduate Institute of Linguistics
No. 1, Sec. 4, Roosevelt Road 106, Taipei, Taiwan R.O.C
3 Department of English, National Taiwan Normal University
NO. 162, Section 1, He-ping East Road 106, Taipei, Taiwan R.O.C
javaxu@gmail.com
{jiafei, churen}@gate.sinica.edu.tw
shukai@gmail.com

Abstract. This paper reports our recent work to use visualization to present semantic relation for Chinese WordNet. We design a visualization interface, named CWN-Viz, based on “TouchGraph”. There are three import design features of this visualization interface: First, visualization is driven by word-form, the most intuitive lexical search unit in Chinese. Second, the CWN-Viz allows visualization of bilingual semantic relations by incorporating Sinica BOW (Bilingual Ontological WordNet) information. Third, the semantic distance of each relation is calculated and used in both clustering and visualization.

Keywords: Chinese Lexical Knowledgebase, Chinese WordNet, Semantic Relation, Visualization.

1 Introduction

George Miller [1] thought that synonym sets can be used to anchor the representation of the lexicon concepts and describe the (mental) lexicon. This was the original motivation to construct WordNet. Recently, there are many research teams to deal with semantic relations by the knowledge-base of WordNet. With the bilingual Chinese-English WordNet Sinica BOW [14], the Chinese WordNet Group at Academia Sinica has been worked on dividing and analyzing Chinese lemmas, senses and their semantic relations.

Regard semantic relations as the foundation, the establishment of the reliability and relevant problems for the Chinese WordNet. In these data, we would put emphasis on presenting visualization system and show their semantic relations of whole senses. Finally, we will make use of calculating principle in order to cluster related senses and present several different similar synonym groupings.
Based on these viewpoints, we will obtain several different similar synonym groupings. And then, we can auto-create more and more different similar synonym groupings. In the same time, we also can establish huger Chinese semantic relation WordNet.

2 Chinese WordNet

The relationship between language and meaning has always been one of the problems that people think about all the time since human languages and cultures started. “Word” is the minimum meaning unit in the human languages. Dividing and expressing the meaning of a “word” and the interaction between accessing senses and expressing knowledge are the most fundamental researches. Sense division and expression need to be established basing on a complete set of lexical semantic theories and on the basic frames of ontology. In the Institute of Linguistics at Academia Sinica, under the direction of Chu-Ren Huang, the Chinese WordNet Group (CWN Group) have been working on the research called “Chinese meaning and Sense,” This research provides a explicit data by analyzing Chinese lexical senses manually.

Huang [2] proposed the criteria and operational guidelines for process of dividing lexical senses. Besides, the criteria are also the basis for constructing a Chinese sense knowledgebase and codifying the Dictionary of Sense Discrimination. The entries in the Dictionary of Sense Discrimination can be singular word, two words or multiple words and are limited to the common words in modern Chinese. As shown in Fig. 1, this dictionary lists the complete information of each entry, including the phonetic symbols (Pinyin and National Phonetic Alphabets), sense definition, corresponding synset, part-of-speech (POS), example sentences and explanatory notes.

WordNet was the first application that integrated all the different linguistic elements, such as sense, synonyms (synset), semantic relation and examples. And they have designed an online interface for their current version— WN3.0. In here, we want to have the comparison of the data structures between WN3.0 and Chinese WordNet.

As shown in Fig.2, in WN, the information of each lexicon is simply listed its sense, synonyms (synset), semantic relations with other English synsets and some examples. As shown in Fig.3, the data structure in Chinese WordNet is divided into the parts of Chinese lexicon, Chinese lexical knowledge, and Linking to related language resource. In the section of Chinese lexical knowledge, similar to WN, each lexicon is accompanied with its sense, domain, English synset and synset number from WN1.6, semantic relations with other lexicons and example sentences, but common lexicons do not belong to any domain. Synset is used to link the other English resources in the system. The uniqueness and value of Chinese WordNet are to present the analysis result of after the Sense division and link the result with some English resources.
Fig. 1 Example of Chinese Lexical Lemma.

Fig. 2 Data Structure of WN.
In other words, Chinese WordNet equips the function for searching cross-lingual information because Chinese WordNet integrates with the varied English resources, such as WN and SUMO (Suggested Upper Merged Ontology). Therefore, users can easily compare the concept differences between Chinese and English via Chinese WordNet.

Normally, a text-searching system can only search a target document that include the words as its content, but it cannot be searched by the word senses or other relevant information of the words. Such searching function obviously cannot fulfill the requirements of linguistic researches. Through the experience from the pervious relevant linguistic researches, it is possible to collect a lot of varied lexical information, which is considered and required for the linguistic research purpose.

This research is based on analyzing the Chinese lexicons. After conscientiously analyzing and researching, each Chinese lexicon accompanies with the information about lemma, sense, domain, sense definition, semantic relation, synset, example sentences, explanatory note and so on. Such conscientious analysis is helpful to preserve the lexical knowledge systematically and fulfill the different needs for the relevant linguistic researches.

From the beginning of year 2003 to the beginning of September 2007, totally CWN Group had analyzed 6653 lemmas and identified 16693 senses. In order to refer to the data from the cumulative knowledgebase clearly, in 2005, we started to use "edition" to divide the content in Chinese WordNet. The data cumulating until the end of year 2003 was the first edition. The research result cumulating until the end of year 2004 was the second edition. The third edition was the data cumulating until the end of year 2005 and was presented to the public on April, 2006. Now, the fourth edition was our
sense division cumulating until the end of year 2006 and has already published on April, 2007.

3 Visualization

In this study, we want to follow a well-accepted design paradigm to create a working prototype of a visualization suite for Chinese WordNet which has sense division of the data, as well as the ability to focus on specific synsets of interest and get some details.

Before, information visualization technique was applied in computer science or biology science and shown relationship constructions of large data. Ware [3] suggests five advantages of effective information visualization:

1) Comprehension: Visualization provides an ability to comprehend huge amounts of data.
2) Perception: Visualization reveals properties of the data that were not anticipated.
3) Quality control: Visualization makes problems in the data immediately apparent.
4) Detail + Context: Visualization facilitates understanding of small scale features in the context of the large scale picture of the data.
5) Interpretation: Visualization supports hypothesis formation, leading to further investigation.

Recently, this technique was applied in NLP gradually such as Visual Thesaurus [15] and WordNet Explorer [4]. These studies only focus on usage and showing partial information of WordNet. However, so few studies focus on WordNet for designing, completely showing the relation between lemma and senses.

Following similar design, we construct a visualization interface, named CWN-Viz, based on “TouchGraph”, a pen source graph layout system that has been extended and adapted to suit our requirements. The interface can completely show all lemmas, senses, and semantic relations for a word form recorded in Chinese WordNet. An important design feature of this interface is the iconic representation of semantic distance. We propose a set of principles to measure the distance of each semantic relation by calculating the node distances connected by that semantic relation. The distance information is used in clustering of related information and presented in this visualization interface such as below Fig.4 to Fig.6.
Fig. 4 The basic visualization construction.

Fig. 5 Semantic relations of visualization construction.
4 Analysis

According to above CWN analysis and visualization construction, we develop the calculating principle in order to cluster for different lemmas and several senses and cluster their semantic relations. First, we regard a keyword root as a center node and extend two levels. Based on the first level nodes to calculate sub-roots of the sub-trees such as A, B, C, and D, respectively expand and make several sub-trees. Through calculating the numbers of semantic relations of these sub-trees, we can evaluate the relationship score for each sub-tree. When we calculate the relationship score for each sub-tree such as Fig.7, we would like present the calculating matrix for each cluster such as Table1. Finally, we select the most nodes of the numbers semantic relations until all nodes were selected. Consequently, we can obtain Cluster1 for [A, B] and Cluster2 for [C, D].
Following these constructs, we take CWN analysis into this model and the results were shown as:

**Table 1.** The calculating matrix for each cluster.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 7 The clusters of semantic relations of visualization construction.
We name this visualization system as CWN-Viz. We make the idea of CWN analysis and visualization construction into the website version. The available website address is http://cwn.ling.sinica.edu.tw/cwnviz/. In this website, there are several resources such as Chinese WordNet analysis, SinicaBOW, Sinica Corpus, WordNet 1.6 version, WordNet 1.7 version, the textbook of elementary school and Chinese Dictionary of Ministry of Education. So far, there are 3000 lemma, 5568 senses, 1260 facets, 9828 nodes and 11978 relations in this website.

In here, we want to explain briefly for Fig.8 and Fig.9:

正 zheng3 right
正 2 The second lemma of 正
正 2(0620) The sixth senses of 正 2
美麗 mei3 li4 beautiful
美麗(0101) The first meaning facet of the first sense of 美麗
In this system, blue line represents synonym, red line represents antonym, green line represents hyper or hypo and yellow line presents near synonym.

In this website, we can look up some words and check their semantic relations. Also, we can select different resource to show different content in here. The website is such as Fig.9 and if we select Chinese WordNet analysis or Sinica BOW as the content resource, we can get some information in here such as Fig.10 and Fig.11.

Fig. 9 The website of visualization construction.
Fig. 10 The visualization of Chinese WordNet analysis for keyword.

Fig. 11 The visualization of Sinica BOW for keyword.
5 Conclusion

WordNet by nature is a representation of a network of lexical relations in the mental lexicon. It is no surprising that many previous works have attempt to make these relational network more explicit and easier to understand and process by visualization tools. Work such as VISdic [5] and University of Toronto’s WordNet Explorer all made substantial contribution to this line of study. Our current study aims integrate sharable tools from the visualization field with WordNet studies. We also design the tool such that cross-lingual lexical semantic relations can be visualized in parallel. It is hoped that such a tool will be facilitate linguists’ use and understanding of WordNets, as well as facilitate international initiatives such as the Global WordNet Grid. One area that has not been explored sufficiently so far is the calculation of semantic distances according to visualized relations. According to above analysis, calculating principle and cluster, we can obtain visualizations such as Fig.8 to Fig.11. Ideally, we will create more related clusters and show their semantic relations. Further, we want to show these related clusters by different color and link related senses by different line in the interface. Also we provide more details for each sense. Fig.12 will be presented our prototypical type for this study.

![Fig. 12 The interface for semantic relations of visualization with sense division.](image)
References


Resource

   http://BOW.sinica.edu.tw
16. Chinese WordNet
   http://cwn.ling.sinica.edu.tw/
17. Sinica Corpus: Academia Sinica Balanced Corpus of Mandarin Chinese
   http://www.sinica.edu.tw/SinicaCorpus/
18. TouchGraph.
   http://www.touchgraph.com/index.html
19. WordNet
   http://wordnet.princeton.edu/
20. WordNet Explorer (Prototype)
   http://www.cs.toronto.edu/~ccollins/wordnetexplorer/index.html#controlpanel
Using WordNet in Extracting the Final Answer from Retrieved Documents in a Question Answering System

Mahsa A. Yarmohammadi, Mehrnoush Shamsfard, Mahshid A. Yarmohammadi, and Masoud Rouhizadeh

Natural Language Processing Laboratory, Shahid Beheshti University, Tehran, Iran
m_yarmohammadi@std.sbu.ac.ir, m-shams@sbu.ac.ir, yarmohammadi@modares.ac.ir, m.rouhizadeh@mail.sbu.ac.ir

Abstract. In this project we propose a model for answer extraction component of a question answering system called SBUQA. Methods which extract answers based on only the keywords ignore many acceptable answers of the question. Therefore, in our proposed system we exploit methods for meaning extension of the question and the candidate answers and also make use of ontology (WordNet). In order to represent the question and the candidate answers and comparing them to each other, we use LFG - Lexical Functional Grammar, a meaning based grammar that analyses sentences in a deeper level than syntactic parsing- and obtain the f-structure of the sentences. We recognize the appropriate f-structure pattern of the question and based on that, the f-structure patterns of the answers. Then, the answer's pattern is matched to the pattern of the candidate answer by the proposed matching method, called extended unification of f-structure. Finally, the sentences which acquire the minimum score to be offered the user are selected; the answer clause is identified in them and displayed to the user in descending order.

Keywords: Question answering systems, Answer extraction, Information retrieval, Natural language processing

1 Introduction

Information Retrieval (IR) systems are the systems in which the user enters his/her query in form of separated keywords, and the search engine retrieves all the related documents from its knowledge base in a limited time. Most of retrieved documents are just syntactically –and not semantically- related to the user query.

Users need exact and accurate information and don't like to waste their time by reading all retrieved documents to find the answer, and IR systems are not sufficient for this reason. So, a new kind of IR named Question Answering (QA) systems appeared from the late 1970's and early 1980's. In these systems, the user ask his/her natural language question with no restriction in its syntax or semantic. The system is responsible for finding the exact, short, and complete answer at the shortest possible time. To do this, QA system applies both IR and NLP techniques. In this article, we propose a method for answer extraction component of a question answering system.
Methods which extract answers based on only the keywords ignore many acceptable answers of the question. Therefore, in our proposed system we approach to methods for meaning extension of the question and the candidate answers and also make use of ontology (WordNet). In order to match the question and the candidate answers we use Lexical Functional Grammar and the benefits of its functional-structure representation, and propose a unification algorithm. The question answering system we have designed and implemented for this purpose is named SBUQA\(^1\).

In the following sections, we first introduce the overall architecture of QA systems, Lexical Functional Grammar and advantages of using this grammar in QA systems. Then, we present SBUQA. Finally, we describe the implementation and evaluation of SBUQA and mention future works.

## 2 QA Systems Architecture

A question answering system that is based on searching among a set of documents, is composed of three main components [3]:

1) Getting user question and processing it that converts question given in natural language to query (or queries) to be used in the information retrieval component.

2) Retrieving documents (search engine), that retrieves related documents - documents that probably contain the answer- based on the input query.

3) Extracting final answer from retrieved documents, that extracts sentence or expression or text part containing the answer among documents.

A survey on existing question answering systems reveals that all of these systems include the above three components, but use different methods performing them.

## 3 Lexical Functional Grammar

based formalism which analyses sentences at a deeper level than syntactic parsing [1]. LFG views language as being made up of multiple dimensions of structure. The primary structures that have figured in LFG research are the structure of syntactic constituents (c-structure) and the representation of grammatical functions (f-structure). For example, in the sentence "The old woman eats the falafel", the c-structure analysis is that this is a sentence which is made up of two pieces, a noun phrase (NP) and a verb phrase (VP). The VP is itself made up of two pieces, a verb (V) and another NP. The NPs are also analyzed into their parts. Finally, the bottom of the structure is composed of the words out of which the sentence is constructed. The f-structure analysis, on the other hand, treats the sentence as being composed of attributes, which include features such as number and tense or functional units such as subject, predicate, or object. This type of analysis is useful in that it is a more abstract representation of linguistic information than a parse tree structure. In addition, long distance dependencies, which are very common in interrogative sentences and fact seeking questions, are resolved in order to have a complete and correct f-structure.

---

\(^1\) Shahid Beheshti University Question Answering system
analysis. This makes LFG analysis useful for QA tasks because it identifies the focus of the question and also which functional role (e.g. subject and object) the focus can fulfill. LFG analysis also provides valuable information for the detailed interpretation of complex questions which can potentially form a significant component in answering them correctly.

In our proposed system, we use LFG f-structure for representing and matching the question and its candidate answers.

4 SBUQA System

From overall view, the third component of a QA system, gets the user question and a set of retrieved text documents as the input, and shows user the answer(s) extracted from the document set as the output. The document set is composed of text documents that are the output of the second component of the QA system (search engine). Figure 1, shows the architecture of SBUQA.

![Figure 1. The architecture of SBUQA](image)

The components of the system and relationships between them are described in the following.
4.1 Getting Question and Building its f-structure

The system gets the user natural language question and sends it to the LFG parser to build its f-structure representation. This representation makes one of the inputs of “Building f-structures of Answer Instances” component.

4.2 Document Preprocessing and Building its Sentences f-structures

Documents retrieved by the search engine are saved as text files in the system’s document bank. These documents are preprocessed by JAVARAP\textsuperscript{2} tool, so that the sentences are separated and pronouns are replaced by their referents. Then, the sentences of each document are sent to the LFG parser to represent as f-structures (called $f_{SC}$). These representations make one of the inputs of “Extended Unification Algorithm and Answer Scoring” component.

4.3 Building f-structures of Answer Instances

We have defined some templates -represented in f-structure format- for Wh questions (called $f_{STQ}$). Question f-structure is compared to $f_{STQ}$s and is matched with one of them, say X. For each $f_{STQ}$, we have defined one or more answer template(s) represented in f-structure format (called $f_{STA}$). $f_{STA}$s of the X (the $f_{STQ}$ matching with user question) are filled with question keywords and make answer instances (called $f_{SA}$). These instances are the other inputs of “Scoring the answer” component. Question and answer templates are described in section 4.5.

4.4 Extended Unification Algorithm and Answer Scoring

$f_{SC}$ of each sentence of document (input from document preprocessing component) is compared to $f_{SA}$s (input from “Building f-structures of Answer Instances” component). The comparison is done by the Extended Unification Algorithm, introduced in section 4.5 and sentences are scored. Finally, sentences acquiring the score more than a defined threshold are selected, ordered by their scores, and shown to the user.

4.5 Question and Answer Templates

We define templates for questions and related answers based on the following categorization of English sentences [2]:

1. Active sentences with transitive verb, containing subject, verb and optional object.
2. Active sentences with intransitive verb, containing subject and verb.

\textsuperscript{2} http://www.comp.nus.edu.sg/~qiul/NLPTools/JavaRAP.html
3. Passive sentences, containing adverb which is representative of the omitted subject, verb and object.
4. Sentences containing copula.

Each of the above sentences can contain some complement or adverb.

We tried to define question templates for wh questions so that they can cover all standard forms of questions. We divided five types of wh questions in two groups: one group containing who, where and when questions and the other group containing what and which questions. Then by considering four forms described above, we defined the following templates for wh questions:

### 4.5.1 Who, Where, and When Questions

The following four templates are the question templates for forms 1 and 2 (active sentences). These templates are numbered from I to IV. The FOCUS property indicates the type of wh question. The PRED property indicates the main verb of the sentence. TENSE indicates the tense of the verb, OBJ represents the object, ADJ represents the adjuncts especially adverb, SUBJ indicated the subject and XCOMP represented the complement. The MODAL property in template II indicates that the sentence contains a modal verb.

Template I is used for active interrogatives that contains only the main verb and template II covers active interrogatives that contain modal verb in addition to main verb.

**Template I**

\[
\text{FOCUS} \quad \begin{cases} \text{PRED} \quad \text{verb1} \quad \text{PRO} \\ \text{TENSE} \quad tl \\ \{\text{OBJ} \quad \text{obj1}\} \\ \{\text{ADJ} \quad \text{adj1}\} \\ \text{SUBJ} \quad \{\text{PRED} \quad \text{I}\} \end{cases}
\]

**Template II**

\[
\text{FOCUS} \quad \begin{cases} \text{PRED} \quad \text{can, might, must...} \\ \text{MODAL} \quad \text{proposition with*} \end{cases}
\]

**Template III**

\[
\text{FOCUS} \quad \begin{cases} \text{PRED} \quad \text{do, have} \\ \text{TENSE} \quad tl \\ \{\text{OBJ} \quad \text{obj1}\} \end{cases}
\]

**Template IV**

\[
\text{FOCUS} \quad \begin{cases} \text{PRED} \quad \text{have} \\ \text{TENSE} \quad tl \\ \{\text{OBJ} \quad \text{obj1}\} \\ \{\text{ADJ} \quad \text{adj1}\} \end{cases}
\]

Templates III and IV covers interrogatives that use auxiliary verb do or have.
Template for passive interrogatives (form 3) is as template V. The PASSIVE property with + represents the passive sentence. Template for copula interrogatives (form 4) is as VI.

For each of the defined question templates (fTQA), one or more answer templates (fTA) are defined. As mentioned in section 4-2-3, if the user question matches with one of the fTQA, the fTA for that are filled with words of the question in order to make answer instances and are used in the extended unification algorithm with sentences of candidate answer.

Template for answer of active interrogatives – that matches with forms 1 and 2- are as the followings:

Answer template I_A is defined based on question template I_Q. As the same, the answer template II_A is defined for question template II_Q, III_A for III_Q and IV_A for IV_Q.
A question in active form can be answered by a passive sentence; so the template for passive answer ($V_A$) is added to answer templates for form 1 and 2. Answer templates for questions that match form 3 (passive sentences) are as follows:

Answer template $V_A$ is defined based on question template $V_Q$. Here it is possible that the question in passive form have answer in active form. Hence four answer templates for active sentences ($I_{A}$, $II_{A}$, $III_{A}$, $IVA$) are also added to answer templates of form 3.

Answer template of the questions matched with form 4 (copula) is like the following. This template is defined based on the question template $VI_Q$.

4.5.1 What, and Which Questions

Templates offered for who, where and when questions all are applicable to what and which questions. In addition to them, defining some additional templates for these two types of questions is possible. If a word (or expression) appears after words which and what, that actually is the topic of the question, expression

is replaced by

in all of the previous templates.

Answer templates for what and which question, are the same templates defined for who, where and when questions.
4.6 Extended f-Structure Unification

Answer extraction is a result of unifying the f-structure of the candidate answer and instance of the answer (that is generated based on the question). Experiments show that unification strategy based on exact matching of values is not sufficiently flexible [5]. For example, sentence "Benjamin killed Jefferson." is not answer to question "Who murdered Jefferson?" by exact matching. In our proposed system, we considered approximate and semantic matching in addition to exact and keyword-based matching. Approximate matching is performed by ontology-based extended comparison between different parts of the question template and the candidate answer template (including subj, obj, adjunct, verb and ...) and comparing of their types.

In Our unification algorithm, by slot we mean various parts of templates (including subj, obj, adjunct, verb and ...) accompanied with their types, and by filler we mean values (instances) that slots are filled with them.

For determining the level of matching between fA and fC, we proposed a hierarchical pattern based on the exact matching, approximate matching, or no matching between slots and fillers of the two structures. Levels of scoring the candidate answer based on the matching of fA and fC is as follows:

A) Existence of all slots of fA in fC and
   - Exact matching of the fillers.
   - Approximate matching of the fillers.
   - No matching of the fillers.

B) Existence of all slots of fA in fC, plus additional slots in fC and
   - Exact matching of the fillers.
   - Approximate matching of the fillers.
   - No matching of the fillers.

C) Existence of some (no all) slots of fA in fC and
   - Exact matching of the fillers.
   - Approximate matching of the fillers.
   - No matching of the fillers.

D) Existence of some slots of fA in fC, plus additional slots in fC and
   - Exact matching of the fillers.
   - Approximate matching of the fillers.
   - No matching of the fillers.

For approximate matching of the fillers, a hierarchical pattern as the following is defined:

Approximate matching of the fillers of type verb:
- Value of fC is synonym of value of fA.
- Value of fC is troponym of value of fA.
- Value of fA is troponym of value of fC.
- Value of fA is hypernym of value of fC.
- Value of fC is hypernym of value of fA.

Approximate matching of the fillers of other parts of the sentence (obj, subj, adjunct):
- Value of fC is synonym of value of fA.
- Value of \( f_{A} \) is hypernym of value of \( f_{C} \).
- Value of \( f_{C} \) is hypernym of value of \( f_{A} \).
- Value of \( f_{C} \) is meronym of value of \( f_{A} \).
- Value of \( f_{C} \) is holonym of value of \( f_{A} \).

5 SBUQA Implementation & Evaluation

SBUQA is implemented using Java programming language as a Java Applet and is developed in Oracle JDeveloper 10g IDE. The software is composed of several functions and built-in or user-defined libraries. One of the most important libraries used, is JWNL\(^3\), the Java API for WordNet.

The software package, interacts with some available tools such as Probabilistic LFG f-structure parser – developed by National Centre for Language Technology (NCLT) in Dublin City university- and JAVARAP anaphora resolution tool that is used in document preprocessing component.

User enters his/her question via the user interface. User can select option "Find answer from the document set" to limit the searching of answer in a previously prepared document set. For preparing this document set, user question is given to google search engine and some online question-answering systems (AnswerBus, Start, Ask, …) and some first documents retrieved by these systems, are saved as text files in the system's document base. We have not designed any graphical user interface for this process yet and user must prepare the document set manually. If user selects the second option, "Find answer from the following input", a text area is opened and user can type his text and final answer is searched among this text.

After entering the question and choosing the source, user clicks Ask button and the process of searching and extracting final answer starts. After this process, possible extracted answer(s) are displayed in "Possible Answers" section in decreasing order of assigned scores.

A sample of running the software for question "Where was George Washington born?" is shown in figure 2.

For evaluating the operation of the proposed system in finding final answer, we selected 10 questions (two questions of each type of where, who, what, when and which) from TREC question set. These questions selected in a way that cover various kinds of question templates. For each question, sentences of documents retrieved by google search engine and AnswerBus, Start and Ask online question-answering systems are extracted. Level of matching of these sentences with answer templates, are determined using the implemented tool. If the sentence matches with one of the templates, the answer part is extracted from the sentence using the tool and correctness or incorrectness of it is determined.

Based on the evaluation, the precision of matching level A is equal to 0.78, level B is equal to 0.67, level C is equal to 0.50, and the precision of level D is equal to 0.33. The recall of the system is equal to 0.54.

\(^3\) Java WordNet Library
Also, each test question is evaluated based on some QA systems metrics: First Hit Success (FHS), First Answer Reciprocal Rank (FARR), First Answer Reciprocal Word Rank (FARWR), Total Reciprocal Rank (TRR), and Total Reciprocal Word Rank (TRWR).

Possible values for FHS, FARR, FARWR, and TRWR are in the range of 0 to 1 and the ideal value in an errorless QA system is equal to 1. Possible values for TRR are from 0 to $\infty$ (it doesn’t have an upper bound). The greater value for TRR, indicates that more correct answers are extracted. System evaluation based on these metrics, gives 0.78 for FHS and FARWR, 0.82 for FARR, 1.33 for TRR, and 0.72 for TRWR, that are good (greater than the average) values.

6 Conclusion and Future Works

The SBUQA system that is proposed for the third component of question-answering systems, operates based on f-structure of the question and candidate answers and extended unification based on ontology (WordNet). According to the evaluation measures of question-answering systems, the SBUQA system resulted a good (better than average) operation in retrieving final answer.

The proposed system is designed for wh questions in open domain. Further extensions can cover yes/no questions and other types of questions. f-structure is beyond some shallow representations that are dependent to language. Although languages are different in shallow representations, but they can be represented by the
same (or very similar) syntactic (and semantic) slot-value structures. This feature of $f$-structure, make it possible to use the algorithms introduced in the proposed system for other languages including Persian. Now it is not possible to implement the system for Persian, because lack of usable and available tools for processing Persian language (such as parser and WordNet ontology for Persian). But we consider this as future extensions of the system.

References

Towards the Construction of a Comprehensive Arabic WordNet

Hamza Zidoum
Sultan Qaboos University
Department of Computer Science
Po Box 36, Al Khod 123 Oman
zidoum@squ.edu.om

Abstract. Arabic is a Semitic language spoken by millions of people in more than 20 different countries. However, not much work has been done in the field of online dictionaries or lexical resources. WordNet is an example of a lexical resource that has not been yet developed for Arabic. WordNet, a lexical database developed at Princeton University, has seen life 15 years ago. Ever since then, it has proved to be widely successful and extremely necessary for today’s demands for processing natural languages. Accordingly, the motivation of developing an Arabic WordNet became strong. In this paper we tackle some of the challenges inherent to constructing an Arabic lexical reference system. The paper goes through some solutions adopted in existing WordNets and presents justifications for adopting the Arabic WordNet's (AWN) philosophy. We address the nominal part of Arabic WordNet as the first step towards the construction of a comprehensive Arabic WordNet. The nominal part means nouns as a part of speech.

Key Words: WordNet, Synsets, Arabic Processing, Lexicon

1 Introduction

WordNet is an online lexical reference system, which groups words into sets of synonyms and records the various semantic relations between these synonym sets. It has become an important aspect of NLP and computational linguistics [1]. Many WordNets were constructed for different languages [11, 13, 14, 15, 16, 19, 20, 22, 24, 29, and 30].

WordNet design is inspired by current psycholinguistic theories of human lexical memory [30]. It groups words into sets of synonyms called synsets, which are the basic building blocks of WordNet [30]. A synset is simply a set of words that express the same meaning in at least one context [1, 30]. WordNet also provides short definitions, and records the various semantic relations between these synonym sets [6]. Nouns, verbs and adjectives are organized into synset, each representing one underlying lexical concept [1, 2, 3, and 4]. The lexical database is a hierarchy that can be searched upward or downward with equal speed. WordNet is a lexical inheritance system [2]. The success of WordNet is largely due to its accessibility, quality and potential in terms of NLP. WordNet was successfully applied in machine translation, information retrieval, document classification, image retrieval, and
conceptual identification [10]. Different WordNets can be aligned, resulting in the possibility of translation between different languages, as so called machine translation. Information retrieval can be achieved by improving the performance of query answering systems using lexical relations. Since WordNet links between semantically related words, semantic annotation and classification of texts and documents are possible [16]. The visual thesaurus is a dictionary and a thesaurus with an interesting interface. It is an excellent way of learning the English vocabulary and understanding how the English words link together. It has 145,000 words and 115,000 meanings and shows 16 kinds of semantic relationships. The user can as well hear the pronunciation of the word using a British or an American accent. Once the user enters a word, it is kept in the center, and all of the related words surround it. The user can click on a word to bring it to the center, roll over a word to learn more about it, and print the output chart [17]. Another interesting application is "READER". A person reading a text can click on a word, which is linked to a lexical database "WordNet", and reads its meaning in the given context [21]. An English dictionary, thesaurus and word finder program called WordWeb was developed based on the database from Princeton WordNet. It shows synonyms, antonyms, types and parts of a word. It has the advantage of integrating a dictionary and a thesaurus, unlike similar programs, where the dictionary and thesaurus are separate programs [5].

As shown from WordNet's applications, WordNet is inevitable for any language that aims to be part of today's ever-evolving NLP related applications. Surprisingly relatively few efforts have been made to develop an original Arabic WordNet. Filling this gap by developing an Arabic WordNet is a challenging and non trivial task. This project aims towards constructing a Nominal Arabic WordNet for Modern Standard Arabic, which will be the starting point of developing a complete Arabic WordNet. Our goal is to develop an Arabic WordNet freely distributed to the community. Our objectives are (1) Producing an Arabic WordNet, which contain nouns, verbs and adjectives, (2) Collect basic lexical data from available resources, and (3) Create a set of computer programs that would accept the user's queries and display output info to the user.

This paper presents Micro Arabic WordNet. This is a first step towards developing a complete Arabic WordNet. In this paper, we concentrate on using a subset of nouns for implementing this system. The other parts of speech e.g. verbs and adjectives, are considered as future work.

The next section gives a definition of Arabic language and its characteristics. Section 3 introduces some system requirements and specifications, general approaches for constructing WordNet, details of other WordNets, reasons of adopting the Arabic WordNet's (AWN) philosophy and challenges. Section 4 explains the different aspects of system design, where a system architecture, dataflow diagrams, entity-relation diagram, data structures and interface designs are sketched. Section 5 lists the sample data used in the database. It presents our test cases and our observations regarding those cases, and provides the performance tests. Finally, we include a cross-check validation of the requirements and discus the work still to be done on the system in the future.
2 Arabic WordNet

Semitic languages are a family of languages spoken by people from the Middle East and North and East Africa. It’s a subfamily of the Afro-Asiatic languages. Examples of Semitic languages are: Arabic, Amharic (spoken in North Central Ethiopia), Tigrinya (also in Ethiopia), and Hebrew. The name “Semitic” come from Shem son of Noah. Some characteristics of Semitic languages are:

1. Word order is Verb Subject Object (VSO),
2. Grammatical numbers are single, dual and plural, and
3. Words originate from a stem also called the “root”.

Unlike English, Arabic script is written from right to left. Diacritics are indicated above or below the letters in Arabic words. Arabic morphemes are often based on insertion between the consonants of a root form. Roots are verbs and have the form (f’a l). Arabic is cursive, written horizontally from right to left, with 28 consonants. Arabic is the only Semitic language having "broken plurals".

It is important to state that the most distinctive feature of this work is the insistence on maintaining language specific concepts and the intention of developing an Arabic WordNet which exhibits its richness rather than be driven by other incentives such as national security concerns, etc...

In the field of lexical semantics, terms such as 'word' which we would usually define as "the blocks from which sentences are made" [30], are defined differently. It is therefore necessary to define such terms in order to be able to comprehend the following concepts.

2.1 Word

A word is an association between a lexicalized concept and an utterance (or inscription) that plays a syntactic role [1]. For clarity, "word form" is used to refer to the physical utterance or inscription and "word meaning" to the lexicalized concept. Associations between word forms and word meanings are many:many. Some word forms could have several meanings (Polysemy), and some word meanings could be expressed by several word forms (Synonymy) [1].

![Synonymy and Polysemy](image_url)
2.2 Semantic Relations

Semantic relations are very important in lexical semantics. However, prior to the appearance of WordNet, they were implicit in conventional dictionaries [28]. Now they are explicit in WordNet, and play as the source of WordNet’s richness. Semantic relations associate between synsets & words. Before they are listed, an important concept must be put forward. It is the distinction between lexical semantic relations (table 1) and conceptual semantic relations (table 2). The former are between word forms such as, synonymy and antonymy whereas the latter are between synsets such as, hyponymy and meronymy [30].

Table 1. Lexical Semantic Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Relation in Arabic</th>
<th>Definition</th>
<th>Example</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymy</td>
<td>الترائف</td>
<td>Similarity of meaning; two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made.</td>
<td>Location and place</td>
<td>Lexical</td>
</tr>
<tr>
<td>Antonymy</td>
<td>التضاد</td>
<td>The antonym of a word x is sometimes not-x, but not always.</td>
<td>Rich and poor</td>
<td>Lexical</td>
</tr>
</tbody>
</table>

Table 2. Conceptual Semantic Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Relation in Arabic</th>
<th>Definition</th>
<th>Example</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyponymy/Hypernymy</td>
<td>انضواء/اختواء</td>
<td>IS A relation; a hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate and from any other hyponyms of that superordinate.</td>
<td>Maple and tree</td>
<td>Semantic</td>
</tr>
<tr>
<td>Meronymy/Holonymy</td>
<td>جزء/كل</td>
<td>HASA relation; part-whole relation</td>
<td>Finger and hand</td>
<td>Semantic</td>
</tr>
</tbody>
</table>

In [24] author proposed the idea of "Bootstrapping an Arabic WordNet Leveraging Parallel Corpora and an English WordNet". She studied the feasibility of meaning definition projection of English words onto their Arabic translation. She concluded that it is feasible to automatically bootstrap an Arabic WordNet taxonomy given less than perfect translations and alignments leveraging off existing English resources. The results were encouraging, as they are similar to those of researchers built EuroWordNet.
Supported by the United States Central Intelligence Agency, a group of researchers, some of who were involved in the construction of other WordNets such as, Princeton WordNet and Euro WordNet, decided to undertake the task of developing an Arabic WordNet [18]; for reasons such as, Arabic being a language spoken in more than 20 countries and the fact that it represents vital interest to US national security [19]. The project is still under construction, as it is due to finish in 2007 [19].

3 General approach for constructing WordNet

There are two main strategies for building WordNets: (1) Expand approach: translate English (or Princeton) WordNet synsets to another language and take over the structure. This is an easier and more efficient method. The outcome of this approach is of compatible structure with English WordNet. However, the vocabulary is biased by PWN, and (2) Merge approach: create an independent WordNet in another language and align it with English WordNet by generating the appropriate translation. This is more complex and requires a lot of work and effort. Language specific patterns can be maintained. But, it has different structure from WordNet [19].

Arabic is a totally different language from English, obviously the expand approach will not be appropriate. Moreover, it is undesirable for the Arabic WordNet to be biased by the English WordNet. Therefore, we are going to use the merge approach, since Arabic's specific patterns could be maintained. Arabic WordNet being developed in [19] is centered on enabling future machine translation between Arabic and other languages which justifies the use of tools such as the SUMO. Some aspects have been adopted in our project because SUMO for instance is a good software engineering practice (increasing the number of users). However, it is necessary to state that the most distinctive feature of our project is the insistence on maintaining language specific concepts and the intention of developing an Arabic WordNet which exhibits its richness rather than be driven by other incentives.

The user interface specification can be described by the following:

**Input:**
- Arabic
- Noun
- Singular

**Processing:**
- Search for the word in the Lexical Database (AWN). Display *synset* and *gloss*
- Display relations at users demand

**Output:**
- Synset of word in Arabic
- Gloss of synset
- Display Different relations from current synset i.e. synonyms, antonyms, hyponyms and hyponyms.
WordNet, since it is a lexical database, attempts to approximate the lexicon of a native speaker [30]. The mental lexicon, which is the knowledge that a native speaker has about a language, is highly dense in connectivity, i.e. there are many associations between words. Therefore, constant additions of relations are needed to improve the connectivity of a WordNet. This requires intensive research to discover relations which are not commonly used, since they are the ones which have a lower priority of inclusion into the database. Moreover, according to [30] "one of the central problems of lexical semantics is to make explicit the relations between lexicalized concepts". A lexicalized concept is a concept that can be expressed by a word [28].

One of the challenges in this project specific to Arabic is the fact that Arabic texts today tend to be written without diacritics, leaving the task of disambiguation to the reader's natural mental ability, which is a very complicated one when attempted to implement through a computer. For example, the form (كتاب) could be either intended to mean (كتاب) "kuttab" which is 'a group of writers' or (كتاب) "kitab" which is 'a book'. This example clearly demonstrates how missing diacritic marks compound lexical ambiguity.

Finally, it is a frequent criticism that much of the infrastructure for computational linguistics research or the development of practical applications are lacking. Low involvement of Arab linguists also compounds the challenge, driving some researchers to find alternative means which sometimes might degrade the desired quality of the outcome.

4 System Architecture

The system has two main components: (1) a User System Component, and (2) Lexicographer System Component

The latter is necessary since there is a lack in electronic Arabic resources and available lexicographers that are necessary for the population of the Arabic WordNet. The user system component retrieves information from Arabic_wordnet Database. It addresses the normal user's need, who is interested in finding out the different synsets and relations for a given word.
Fig. 3. Data Flow Diagram of displaying overview (level 2)

Fig. 4. Data Flow Diagram of displaying synset's relation (level 2)
The lexicographer system component stores information into \texttt{Arabic\_wordnet} Database. It handles the lexicographer (linguist) requirements, who basically insert new synsets and relations. Data flow starts when the lexicographer chooses an action. The lexicographer could either choose to add a synset or to edit an existing one. Dashed arrows are optional paths which the lexicographer could choose to follow.

\textbf{Fig. 5. Lexicographer System Component}
5 System Implementation and Validation

In this project we used MySQL as a database management system for many reasons. MySQL is a widely used software by many large companies for keeping thousands even millions of records. Since there are thousands of words in a language, the need for software to handle such a big number of records has emerged. MySQL is also web accessible. Querying MySQL is straightforward and easy. MySQL Query Browser is free software, which renders MySQL database with an interface similar to Microsoft Access DBMS. It also checks for users' query correctness.

The programming language used is Java Programming Language accessed by SunOne Java. Java is platform independent. It allows creating attractive graphical interfaces. It handles Unicode characters as well. As we are going to manipulate Arabic words, ASCII characters do not suffice our purpose. Unicode characters are the solution for writing and reading Arabic text. Being an object-oriented language; Java can provide a better structure and interface of the system. Java is also popular for its rich library which facilitates string manipulation.

5.1 Testing

Testing data was carefully chosen to cover all test cases. The test cases are shown in table 3.
Table 3. Test Cases Results

<table>
<thead>
<tr>
<th>Test case</th>
<th>Covered in System</th>
<th>Proved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input is correct</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Input is wrong</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Input is correct and relation exist</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Input is correct and relation does not exist</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 4. Test Cases Results

| Input                                      |             |

<table>
<thead>
<tr>
<th>Expected synsets output</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. مادة، جوهر، عين -- (أصول الشيء التي منها يتكوين)</td>
<td></td>
</tr>
<tr>
<td>2. عين، مقبلة، طرف، بصر -- (عضو الإجبار للإنسان)</td>
<td></td>
</tr>
<tr>
<td>3. عين، نفس، مصيبة -- (نفس الشيء)</td>
<td></td>
</tr>
<tr>
<td>4. عين، بئوض، بئع -- (يبدو الماء ينفع من الأرض ويجري)</td>
<td></td>
</tr>
<tr>
<td>5. عين، جاسوس -- (الشخص الذي يطلع على الأخبار السرية)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Real synsets output</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. مادة، جوهر، عين -- (أصول الشيء التي منها يتكوين)</td>
<td></td>
</tr>
<tr>
<td>2. عين، مقبلة، طرف، بصر -- (عضو الإجبار للإنسان)</td>
<td></td>
</tr>
<tr>
<td>3. عين، نفس، مصيبة -- (نفس الشيء)</td>
<td></td>
</tr>
<tr>
<td>4. عين، بئوض، بئع -- (يبدو الماء ينفع من الأرض ويجري)</td>
<td></td>
</tr>
<tr>
<td>5. عين، جاسوس -- (الشخص الذي يطلع على الأخبار السرية)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expected related synsets output</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. مادة، جوهر، عين -- (أصول الشيء التي منها يتكوين)</td>
<td></td>
</tr>
<tr>
<td>2. عين، مقبلة، طرف، بصر -- (عضو الإجبار للإنسان)</td>
<td></td>
</tr>
<tr>
<td>3. عين، نفس، مصيبة -- (نفس الشيء)</td>
<td></td>
</tr>
<tr>
<td>4. عين، بئوض، بئع -- (يبدو الماء ينفع من الأرض ويجري)</td>
<td></td>
</tr>
<tr>
<td>5. عين، جاسوس -- (الشخص الذي يطلع على الأخبار السرية)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Real related synsets output</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>عين، مقبلة، طرف، بصر -- (عضو الإجبار للإنسان وغيره)</td>
<td></td>
</tr>
<tr>
<td>جسم، جسد، بن -- (كل شخص يدرك من الإنسان والحيوان والنبات)</td>
<td></td>
</tr>
</tbody>
</table>
The functionality of the system is designed to handle all cases and all errors. Testing data was carefully chosen to cover all test cases. The test cases are shown in table 4. Apparently, the functionality of the system is designed to handle all cases and all errors.

5.2 Performance

In this section, we will discuss the approximation of data retrieval in our system. Another test was made to approximate processing times on Intel(R) Pentium(R) 4 CPU 3.20 GHz 3.19 GHz, 0.99 GB of RAM. The results using the timer in the source code are summarized in table 5.

<table>
<thead>
<tr>
<th>Number of test case</th>
<th>Retrieval time of synsets</th>
<th>Retrieval time of relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>15 milliseconds</td>
<td>78 milliseconds</td>
</tr>
<tr>
<td>Case 2</td>
<td>0 milliseconds</td>
<td>-</td>
</tr>
<tr>
<td>Case 3</td>
<td>16 milliseconds</td>
<td>31 milliseconds</td>
</tr>
<tr>
<td>Case 4</td>
<td>16 milliseconds</td>
<td>15 milliseconds</td>
</tr>
</tbody>
</table>

The results using the timer in MySQL Query Browser are given in table 6.

<table>
<thead>
<tr>
<th>Number of test case</th>
<th>Retrieval time of synsets</th>
<th>Retrieval time of relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>0.0043 seconds</td>
<td>0.0050 seconds</td>
</tr>
<tr>
<td>Case 2</td>
<td>0.0096 seconds</td>
<td>-</td>
</tr>
<tr>
<td>Case 3</td>
<td>0.0125 seconds</td>
<td>0.0454 seconds</td>
</tr>
<tr>
<td>Case 4</td>
<td>0.0125 seconds</td>
<td>0.0107 seconds</td>
</tr>
</tbody>
</table>
5.3 Future Work

Our current system tackles nominal singular input. In the future, we are planning to implement verbs and adjectives as well as plural input. There will be separate tables in the database for verbs and adjectives. Also, an algorithm will be designed to generate the singular form of a given plural word since only the singular forms will be stored in the database. Moreover, to make the system comprehensive it is planned to integrate a morphological analyzer component to the system to generate the sound form of the word if given an inflected form or a derived form.

Another issue that is planned to be tackled in the future is the problem of diacritics, a problem unique to Arabic. Lemmas with the same orthographical representation when stripped of diacritic marks will have to be disambiguated if the user attempts to search for one of them.

To enrich our database as much as possible, it is desired in the future to cover Classical Arabic in addition to the Modern Standard Arabic which is currently being covered. For additional functionality, and specific to Semitic languages, it would be convenient to have a search by roots or to display the words derived from the same root. Also, an important plan for our system is that we are currently upgrading our system to be a web application. To facilitate this step, we have taken all the necessary precautions. We have used open source tools like MySQL and Java, and hence developing the application as a Servlet or an Applet is not a big challenge. The operating system that will be used is Linux with Apache as a server. It is notable that the domains: arabicwordnet.com, arabicwordnet.org and arabicwordnet.net have been registered.

As well as the system being a textual-based system, we are looking forward to implement a graphical-based Arabic WordNet. The synsets and relations can be represented as hierarchies, trees or even radial diagram.

6 Conclusion

We have realized after doing research on the possibility of collecting validated data, that it is extremely difficult to populate the database because of the lack of machine readable dictionaries and available lexicographers. We decided therefore to include a lexicographer interface in addition to the original intended user interface. We have also built the system in a structure that enhances scalability. After upgrading the system to a web application (as it has been discussed in the previous chapter) we will aim to contact lexicographers in universities around the world to contribute in the construction of the Arabic WordNet. In the analysis phase we collected some user, lexicographer and system requirements and analyzed them. A general idea of the system architecture has been developed. The design phase included the system architecture, data flow diagrams, entity-relation diagram, data structures and interface design. The implementation phase discussed the different tools used in implementing the system. We define different functions in this phase as well as stating their pseudocode. The testing phase included database data as well as testing data, and a discussion. We also tested the performance of the system and stated the statistics. We
anticipated the realization of a Comprehensive Arabic WordNet once it is published on
the web (current project).

References

   Lexicography 3(4), 245–264 (1990)
   Lexicography 3(4), 265–277 (1990)
   3(4), 278–301 (1990)
   http://wordweb.info
   http://www.fathom.com/feature/1140/
    http://www.princeton.edu/pr/pwb/01/1203/1c.shtml
    http://www.illc.uva.nl/EuroWordNet/
    http://www.cs.cmu.edu/~alavie/Sem-MT-wshp/Wintner+Yona_presentation.pdf (2003,
    Sep.)
    Conference (2006)
20. Elkateb, S., Black, B.: Arabic, some relevant characteristics. Presentation.
    http://wordnet.princeton.edu/reader
24. Diab, M.T.: Feasibility of Bootstrapping an Arabic WordNet Leveraging Parallel Corpora
    and an English WordNet. In: Proceedings of the Arabic Language Technologies and
    Resources. NEMLAR, Cairo (2004)
    http://www.sil.org/linguistics/GlossaryOfLinguisticTerms
# Author List

<table>
<thead>
<tr>
<th>Author</th>
<th>Pages</th>
<th>Author</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agić, Ž.</td>
<td>349</td>
<td>Hao, Y.</td>
<td>453</td>
</tr>
<tr>
<td>Agirre, E.</td>
<td>474</td>
<td>Hatvani, Cs.</td>
<td>311</td>
</tr>
<tr>
<td>Alkhaliifa, M.</td>
<td>387</td>
<td>Hobbs, J.</td>
<td>111</td>
</tr>
<tr>
<td>Almási, A.</td>
<td>462</td>
<td>Hong, J.</td>
<td>506</td>
</tr>
<tr>
<td>Álvarez, J.</td>
<td>3</td>
<td>Horáčík, A.</td>
<td>194, 200</td>
</tr>
<tr>
<td>Angioni, M.</td>
<td>21</td>
<td>Hotani, C.</td>
<td>209</td>
</tr>
<tr>
<td>Atserias, J.</td>
<td>3</td>
<td>Hsiao, P.</td>
<td>220</td>
</tr>
<tr>
<td>Azarova, I. V.</td>
<td>35</td>
<td>Huang, C.</td>
<td>209, 220, 474, 506</td>
</tr>
<tr>
<td>Balkova, V.</td>
<td>44</td>
<td>Ion, R.</td>
<td>441</td>
</tr>
<tr>
<td>Barbu, E.</td>
<td>56</td>
<td>Isahara, H.</td>
<td>101, 419, 474</td>
</tr>
<tr>
<td>Bekavac, B.</td>
<td>349</td>
<td>Jaimai, P.</td>
<td>101</td>
</tr>
<tr>
<td>Bertran, M.</td>
<td>387</td>
<td>Kalhusk, N.</td>
<td>334</td>
</tr>
<tr>
<td>Bhattacharyya, P.</td>
<td>321, 360</td>
<td>Kalez, S.</td>
<td>321</td>
</tr>
<tr>
<td>Bijankhan, M.</td>
<td>297</td>
<td>Kanzaki, K.</td>
<td>474</td>
</tr>
<tr>
<td>Black, W.</td>
<td>387</td>
<td>Ke, X.</td>
<td>220</td>
</tr>
<tr>
<td>Bosch, S.</td>
<td>74, 269</td>
<td>Kerner, K.</td>
<td>229</td>
</tr>
<tr>
<td>Bozianu, L.</td>
<td>441</td>
<td>Kirk, J.</td>
<td>387</td>
</tr>
<tr>
<td>Broda, B.</td>
<td>162</td>
<td>Koëva, S.</td>
<td>239</td>
</tr>
<tr>
<td>Buitelaar, P.</td>
<td>375</td>
<td>Kopra, M.</td>
<td>321</td>
</tr>
<tr>
<td>Butnariu, C.</td>
<td>91</td>
<td>Krstev, C.</td>
<td>239</td>
</tr>
<tr>
<td>Calzolari, N.</td>
<td>474</td>
<td>Kunze, C.</td>
<td>281</td>
</tr>
<tr>
<td>Carrera, J.</td>
<td>3</td>
<td>Kuo, T.</td>
<td>209</td>
</tr>
<tr>
<td>Ceaușu, A.</td>
<td>441</td>
<td>Kutí, J.</td>
<td>254, 311</td>
</tr>
<tr>
<td>Charoenporn, T.</td>
<td>101, 419</td>
<td>le Roux, J.</td>
<td>269</td>
</tr>
<tr>
<td>Clark, P.</td>
<td>111</td>
<td>Lemnitze, L.</td>
<td>281</td>
</tr>
<tr>
<td>Climent, S.</td>
<td>3</td>
<td>Lüngen, H.</td>
<td>281</td>
</tr>
<tr>
<td>Cranier, I.</td>
<td>120, 178</td>
<td>Maks, I.</td>
<td>485</td>
</tr>
<tr>
<td>Csirik, J.</td>
<td>311</td>
<td>Mansoory, N.</td>
<td>297</td>
</tr>
<tr>
<td>Demontis, R.</td>
<td>21</td>
<td>Marchetti, A.</td>
<td>474</td>
</tr>
<tr>
<td>Deriu, M.</td>
<td>21</td>
<td>Marina, A. S.</td>
<td>35</td>
</tr>
<tr>
<td>Derwojedowa, M.</td>
<td>162</td>
<td>Martí, M. A.</td>
<td>387</td>
</tr>
<tr>
<td>Elkateb, S.</td>
<td>387</td>
<td>Mbame, N.</td>
<td>304</td>
</tr>
<tr>
<td>Farreres, J.</td>
<td>387</td>
<td>Melo, G.</td>
<td>147</td>
</tr>
<tr>
<td>Farwell, D.</td>
<td>387</td>
<td>Miháltz, M.</td>
<td>311</td>
</tr>
<tr>
<td>Fellbaum, C.</td>
<td>74, 111, 269, 387, 474</td>
<td>Mohanty, R. K.</td>
<td>321</td>
</tr>
<tr>
<td>Finthammer, M.</td>
<td>120, 178</td>
<td>Mokarat, C.</td>
<td>101</td>
</tr>
<tr>
<td>Fišer, D.</td>
<td>185</td>
<td>Monachini, M.</td>
<td>474</td>
</tr>
<tr>
<td>Gyarmati, Á.</td>
<td>254</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Moropa, K. 269 Ștefănescu, D. 441
Neri, F. 474 Storrer, A. 281
Nimb, S. 339 Su, I. 209, 220
Oliver, A. 3 Sukhonogov, A. 44
Orav, H. 334 Szarvas, Gy. 311
Pala, K. 74, 194 Szauter, D. 462
Pandey, P. 321 Szpakowicz, S. 162
Parm, S. 334 Tadić, M. 349
Pease, A. 387 Tesconi, M. 474
Pedersen, B. S. 339 Tuńś, D. 441
Piasecki, M. 162 Tuveri, F. 21
Poesio, M. 56 Vajda, P. 254
Prószéký, G. 311 VanGent, J. 474
Raffaelli, I. 349 Váradi, T. 311
Raffaelli, R. 474 Varasdi, K. 254
Ramanand, J. 360 Veale, T. 91, 453
Rambousek, A. 194, 200 Vider, K. 334
Reiter, N. 375 Vincze, V. 462
Rigau, G. 3, 474 Vitas, D. 239
Riza, H. 101 Vliet, H. 485
Robkop, K. 419 Vossen, P. 200, 387, 474, 485
Rodríguez, H. 387 Weikum, G. 147
Rouhizadeh, M. 406, 520 Xu, M. 506
Segers, R. 485 Yablonsky, S. 44
Shamsfard, M. 406, 413, 520 Yarmohammadi, M. A. 406, 520
Sharma, A. 321 Zawislawska, M. 162
Sinopalnikova, A. A. 35 Zidoum, H. 531
Sornlertlamvanich, V. 101, 419 Zutphen, H. 485
Spohr, D. 428