Why are wordnets important?

Veronika Vincze[1], György Szarvas[2], János Csirik[2]

[1] University of Szeged, Department of Informatics
Árpád tér 2. Szeged, H-6720
HUNGARY
vinczev@inf.u-szeged.hu

[2] Research Group on Artificial Intelligence of the
Hungarian Academy of Sciences and University of Szeged
Aradi vértanúk tere 1. Szeged, H-6720
HUNGARY
{szarvas, csirik}@inf.u-szeged.hu

Abstract: - Wordnets are lexical databases in which words are organized into clusters based on their meanings, and they are linked to each other through different semantic and lexical relations. The first wordnet called the Princeton WordNet was created for English, which were followed by various wordnets created within the framework of the EuroWordNet and BalkaNet projects, among others. Here we focus on the development of wordnets in general and of the Hungarian WordNet (HuWN). The process of constructing HuWn is illustrated by examples, some language-specific and language-independent problems encountered during the construction process are discussed, and then basic statistical data on HuWN are presented as well. Finally, two subontologies of HuWN, namely, the financial domain ontology and the legal domain ontology are also presented, and possible applications of WordNets are outlined.

Key-Words: - **natural language processing, wordnet, word-senses, by content analysis and clustering, interlingual processing, semantic search**

1 Introduction

Wordnets are lexical databases in which words are organized into clusters based on their meanings, and they are linked to each other through different semantic and lexical relations, yielding a conceptual hierarchy (i.e. lexical ontology) of words. Originally, they were designed to represent how linguistic knowledge is organized within the human mind [1]. The first wordnet called the Princeton WordNet was created for English [1], which was followed by numerous wordnets all around the world. Wordnets for European languages have been developed mostly within the framework of the EuroWordNet and BalkaNet projects [2, 3], among others..

Wordnets can differ in size, but they – especially the Princeton WordNet – are usually considered to be the largest database containing linguistic information for the given language. Thus, they can be used in various applications within the field of computational linguistics: word sense disambiguation, machine-assisted translation, document clustering, and so on.

This paper is structured as follows. First, we focus on the inner structure of wordnets, that is, basic relations that constitute the hierarchy are illustrated by examples. Then examples of wordnets (being) created all over the world are presented, with special attention to the development of the Hungarian WordNet (HuWN). Some language-specific and language-independent problems encountered during the construction are also discussed. HuWn also contains two subontologies, namely the financial domain ontology and the legal domain ontology, which are also presented together with some basic statistical data of HuWN. Finally, possible applications of WordNets in computational linguistics are discussed.

2 Representing conceptual hierarchies by wordnets

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 wordnets is that lexical information is organized in terms of meaning; that is, a synset (the basic unit of wordnets) contains words of the same part-of-speech which have approximately the same meaning. Thus, it is synonymy that functions as the essential principle in the construction of wordnets [1]. An example of a synset is the following:

{bicycle:1, bike:2, wheel:6, cycle:6}

Literals forming one synset are numbered as a word can have several meanings and it is important to represent that a word is synonymous with other words in one given sense. Thus, *cycle* occurs in five other synsets, including:

{cycle:1, rhythm:3, round:2}
{Hertz:1, Hz:1, cycle per second:1, cycles/second:1, cps:1, cycle:4}
{cycle:5, oscillation:3}

Synsets are connected to each other by means of semantic and lexical relations, yielding a hierarchical network of concepts. *Semantic relations* hold between concepts. In other words, not the forms but their meanings are related. Such relations include hyponymy and meronymy. On the other hand, *lexical relations* connect different word forms. For instance, synonymy, antonymy and different morphological relations belong to this group [1]. Next, we will focus on the basic relations of wordnets – we provide definitions and illustrate them using nominal synset examples.

Hypernymy has a crucial role in forming the conceptual hierarchy in wordnets. A concept is a hypernym of another concept if it is a more generic term and the latter can be seen as an instance of the former (i.e. the IS-A relation holds between them) [1]. For example:

{substance:1, matter:1} is *hypernym* of {fluid:2}, which is *hypernym* of {gas:2}

{furniture:1, piece of furniture:1, article of furniture:1} is *hypernym* of {wardrobe:1, closet:3, press:6}

Based on this relation, synsets can be organized into a conceptual hierarchy represented by a tree. Hypernymy is a transitive relation; that is, a synset usually has one direct hypernym, and it may have several hypernyms on different levels of the hierarchy. For instance, the direct hypernym of {bicycle:1, bike:2, wheel:6, cycle:6} is {wheeled vehicle:1}, but its indirect hypernyms include {container:1}, {artifact:1, artefact:1} and {entity:1}. On the other hand, {bicycle:1, bike:2, wheel:6, cycle:6} is a hypernym of {mountain bike:1, all-terrain bike:1, off-roader:1} and {bicycle-built-for-two:1, tandem bicycle:1, tandem:1}, among others. This is illustrated in the following figure:

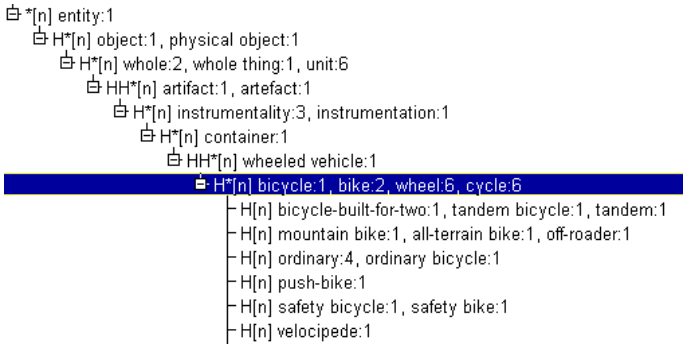

Fig. 1. Hypernyms and hyponyms of {bicycle:1, bike:2, wheel:6, cycle:6}

Holonymy and meronymy encode part-whole relations in wordnets. A concept is a meronym of another one if the former is a part of the latter (i.e. the HAS-A relation holds between them) [1]. In the Princeton WordNet, holonymy is encoded by three different relations [4], and in EuroWordNet there are two other relations besides these [2]. First, *holo_part* tells us that a thing is a component part of another thing:

{bicycle:1, bike:2, wheel:6, cycle:6}is *holo_part* of {pedal:2, treadle:1, foot pedal:1, foot lever:1}

Second, *holo_member* tells us that a thing or person is a member of a group:

{fleet:3} is *holo_member* of {ship:1}

Third, *holo_portion* refers to the *stuff* that a thing is made from [4], but this relation links a whole and a portion of the whole in EuroWordNet [2]:

{joint:6, marijuana cigarette:1, reefer:1, stick:5, spliff:1} is *holo_portion* of {cannabis:2, marijuana:2, marihuana:2, ganja:2}
{bread:1} is *holo_portion* of {piece:8, slice:2}(EuroWN)

Fourth, *holo_madeof* encodes the *stuff* a thing is made from in EuroWordNet:

{paper:1} *has_holo_madeof* {book:2, volume:3}

Fifth, *holo_location* denotes a thing that is located within another place:

{oasis:1} *has_holo_location* {desert:1}

Holonymy and meronymy also allow us to visualize the relations between synsets as a tree structure. Here

Figure 2 shows the parts of a bicycle (and the parts of a bicycle wheel):

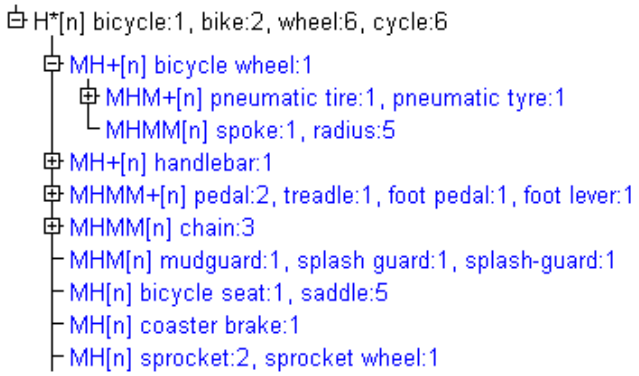

Fig. 2. Meronyms of {bicycle:1, bike:2, wheel:6, cycle:6}

Since a thing can function as a part of more than one thing – e.g. many vehicles have wheels –, it can have more than one holonym. This means that in a holonymic hierarchy, a leaf could belong to more than one tree. However, in this case it is more advisable to represent the hierarchy in a meronymic tree, where the top node is the part and the leaves of the tree are the entities that have the top node as a part of them. The following figure represents those entities that contain a handle as a part:

Fig. 3. Holonyms of {handle:1, grip:2, handgrip:1, hold:8}

3 Wordnets in the world

The first wordnet was created for the English language at Princeton University, so it is called the Princeton WordNet. It has been developed continuously since the 1990s, and it is now the largest lexical database that is available for the English language, and it can be readily adapted to various computational applications. As of 2006, Princeton WordNet contains about 150,000 words in approximately 115,000 synsets [5].

The EuroWordNet database has similar structure to the Princeton WordNet, but there are some noticeable differences between the basic principles that were applied during its construction [2]. First of all, it is a multilingual project; that is, synsets for Dutch, Italian, Spanish, German, French, Czech and Estonian are included in the database, which are linked to each other by means of an interlingual index. Second, there are some relations between synsets that were either not included in Princeton WordNet or they were interpreted differently from it (see the example of *holo_portion* above). WordNets for different languages differ in size, but there is a shared top ontology of 63 semantic distinctions and a shared set of 1024 concepts available for all languages [6].

The BalkaNet project sought to extend EuroWordNet with lexical databases created for languages of the Balkan Peninsula, namely Bulgarian, Greek, Turkish, Serbian and Romanian [3, 7]. The Base Concepts of EuroWordNet were expanded to 8516 concepts, which are present in each wordnet (BalkaNet Concept Set). Another innovation of the project is that PWN 2.0 was used as the base ontology, on the basis of which other wordnets were developed (instead of PWN 1.5). For this project, a freely available editor called VisDic was developed, and databases for each language were stored in XML format [8].

Since then, other wordnets have been created and developed for several languages. The languages covered include Arabic, Croatian, Chinese, Danish, Slovene, Polish, Russian, Persian and those of Africa and India [9].

4 The Hungarian WordNet project

The Hungarian WordNet (HuWN) was developed by the Research Institute for Linguistics of the Hungarian Academy of Sciences, the Department of Informatics of the University of Szeged, and MorphoLogic Ltd. in a 3-year project [10, 11]. As a result, HuWN now consists of over 40.000 synsets, out of which 2.000 synsets form

part of a subontology in the business domain. The number of synsets belonging to different parts-of-speech is shown here:

Part-of-speech	Number of synsets
Noun	33 778
Verb	3 310
Adjective	4 083
Adverb	1 038
Total	42 209

Table 1. The distribution of parts-of-speech in HuWN

Here the Princeton WordNet 2.0 served as a basis for the construction of HuWN; that is, synsets belonging to the BalkaNet Concept Set were selected from PWN 2.0 and then translated into Hungarian. These were then edited, corrected and extended with other synonyms using the VisDic editor. The set of concepts to be included in HuWN were expanded concentrically later on. That is, descendants of the existing synsets were treated as synset candidates. The final decision on their status (whether they should be included or not) was influenced by several factors such as the frequency of the concept or its presence in other WordNets [11].

The following relation types were borrowed from the Princeton WordNet: hypo- and 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). In addition, some new relation types were introduced, partly because of language specific phenomena and partly for other, language-independent reasons [11, 12].

As for the first type, new verbal relations should be mentioned. Since in Hungarian, it is verbs as lexical units that bear aspectual information (as opposed to English for example, where aspect is mostly related to tense and grammatical structure), it is neccessary to represent this information in the verbal network as well. For this reason, an abstract node *nucleus* is introduced for each event, which functions as an idealized eventuality consisting of three parts: preparatory phase, culmination (telos) and consequent state [12]. Subevents of the idealized eventuality are linked to the nucleus through the new relations *is_preparatory_phase_of*, *is_telos_of* and *is_consequent_state_of* as it is shown in the following example:

{szárad} 'is drying' *is_preparatory_phase_of* NUCLEUS MEGSZÁRAD 'get dry'
{megszárad} 'get dry' *is_telos_of* NUCLEUS MEGSZÁRAD
{száraz} 'dry' (adjective) *is_consequent_state_of* NUCLEUS MEGSZÁRAD

In addition to these language-specific relations, some language independent relations were introduced in HuWN as well [12]. As for adjectives, they were generally represented in a bipolar cluster structure, but there were certain groups of adjectives that did not fit into this pattern. For instance, *pozitív* 'positive', *negative* 'negative' and *semleges* 'neutral' seem to be focal points in the same domain though it is only *pozitív* and *negatív* that are real antonyms. What is more, s*emleges* expresses a value situated right the middle of the scale determined by *pozitív* and *negative*. Thus, the new relation that connects *semleges* and *pozitív* on the one hand and *semleges* and *negatív* on the other hand is called *scalar middle* [12]:

{semleges} is the *scalar middle* of {pozitív}
{semleges} is the *scalar middle* of {negatív}
{pozitív} is the *near_antonym* of {negatív}
{negatív} is the *near_antonym* of {pozitív}

The third type of new relations introduced in HuWn expresses a certain connection between nouns and adjectives. For some adjectives it is the case that they can only modify nouns of a certain type; that is, those belonging to a certain semantic class [12]. An example is *egynyári* 'annual', *kétnyári* 'biennial' and *évelő* 'perennial', which may only refer to plants. To capture this information, the new relation *partitions* was introduced, which connects the above-mentioned adjectives with the synset {növény} 'plant':

{egynyári} *partitions* {növény}
{kétnyári} *partitions* {növény}
{évelő} *partitions* {növény}

Overall, then, the creation of the Hungarian WordNet not only means another wordnet for the community to study and apply, but it also enriches the theoretical and linguistic background behind wordnets because it introduces some new relations that could be usefully applied to wordnets of other languages of the world.

5 Extending the Hungarian WordNet with concepts taken from economics and law

Besides the construction of general purpose language ontologies, developing domain ontologies for specific terminologies is essential since the vocabularies of general language ontologies are rarely capable of covering the specific language terminology of a special scientific or technical domain. For this reason, two

subontologies of the Hungarian WordNet were created, namely, an economic and a legal one.

5.1 Economic subontology

Nowadays, one of the most dynamically developing areas is the domain of finance and business, which makes heavy demands on applications in language technology. The importance of communication between business partners with different native languages can hardly be overestimated since Hungary became a member state of the European Union. The sudden increase in the quantity of business news requires the constant development of information extraction tools designed for this domain. Domain ontologies specifically tailored to the special terminology of a domain can serve as a basis for information extraction systems.

To construct a business domain ontology, first of all the typical terms used in business communication must be identified. When collecting these terms, our group made use of two different strategies [11].

First, our linguists read business and financial news on the one hand and websites on political and economic issues on the other. They scanned these texts for business term candidates, which were collected into lists based on their part-of-speech. Elements of the lists were transformed into synset candidates automatically, and the linguists in our group then decided whether or not to include them in the domain ontology. If the synset was already present in the general ontology, it was obviously disregarded; that is, it was not duplicated. If the synset candidate was to be included in the economic subontology, it was linked to its English equivalent in PWN 2.0 (if any), and it was inserted into the already existing hierarchy.

Second, our group selected 32 concepts belonging to the domains of economy, enterprise and commerce from PWN 2.0, which appeared to be useful for the construction of the domain ontology. This strategy sought to provide more complex encyclopedic knowledge in this field. These concepts and their hyponyms (that is, their subtrees) were then automatically translated into Hungarian, transformed into synsets and then checked manually by our linguists.

The financial domain ontology of the Hungarian WordNet contains about 2800 synsets.

5.2 Legal subontology

Our research group – within the framework of an earlier project – implemented the business subontology of the general, Hungarian wordnet. Presently, making use of the lessons learned from the above task, the group is working to create a subontology of a future Hungarian legal wordnet. This work embraces the organization of concepts related to financially liable offences in a hierarchy and the creation of a terminological ontology and dictionary.

This initiative plays a significant role in the integration of the Hungarian legal system into the international legal system, or more precisely into that of the European Union since it lays the foundations for establishing a legal database that provides the lexical background for the approximation and harmonization of laws with the EU.

Experts in law, informatics and linguistics are participating in this work in the following order: first source law texts are procured, collected and organized, then a frequency list of concept candidates is generated from which concepts pertaining to the target ontology; that is, financially liable offences are selected. Afterwards, concepts are defined and the ancillary information necessary for law interpretation is included. Then concepts are organized in a hierarchy that accurately reflects legal as well as lexical (semantic) relations, which is followed by the phase when the data is transformed to XML for viewing and editing lexical databases. The process is then concluded with the control phase when the structure and content of the network are finalized.

As regards its relations and structure, this network of concepts applies those of a general purpose ontology, but due to the specific features of legal terminology, it does not always rely just on linguistic considerations, e.g. when definitions are to be formed (in the case of a general purpose ontology, definitions contain a hypernym of the concept to be defined, but a legal ontology does not always follow this rule; it often makes use of lists (of hyponyms or meronyms) in defining a concept), or when relevant legal content (necessary information for the interpretation of law, e.g. dates, quantities and paragraphs) needs to be explicitly stated.

6 Some possible applications of wordnets

From a computational linguistics viewpoint, wordnets are well-structured databases in which thousands of words and senses are organized into a semantic network. Since their inner structure is much more complex than that of ordinary dictionaries or thesauri, their possible applications extend to various fields in the computational linguistics domain. In this section we will mention several potential applications of wordnets.

6.1 Word-sense disambiguation

Word-sense disambiguation (WSD) attempts to resolve ambiguities (homonyny, polysemy) in texts. It is an essential intermediate task for many applications in natural language processing (e.g. human-machine

interaction, text comprehension, machine translation and information retrieval and extraction).

To perform a WSD task successfully, words to be disambiguated should be selected and the possible senses of those words should be given and accurately defined. Since wordnets generally contain various (if not all) senses of a word, they can be readily used as a source for sense definitions. For instance, *cycle* has six different senses, hance there are six different sense definitions in the Princeton WordNet:

> {**cycle:1**, rhythm:3, round:2}: an interval during which a recurring sequence of events occurs
> {**cycle:2**}: a series of poems or songs on the same theme
> {**cycle:3**}: a periodically repeated sequence of events
> {Hertz:1, Hz:1, cycle per second:1, cycles/second:1, cps:1, **cycle:4**}: the unit of frequency; one Hertz has a periodic interval of one second
> {**cycle:5**, oscillation:3}: a single complete execution of a periodically repeated phenomenon
> {bicycle:1, bike:2, wheel:6, **cycle:6**}: a wheeled vehicle that has two wheels and is moved by foot pedals

Since these definitions are already present in wordnets, the time-consuming task of defining possible senses can be reduced to the simple selection of them from wordnets when preparing for a WSD task.

The inner structure of wordnets can also make the disambiguation process much easier. In wordnets, concepts are organized into synsets; that is, words and their synonyms are grouped together. If a synonym occurs in the context of the word to be disambiguated, it functions as an indicator for the sense whose synset contains the synonym and the word in question as well. The following example nicely illustrates this:

> Welcome in the official Mountain **Cycle** website. Find here every news, **bike** range, race team, warranty and lots of others things. Let's go!

The sentence contains the word *cycle*, whose precise sense is to be determined. In the following sentence the word *bike* can be found, which suggests that the sense definition designed for the synset {bicycle:1, bike:2, wheel:6, cycle:6} is to be selected here since it is this synset that covers both words.

Within the framework of the Hungarian WordNet project, the first Hungarian WSD corpus was built [13]. Sense distinction was made on the basis of HuWN synsets. On the other hand, whenever a necessary sense was missing from HuWN, the database was extended with that sense. For results, statistics and further details on the corpus, see [13].

6.2 Subject encoding and document clustering

Subject encoding and text document clustering can definitely improve the browsability of large text collections. Automated topic encoding and document clustering can benefit from wordnets by exploiting several lexical and semantic relations (like synonymy, hypernymy) that can reveal non-trivial similarities. This can result in significant improvement of the quality of clusters [14].

6.3 Machine Assisted Translation and Machine Translation

Wordnet projects seek to ensure interoperability between wordnets of different languages by using an International Language Index (ILI), which allows synsets belonging to the same concept to be readily accessible in every database. Thus multilingual wordnets can be used as large dictionaries where a set of words in one language corresponds to another set of words in the other language. This feature can be exploited in several machine-assisted translation applications.

With the help of wordnets, intelligent dictionaries can be developed which facilitate the task of translating documents. When in doubt, the human translator can look up the appropriate sense of the word to be translated in the wordnet database of the source language, and its equivalent synset in the target language is immediately provided by the application. The only thing the translator should do is to choose from the synonyms covered by the synset in the target language. In this way, machine-aided human translation could be made faster, easier and more cost-effective.

Statistical Machine Translation systems can also benefit from multilingual wordnets, as the interlinked lexical databases provide a rich and useful source of translation candidates (synsets linked through the interlingual index) and substitutes (using the semantic relations encoded in wordnets, such as hypernymy). For instance, the multilingual wordnet developed for Indian languages offers a background for English-to-Indian language and Indian-language to Indian-language machine translation systems [15].

6.4 Multilingual document retrieval and browsing

One of the main goals of the BalkaNet project was to demonstrate the usability of wordnets for multilingual information retrieval. For this purpose the participating

institutes prepared 100-100 concepts for two domains that were designed and implemented in each BalkaNet language. The interlinked synsets then provided a straightforward basis for indexing and retrieving multilingual document collections using queries in any one of the languages involved.

7 Summary

In this paper, several possible applications of wordnets in the field of human language technology were presented along with a detailed description of wordnets (the history of their creation and the basic principles behind their structure). As an illustrative example, the construction of Hungarian WordNet and its two subontologies were outlined. Then four practical applications were chosen to demonstrate the utility and applicability of wordnets in the area of computational linguistics.

8 Acknowledgements

The research presented in this paper was supported by the TUDORKA7 project of the Jedlik Ányos 2007 Programme of the National Office for Research and Technology (NKTH, http://www.nkth.gov.hu/) of the Hungarian government.

References:

[1] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., Introduction to WordNet: an On-line Lexical Database, *International Journal of Lexicography*, Vol. 3, No. 4, 1990, pp. 235–244.

[2] Alonge, A., Bloksma, L., Calzolari, N., Castellon, I., Marti, T., Peters, W., Vossen P., The Linguistic Design of the EuroWordNet Database, *Computers and the Humanities. Special Issue on EuroWordNet*, Vol. 32, No. 2–3, 1998, pp. 91–115.

[3] Tufiş, D. (ed.), *Romanian Journal of Information Science and Technology. Special Issue on BalkaNet*, Vol. 7, No. 1–2, 2004.

[4] Miller, G. A., Nouns in WordNet: A Lexical Inheritance System, *International Journal of Lexicography*, Vol. 3, No. 4., 1990, pp. 245–264.

[5] http://wordnet.princeton.edu/

[6] http://www.illc.uva.nl/EuroWordNet/

[7] Tufiş, D., Cristea, D., Stamou, S., BalkaNet: Aims, Methods, Results and Perspectives. A General Overview, *Romanian Journal of Information Science and Technology. Special Issue on BalkaNet*, Vol. 7, No. 1–2, 2004, pp. 9–43.

[8] Horák, A., Smrž, P., New Features of Wordnet Editor VisDic, *Romanian Journal of Information Science and Technology. Special Issue on BalkaNet*, Vol. 7, No. 1–2, 2004, pp. 201–213.

[9] Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.) *Proceedings of the Fourth Global WordNet Conference. GWC 2008*, University of Szeged, Department of Informatics, 2008

[10] Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M., Szarvas, Gy., Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction, Project report, *Proceedings of the Third International WordNet Conference (GWC2006)*, January 22–26, South Jeju Island, Korea, 2006, pp. 291–292.

[11] Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T., Methods and Results of the Hungarian WordNet Project, in Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.) *Proceedings of the Fourth Global WordNet Conference. GWC 2008*, University of Szeged, Department of Informatics, 2008, pp. 311–320.

[12] Kuti, J., Varasdi, K., Gyarmati, Á., Vajda, P., Language Independent and Language Dependent Innovations in the Hungarian WordNet, in Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.) *Proceedings of the Fourth Global WordNet Conference. GWC 2008*, University of Szeged, Department of Informatics, 2008, pp. 254–268.

[13] Vincze, V., Szarvas, Gy., Almási, A., Szauter, D., Ormándi, R., Farkas, R., Hatvani, Cs., Csirik, J., Hungarian Word-sense Disambiguated Corpus, in *Proceedings of 6th International Conference on Language Resources and Evaluation,* Marrakech, Morocco, 2008.

[14] Hotho, A., Staab, S., Stumme, G., WordNet improves text document clustering, in *Proceedings of the SIGIR 2003 Semantic Web Workshop*, Toronto, Canada, 2003.

[15] Mohanty, R.K., Bhattacharyya, P., Kalele, S., Pandey, P., Sharma, A., Kopra, M., Synset Based Multilingual Dictionary: Insights, Applications and Challenges, in Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.) *Proceedings of the Fourth Global WordNet Conference. GWC 2008*, University of Szeged, Department of Informatics, 2008, pp. 312–333.