

Semantic Processing of a Hungarian Ethnographic Corpus

Miklós Szóts
Applied Logic Laboratory
Hankóczy u. 7.
Budapest, Hungary
szots@all.hu

Sándor Darányi
University of Borås
Swedish School of Library and
Information Science
Borås, Sweden
sador.daranyi@hb.se

Zoltán Alexin
University of Szeged
Department of Software
Engineering
Árpád tér 2., Szeged, Hungary
alexin@inf.u-szeged.hu

Veronika Vincze
University of Szeged
Institute of Informatics
Árpád tér 2., Szeged, Hungary
vinczev@inf.u-szeged.hu

Attila Almási
University of Szeged
Institute of Informatics
Árpád tér 2., Szeged, Hungary
vizipal@gmail.com

ABSTRACT

In this poster, a Hungarian ethnographic database containing linguistic annotation is presented. The corpus contains texts from three domains, namely, folk beliefs, *táltos* texts and tales. All the possible morphosyntactic analyses assigned to each word and the appropriate one selected from them (based on contextual information) are also marked. Syntactic (dependency) annotation is added semi-automatically to the corpus texts at a second phase of the processing. With the help of these enriched linguistic attributes, the texts can be semantically analyzed and clustered. The research and development team is working on a semantic search tool enabling to browse the texts on the basis of their semantic meaning. The proposed technology may result in a new approach to the ethnographic research and may open a new type of access to the databases.

Categories and Subject Descriptors

I.2.7 [Natural Language Processing]: Language Parsing and Understanding

1. INTRODUCTION

The Applied Logic Laboratory and the University of Szeged (Institute of Informatics and Department of Library and Information Science) are developing a technology to perform meaning-based search for natural language texts. Researchers wish to go beyond the world of simple, *type-in-your-keywords* search engines; and to develop a technology and an integrated search engine which performs genuine content-oriented search in natural language documents (textual data), by adapting and combining existing statistical and symbolic techniques in a novel way in order to exploit the user's semantic competence to a considerably greater

extent than traditional search engines do. The kernel of the technology developed during the project is language-independent, while the prototype is going to be developed for patent specifications in English and Hungarian and Hungarian ethnographic texts.

Parallel to the above, a linguistic annotation method and software components are developed. By applying these computational methods and with considerable human efforts a manually checked and corrected ethnographic database is produced. Its importance is twofold: on the one hand, textual databases from the folklore domain have not been (or have hardly been) available in a digitalized version (see e.g. [5] on the difficulties of creating folklore databases). On the other hand, Hungarian folklore texts – to our best knowledge – have not been analyzed with computational linguistic tools.

The processing of the folklore texts follows the traditions of the Szeged Treebank [1]. The format of the ethnographic corpus is TEI XML¹. This database could serve as the input of further semantic processing, like clustering or thematic classification of texts.

In the following, the ethnographic corpus will be presented, some statistical data on the corpus will be shown, the process of the linguistic annotation will be discussed, and the importance of semantic search in ethnographic texts will be emphasized.

2. TOPICS WITHIN THE CORPUS

The corpus contains texts from three domains: folk beliefs (2704 short texts), *táltos* texts (432 texts) and folk tales (1185 texts). Beliefs and *táltos* texts were collected at the beginning of the 20th century and data came from all over the areas of historic Hungary. The original manuscripts can be found in the Museum of Ethnography and they have been published in a volume as well [9]. Tales were collected from the Hungarian Electronic Library (<http://www.mek.hu>) and from various Internet sources, like <http://www.nepmese.hu>.

¹See <http://www.tei-c.org>.

2.1 Folk beliefs

The topics of folk beliefs involve almost every aspect of ordinary life: main stages of human life (birth, christening, getting a husband, illness, death, otherworld), weather, special days, animals. The short – typically one-sentence-long – beliefs are accompanied with some explanation or short stories. In the collection, simple descriptions and incantations in the form of poems can also be found. Certain beliefs occur in different versions. Some folk traditions are also preserved in the beliefs.

Ha a menyasszony cipőjét ellopják a lakodalom éjjelén s lekaparva a talpáról a földet felteszik a füstre – ez a házas társak nyugodt életét megrontja.

'If the bride's shoes get stolen at the night of the wedding, and the soil scratched from their sole is smoked – this will ruin the calm life of the spouses.'

2.2 Táltos texts

The táltos texts tell us about wizards, medicine men and women, táltoses (a mythical figure similar to a shaman, whose task is to heal the body and soul of his/her people) and their abilities and earmarks. The collection comprise short stories in several dialects of Hungarian.

2.3 Tales

The tale collection includes masterpieces of Elek Benedek, who was a famous tale writer. The collection also contains various translated tales from different countries as well as Hungarian folk tales. As for their topic, tales depict stories about kings, queens, princesses, soldiers, rich and poor men, dragons, giants and dwarfs, fairies, ghosts, heaven and hell and many more. Table 1 presents a statistics on the corpus.

Table 1: Statistical data on the corpus

Text type	Number of texts	Number of words
Folk beliefs	2074	65807
Táltos texts	432	44021
Folk tales	1185	1311952
Total	3691	1421780

3. LINGUISTIC ANNOTATION

After the digitalization the texts were segmented into words. Then each word was computationally assigned one or more morphosyntactic codes. The obtained vocabulary was later manually checked and corrected by linguists. Morphosyntactic analysis was based on the categories used in The Concise Dictionary of the Hungarian Language [8]. There were some special words or wordforms that seemed to be problematic from the viewpoint of morphological analysis:

- In the case of regional words, the sense and their parts-of-speech had to be determined. (E.g.: *goroboncás* 'wandering magician', *slájer* 'veil of the bride'.)
- Words and wordforms with substandard orthography: they were paired with their standard forms and the

MSD code of those could usually be copied from the Szeged Treebank. If not, they were provided by our linguists. (E.g.: *ígízis* – standard: *igézés* 'spelling', *abbú* – standard: *abból* 'from that'.)

- Sometimes the word with substandard orthography coincided with another existing (standard) word. They needed special attention during the disambiguation for they already had an MSD code (according to the standard orthography, however, in these texts it was typically the substandard form that occurred. Thus, the standard version of those had to be given together with their MSD code(s). (E.g.: *mellül* – standard 'breast-ESSIVE' – substandard: *mellől* 'from it', *aggyá* – standard: 'brain-TRANSLATIVE' – substandard: *adjál* 'give-IMP'.)

```
<s>Ahun gyűnnek!  
<choice>  
<orig><w>Ahun</w></orig>  
<reg><w>Ahol  
<ana>  
<msd><lemma>ahol</lemma>  
<mecat>[Pd]</mecat></msd>  
</ana>  
<anav>  
<msd><lemma>ahol</lemma>  
<mecat>[Pr]</mecat></msd>  
</anav>  
<anav>  
<msd><lemma>ahol</lemma>  
<mecat>[Pd]</mecat></msd>  
</anav>  
</w></reg>  
</choice>  
<choice>  
<orig><w>gyűnnek</w></orig>  
<reg><w>jönnek  
<ana>  
<msd><lemma>jön</lemma>  
<mecat>[Vmip3p-n]</mecat></msd>  
</ana>  
<anav>  
<msd><lemma>jön</lemma>  
<mecat>[Vmip3p-n]</mecat></msd>  
</anav>  
</w></reg>  
</choice>  
<c>!</c>  
</s>
```

Figure 1: Sample XML containing words with substandard orthography

The corpus currently contains morphosyntactic annotation based on the MSD coding system [2]. All the possible morphosyntactic analyses are assigned to each word, and the appropriate one will be selected from them based on contextual information. In Figure 1 a sample XML fragment is shown. The fragment presents how the ethnographic corpus handles substandard morphology taking into account the newest TEI encoding guideline (TEI P5).

Applying `<choice>` `</choice>` tags we can retain the original orthography of words and add the regular typing as well. The regular typing is marked up by the `<reg>` `</reg>` tags. A morphosyntactic analyses are provided between `<anav>` `</anav>` tags. The disambiguated part-of-speech is provided between `<ana>` `</ana>` tags.

Syntactic annotation is planned to be carried out by the Malt parser [3]. J. Nivre developed a statistical method for learning dependency parsing. His software needs a learning database in a given format and from this training set it can build a probabilistic model for parsing. An other program can execute the learned model on unknown texts. The conversion of the Szeged Treebank to dependency structure is approaching to the end, which will be used as a training database for the parser. In the fall of 2010 a trained dependency parser for Hungarian will be ready for application.

4. SEMANTIC SEARCH

Generally we use only surface level lexical semantics; however, in certain specific questions deep semantics can be used too – e.g. in the domain of technical texts the quantities should be represented precisely. Note, that the use of deep semantics is generally domain dependent.

The difference between lexical and deep semantics is illustrated by the following example. Let us consider the sentences *The prince got a ring from the magician.* and *The magician gave a ring to the prince.* Clearly, these describe the same situation. In our system both will be represented by a structure like the following one:

```
event:      giving1
initiator:  magician1
theme:      ring1
recipient:  prince1
```

Here we describe only the connections of the event denoted by *give* (or *get*) with its dependents. In a deep semantic approach we would need to describe the meaning of these words too – e.g. to give the necessary condition and the result of the event.

Verbs *give* and *get* are considered synonyms since they describe the same frame, thus, their case frame is the same. In our planned semantic lexicon the surface and the deep case frames would be represented together; moreover, if there are some constraints for the possible arguments, they are added to this lexicon too.

The representation of one of the case frames of the verb *get* is a structure like the following one:

```
case_frame1:
literal:  get, got, gotten
category: verb
SUBJ:    recipient
         agentive
ACC:     theme
from:    initiator
         agentive
```

It can be seen that the deep case frame is defined by the thematic roles (see e.g. [4]). In fact, they are basically the same; however, there is a difference between the use of thematic roles in linguistic semantics and in our case. We do not intend to develop or borrow some general system of thematic roles for the whole natural language, but to work out systems of role relations for a specific language usage. For example, if the topic is the production of something, the roles initiator, source, result, ingredient, instrument and goal are used.

The most important step of the semantic search is comparing the representation of the query with fragments of a text. The tokens ("meanings") of the representation of the query are compared with the words in the text, and it is tested whether the syntactic relations between the words match the corresponding semantic relations.

The question is when a word in a query matches a word in the text. Our answer is that the word in the text has to have equal or more information content than the word in the query. The most important cases are as follows:

- synonyms; however, our interpretation of synonymy is more permissive than the generally accepted one – we consider two words (with their case frames) as synonyms if they refer to the same thing or eventuality.
- if A is a kind of B, then B in a query matches A – e.g.: *magic spear* in the text is a relevant hit for *magic weapon*.
- if A (or being A) is a necessary condition of B, then B in a query matches A – e.g.: *find* in the text is relevant to the query *look for*.

Our semantic lexicon is based on sets of synonyms – synsets in WordNet terminology; however, the criterion of synonymy is given by our own definition. The synsets are connected by the relation "having more information content". A top ontology is coupled to the system of synsets. It consists of a general top ontology (the same one for every domain) and a top ontology of the domain. The complexities of connecting synsets to the top ontologies are different in different domains. In the case of production it is not complicated since the role relations are defined in the top ontology of the domain. However, in the domain of fairy tales the ontology of Propp's formalism is the domain ontology [6], and to connect the synsets to it will not be simple.

5. GOALS FOR SEMANTIC PROCESSING

The goals for semantic processing can be summarized as follows:

- Working out the linguistic and methodological foundations of model-based semantic search.
- Elaborating visually based methods for the user interface.

From the point of view of the user, the main novel features of our search engine will be as follows:

- The query is not a Boolean combination of terms (keywords), but well-formed sentences of a controlled natural language, like *The princess sleeps for years*.
- The results given by the search engine to be developed contain phrases which may mean the same thing as the query. The relevance of the results depends on the measure of fit between the meaning of the query and some phrases in the document. Going on with our example: if texts like *The princess sleeps for 100 years* can be found in a document, it will be considered a result of high relevance; however, a tale about a princess who could not sleep because of a pea in her bed will not be given as result.
- The graphical representation of the relevant pieces of text which the engine finds adequate to the query will be shown to the user, so (s)he can easily decide whether (s)he is interested in the document in question.

6. SUMMARY

From the presented processing of folklore texts the semantic search could be the most innovative technology under development. However, other tools like clustering and classification tools can also provide help in information retrieval. Semantic search can be used in different text mining solutions (e.g. information extraction). Researchers have further plans in processing folklore archives, like the semantic tagging of fairy tales, using Propp's morphology [7].

7. ACKNOWLEDGMENTS

This work was supported in part by the Ányos Jedlik Program of the National Office for Research and Technology of the Hungarian government within the framework of the R & D project MASZEKER (Modell Alapú Szemantikus Kereső Rendszer – Model Based Semantic Search System) .

8. REFERENCES

- [1] D. Csendes, J. Csirik, T. Gyimóthy, and A. Kocsor. The Szeged Treebank. In *Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005)*, pages 123–131, September 2005.
- [2] T. Erjavec. *MULTEXT-East morphosyntactic specifications. Version 3*. 2004.
- [3] J. Nivre. *Inductive Dependency Parsing*. Springer Netherlands, 2006.
- [4] T. Parsons. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press, Cambridge, MA, 1990.
- [5] I. Pávai. A néprajzi adatbázis-építés akadályai. *Néprajzi Hírek*, 1(4):86–89, 1996.
- [6] F. Peinado, P. Gervas, and B. Diaz-Agudo. *A Description Logic Ontology for Fairy Tale Generation*. <http://www.fdi.ucm.es/profesor/fpeinado/publications/2004-peinado-description.pdf>, 2004.
- [7] V. Propp. *Morphology of the Folktale*. University of Texas Press; 2 edition (June 3, 2010), US, 2010.

- [8] F. Pusztai, editor. *Értelmező Kéziszótár*. Akadémiai Kiadó, Budapest, 2006.
- [9] K. Verebélyi, editor. *Néphit szövegek*. Magyar Néprajzi Társaság, Budapest, 1998.