SEMANTIC RESOURCES AND THEIR APPLICATIONS IN HUNGARIAN NATURAL LANGUAGE PROCESSING

Theses of the Doctor of Philosophy Dissertation

Márton Miháltz

Supervisor:
Gábor Prószéky, D.Sc.

Multidisciplinary Technical Sciences Doctoral School
Faculty of Information Technology,
Pázmány Péter Catholic University

Budapest, 2010

1. Introduction and Research Aims

Natural language technology (or natural language processing) is a branch of information technology that is interested in developing resources, algorithms and software applications that are able to process ("understand") speech and text formulated in human (natural) languages.

Just as we can distinguish different structural levels in natural languages, we can also define different processing levels in natural language processing. In text processing, these levels could be[1]: segmentation (identifying the sentence and token boundaries within a raw (unprocessed) body of text), morphological analysis/part-of-speech tagging (identifying the morphemes that make up each token, along with all their properties), parsing (identifying structural units of token sequences that make up the sentences), and semantic processing (dealing with the "meaning" of the text: identification of correct word senses of ambiguous words, identifying references within the text or across documents etc.)

In my dissertation, I focused on the latter, semantic aspect of natural language processing, concerning mostly the case of processing texts related to (written in or translated to) Hungarian language.

Semantic processing in NLP may heavily rely on semantic knowledge bases, also called ontologies, that are special databases

[1] Different ways of describing levels of NLP are also possible and there are many other tasks in NLP that are not mentioned here.

that model our knowledge about certain aspects of the real world. In the first part of my work, I focused on examinations concerning one type of ontology formalism called *WordNet*.

WordNet is originally the name of a lexical semantic database developed for the English language at Princeton University [28], [29]. It was built to test and implement linguistic and psycholinguistic theories about the organization of the mental lexicon, modeling the meanings of natural language lexical units (words and multi-words) and their organizational relationships. WordNet can be grasped as a network, where the elementary building blocks are concepts, which are defined by synonym sets (synsets). These are interconnected by a number of semantic relationships, some of them forming a hierarchical network (e.g. the hypernym relationship that would the equivalent of the "is-a" relationship of inheritance networks.)

Soon after the time of its creation, WordNet has proved to be a valuable tool in various natural language processing applications [28], and wordnets for languages other than English have started to be constructed. Projects were launched that aimed to create interconnected the networks of various languages [30], [31].

In the first part of my research, **I was interested in applying and extending existing technologies and finding new methods that aim to aid the creation of a WordNet for Hungarian**. While a reliable semantic resource can only be perfected by human hands, it has been suggested before that this process could be aided by automatic methods [32],[33],[34]. **I experimented with methods to**

extract semantic and structural information from machine-readable dictionaries in order to support the application of the so-called expand model [31] – relying on the conceptual backbone of Princeton WordNet to derive and adapt a wordnet for the semantic characteristics of Hungarian.

The second field of interest in my research focused on **word sense disambiguation** (WSD), which is another aspect of the processing of meaning in natural languages. The aim of WSD is to identify the actual meaning of a semantically ambiguous word in its textual context. The concept of lexical semantic ambiguity is in itself a huge issue in linguistics, covering a spectrum of phenomena ranging from homonymy to polysemy [35], where fine semantic distinctions make it challenging even for humans to define what actual word meanings are. I adopted a pragmatic approach and defined the different senses of a word in language A as the set of possible translations it can have in language B. This approach naturally lends itself for experimentation in machine translation. **I experimented with supervised machine learning methods in the word sense disambiguation of lexical items in a rule-based English-to-Hungarian machine translation system**. Since supervised learning has to rely on a large number of training examples which are costly to produce by human annotators, **I was also interested in developing methods to automate the creation of such training examples by relying on information that can be found in aligned parallel corpora**.

The third subject of my investigations, noun phrase coreference resolution (CR) and possessor identification in Hungarian texts also involved, among other things, the application of (Hungarian) WordNet. The task of NP-CR is to identify groups of noun phrases in a document that refer to the same real-world entities. This task also involves a range of natural language phenomena, of which I attempted to treat the following: coreference expressed by repetition, proper name variants, synonyms, hypernyms and hyponyms, pronouns and zero pronouns.

Possessor identification is a task similar to coreference resolution, but involves the linking of a possessor and possession NP in possessive structures where the two components are separated by several other words and phrases in a sentence.

In both tasks, **I was interested in developing a rule-based system that would integrate different sources of knowledge and different methods for different types of linguistic phenomena in order to achieve high precision and recall, making it suitable for practical NLP applications.**

I also worked on real-life applications of my results in fields like machine translation, information extraction and sentiment analysis. These will be described in more detail in Section 4.

2. Methods Used in the Experiments

During the course of my work, I experimented both with rule-based approaches (designing groups of heuristics, motivated by domain knowledge) and supervised machine learning algorithms. For the

development and evaluation of my methods I generally used hand-annotated example sets and corpora, using precision and recall as main estimates of goodness. I used various NLP tools for pre-processing the various natural language resources (machine-readable dictionaries (MRDs) and corpora) in the course of my work, these will be discussed in detail for each thesis group below.

In the **first part** of my work, I decided to apply the so-called **expand model** of building wordnets, demonstrated by participants in the EuroWordNet project [31]. This involves implementing the English synsets of Princeton WordNet into Hungarian, inheriting the English relations, and then adapting the conceptual hierarchy to suit the specifics of Hungarian. The reason for this choice was the lack of structured semantic resources in Hungarian, required for the other, so-called merge model on the hand, and the possibility of applying automatic methods to speed up the synset translation process on the other hand. It also required the assumption that there would be a sufficient degree of conceptual similarity between English and Hungarian, at least for the part-of-speech of nouns, since they describe physical and abstract entities in a more-or-less common real world (not taking into account cultural differences, of course.)

My goal was to create methods that would aim to attach the entries in the Hungarian side of the available bilingual English-Hungarian MRDs to English synsets in Princeton WordNet. This task involves overcoming two levels of ambiguity. Any Hungarian word *w* may have on average *n* different translations in the bilingual dictionary, and these English equivalents each can belong to *m*

different synsets in Princeton WordNet on average, so the algorithms would need to select the correct synset(s) from *n*m* different possible choices. I used an ensemble of various heuristics that would rely on structural and semantic information found in bilingual and monolingual MRDs in order to get information needed for the disambiguation process.

The **second part** of my work concentrated on the automatic disambiguation of English nouns that have several different possible translations to Hungarian.

I adopted a supervised machine learning approach, where for each ambiguous word, a separate classifier is trained using sense-annotated training examples containing small samples of the contexts of the occurrences. Supervised machine learning methods have shown success in WSD [38], and there are a number of training corpora available for English. Of these, I used the SensEval English lexical sample task dataset [41], the Open Mind Word Expert dataset [40], yielding annotated examples for 45 different polysemous English nouns.

The training data was annotated with Princeton WordNet synsets. In order to have a sense inventory for the English-Hungarian machine translation WSD framework, I manually mapped each English sense to Hungarian translation equivalents. Of the 45 nouns I started with, 34 had less different Hungarian translations than WordNet senses – the Hungarian translation equivalents provided a more coarse-grained sense inventory that subsumed some of the fine-grained WordNet sense distinctions. In the case of 7 further nouns,

all the English senses corresponded to the same Hungarian translation, which meant there was no need for WSD for these, these could be omitted from further experiments. Finally, for 4 nouns the number of English and Hungarian senses was identical. For the rest of the experiment I used 38 nouns where the number of Hungarian equivalents was less or equal to the English senses. On average, each lexical item that was used had 3.97 different senses in WordNet, and after the Hungarian translation, each item had 2.49 different sense tags (Hungarian equivalents), indicating a reduced degree of average ambiguity in the dataset.

The system uses the simple and well-known Naive Bayes classification algorithm, which selects the most probable sense given the joint conditional probabilities of the different senses for the available contextual clues (or features). This learning algorithms was selected after it provided the best precision results in a test of several different supervised learning methods in the Weka environment [58]. The conditional probabilities are estimated from frequencies in the training data. Even though the assumption the algorithm relies on—that contextual features are independent statistical variables—does not hold for natural language data, this method has proved to be successful in WSD [37], [38].

To train the classifiers I used learning features identified from the context of the ambiguous words based on [37] and [39], that can be grouped into two types. The first type of features is taken only from the sentence containing the ambiguous word, with order and relative position being significant. These features represent the syntactic

properties of the context, frequent collocations, modifiers etc. They include the surface form of the ambiguous word, function words from a 2+2 window around the ambiguous word, and content words from a 3+3 window. The other group of features represents the semantic domain, or topic of the entire available context (usually the paragraph containing the ambiguous word). This information is represented by a binary vector that codes the presence of certain frequent content words in the context.

Since semantically annotated training corpora are available only in limited quantity, I needed a solution for scaling the system up. One possibility is to annotate the occurrences of a polysemous item extracted from a corpus with sense tags (target language translations) by hand. However, such corpus annotation is a highly time-consuming, thus costly procedure. Another, more favorable alternative is to use a parallel corpus: appropriate training material can be produced by identifying the translations in sentence-aligned bitexts [45], [48].

The Hunglish Corpus [49] is the largest accurately sentence-aligned English–Hungarian parallel corpus currently available, with 44.6 million English and 34.6 million Hungarian words from 5 genres of text. I processed the English texts in the corpus with a PoS-tagger [46], and used the Humor morphological analyzer [42] and the output of the POS-tagger to get the stem the English word forms, and also to stem the word forms in the Hungarian texts.

I experimented with the polysemous English noun *state* to explore the problems that would arise when producing automatically tagged training corpora for an English-to-Hungarian MT system.

I first identified corpus occurrences containing lexicalized multi-word expressions formed by *state* in the English side. The target word in these collocations always has the same meaning, regardless of context, so the collocation can be unambiguously translated by simple lexical transfer rules. I compiled a list of possible English nominal multi-word lexical items formed by *state* from several lexical resources: a comprehensive English-Hungarian bilingual dictionary [47], Princeton WordNet version 2.1, and the lexical translation pattern database of the MetaMorpho MT system [43]. I also applied *Terminology Extractor* (version 3.0c, Copyright (C) 2002 Chamblon Systems Inc.) to the English side of the corpus to find salient collocations formed by *state* (the output was manually revised). A total of 348 different collocations were identified.

With the help of the bilingual dictionary, I also compiled a list of all the possible Hungarian translations of the noun *state* in its single-word usage, gaining 19 different translations.

I created a sub-corpus of the Hunglish corpus by selecting sentence pairs where the English sentence contained the noun *state* (92,500 sentence pairs). I then grouped these sentence pairs into 3 classes: a) sentence pairs that contained one or more of the known collocations (93%), b) sentence pairs that contained one or more of the known collocations in addition to other occurrences of *state* (3%), and c) sentence pairs that contained only unknown

occurrences (none of the known collocations) (4%). In categories b) and c) I looked for 0, 1 or more occurrences of any of *state*'s 19 known Hungarian equivalents.

Sentence pairs containing exactly 1 translation equivalent on the Hungarian side, without any additional collocational occurrences constituted 2,473 training examples (the most frequent sense represented by 1,296 examples.) Previous experiments showed that this quantity is sufficient for training a high-quality classifier for WSD.

The **third part** of my work focused on the identification of coreference and possession relationships between entities (noun phrases) in Hungarian texts.

In recent work in the field of coreference resolution (CR), data-driven, machine learning-based approached have gained ground over traditional knowledge-based systems [52]. However, such an approach requires an extensive number of hand-labeled training examples, which is not available at present for Hungarian, therefore I had to commit myself to a rule-based approach.

My proposed system relies on several sources of knowledge: the morphological, syntactic and semantic information available from the output of the MetaMorpho MT system's deep parser [43], [44]; rules based on Binding Theory in Hungarian syntax [50] and the results of psycholinguistic research on Hungarian sentence understanding [53], [54]; rules based on semantic information available from the Hungarian WordNet [6]; and finally, I also

employed character-based heuristics, similar to some of those described by [55].

The MetaMorpho parser is used to identify paragraph, sentence and token boundaries, clauses, maximal noun and verb phrases, and to provide morphological, grammatical and semantic information for these units. After pre-processing, the system processes each anaphoric NP in the document from left to right and tries to identify the coreferring antecedent that is closest to it.

In Hungarian, there are three basic **possessive structures**, when the possessor and the possession can be detached. Two of these phenomena (possession predicates, detached dative-case possessors) can be handled by syntactic constraints (as demonstrated by the MetaMorpho [44] parser's grammar), but the third type (zero-pronoun possessor) can only be treated by methods similar to zero pronoun resolution, for which I proposed a rule-based solution.

3. New Scientific Results

Thesis Group I: Methods for the Automatic Construction of Hungarian WordNet Ontology.

I.1. I showed that the expand model can be successfully applied to automatically aid the construction of a wordnet for Hungarian.

The first group of heuristics for automatic synset translation were proposed by [32], [33] for the construction of the Spanish and

Catalan wordnets using the expand methodology. The Variant, Mono and Intersection methods used only structural information in the bilingual MRDs and PWN. A fourth method, proposed by [32] relies on semantic information extracted from a monolingual (explanatory) dictionary: definitions were parsed and a genus proximum word was extracted for each headword. The so-called conceptual distance formula [32] was then applied on the headword and the genus in order to get a PWN synset target for the headword.

To make the application of the last method possible, I processed an electronic version of the Hungarian explanatory dictionary Magyar Értelmező Kéziszótár (EKSz) [36]. I used manually written patterns to extract the genus proximum, synonyms and meronym/holonym terms for the noun headwords from their definition sentences, which were pre-processed by the HuMor Hungarian morphological analyzer [42] and a simple regexp-based tokenizer developed at MorphoLogic.

I used two evaluation methods in order to assess the performance of my own heuristics and the ones proposed by [32], [33] on Hungarian data. In the first method, I manually disambiguated 400 Hungarian nouns, randomly selected from the bilingual MRD, against their possible PWN synsets (total 2,201) and calculated precision and recall of the proposed connections for each heuristic using this set. The methods from [32], [33] in my implementation ranged in precision 49-92%, while [32] reports 61-85% on the manual evaluation of a 10% sample. Following [33], I also experimented with different combinations of the methods. This way

I was able to obtain a preliminary set of 10,786 Hungarian synsets, containing 9,986 words with an estimated average precision of 75%, while [33] reports 6,551 Spanish synsets, containing 7,922 words, with an estimated average precision of 75%.

I.2. I proposed 4 new heuristics for the automatic construction of Hungarian synsets in the expand model. The methods disambiguate Hungarian nouns against English synsets, and rely on the special properties of the Hungarian language and the available resources.

Besides applying the above-mentioned four heuristics to Hungarian, I also created several new heuristics:

- Using a variation of the intersection method, I used synonyms acquired from the monolingual dictionary and available from a thesaurus to assign a Hungarian word to the PWN synset which contains the greatest number of the synonyms' English translations.
- I used the morphological analyzer to identify the head of endocentric N+N compounds, which can be treated as "derivational" hypernyms, making the application of the conceptual distance formula possible. I also applied this method to Hungarian nominal multiword expressions where the last token was a noun.
- I used the Latin equivalents available for a number of EKSz headwords (animal or plant species, taxonomic groups,

diseases etc.) as an interlingua, since PWN synsets directly contain Latin synonyms for such English concepts.

- To increase coverage, in the cases where the application of the conceptual distance formula was not possible due to lack of translation of the genus/synonym in the bilingual dictionary, I used the transitive property of the hypernymy and synonymy relations. I tried to use either the derivational hypernyms, or the extracted hypernyms (genuses) of such synonyms/genus words (in the latter case only if the genus/synonym was not ambiguous in EKSz.)

In a second round of evaluation, I was interested in the precision and recall of my methods in the perspective of the final, human-edited Hungarian WordNet (HuWN) ontology, containing about 42,000 Hungarian synsets, prepared during the Hungarian WordNet project [6], [10], [12]. During the project, a number of human annotators used the results of my synset machine translation heuristics as a starting point, and were free to edit, delete, extend etc. the proposed synsets and restructure the relations inherited from Princeton WordNet 2.0.

I calculated precision as the ratio of the number of translation links (<Hungarian lexical item, Princeton WordNet 2.0 synset> pairs) proposed by the heuristics *and* approved (i.e. not deleted) by the humans annotators, to the total number of links proposed by the heuristics. I defined recall as the ratio of proposed and approved links to all the approved links within the synsets the heuristics

attempted to translate. These measures were calculated for the automatically generated translations for all affected parts of speech in HuWN (nouns, verbs, adjectives). A summary of the results can be seen in Table 1.

	All	**Nouns**	**Verbs**	**Adjectives**
Precision	24.61%	31.53%	13.89%	17.36%
Recall	64.81%	63.77%	64.46%	71.96%

Table 1: Evaluation results of synset translation methods against Hungarian WordNet

Thesis Group II: Supervised word sense disambiguation for English-Hungarian machine translation.

II.1. I proposed a word sense disambiguation system that can be used to improve the lexical translation accuracy of rule-based English-Hungarian machine translation. Without WSD, the baseline MT system would translate polysemous source words to their most frequent sense target language equivalents.

I performed evaluation of the word sense disambiguation classifiers by doing 10-fold stratified cross-validation on the training corpora for the 38 ambiguous nouns. Precision is defined as the ratio of correctly classified instances to all instances to be classified. I took baseline score to be the relative frequency of the most frequent sense in each case.

Evaluation was performed both on the disambiguation of English senses and on the disambiguation of mapped Hungarian translations. In the case of English senses, average precision was 77.99%, the baseline score being 64.16% on average. For the Hungarian translations, the classifiers produced 85.00% precision on average, an average 11.52% improvement over the baseline. In the latter case, all but 10 of the 38 classifiers performed above the baseline, and in only 1 case did the precision fall below the baseline.

II.2. By mapping the English WordNet sense inventory to Hungarian translations, the average number of senses can be reduced and the precision of disambiguation can be improved in comparison to monolingual WordNet senses-based WSD.

The fine-grained sense distinctions in WordNet make it difficult to construct high-performance word sense disambiguation methods when using WordNet synsets as a sense inventory. Since most Hungarian translations possess a degree of polysemy, mapping the WordNet senses to Hungarian translations produced a lower number of sense classes. Mapping the English senses to Hungarian translations improved precision of the classifiers 7.01% overall. In 27 cases out of 38, the precision was higher with Hungarian translations, while in 11 cases precision did not change.

II.3. I showed that annotated training examples for word sense disambiguation in English-Hungarian machine translation can be produced using a large, aligned parallel corpus using considerably less resources than manual corpus annotation. In this approach it is essential to recognize idiomatic multi-word expressions formed with the target word in the corpus.

My experiment with the Hunglish corpus showed that to produce WSD training examples one needs: 1) the set of possible translation equivalents, for example from a bilingual dictionary, 2) a set of multi-word expressions formed by the ambiguous word, from various available lexical resources, or by using corpus-based collocation identification methods. After filtering out ambiguous instances, the large numbers of the Hunglish corpus (2 million sentence pairs) can still provide a sufficient number of labeled examples for training the supervised WSD classifiers (2,473 instances for *state*, plus 1,334.instances are also available that contain a collocation and exactly one translation.)

Thesis Group III: Rule-based coreference and possessor identification in Hungarian.

III.1. I proposed an algorithm based on several knowledge sources and heuristics for recognizing parser errors for the resolution of coreference relationships between noun phrases in Hungarian texts.

Coreference resolution for a given NP in the input document is based on satisfying constraints and evaluating preferences [51]. The algorithm for generating the list of antecedent candidates, filtering the list and finally selecting the winning candidate is specific to the type of the anaphoric NP.

For **proper names**, the list of antecedent candidates consists of all the proper names prior to the anaphor in the entire document. The most likely antecedent candidate is the one having smallest Minimum Edit Distance (MED) from the anaphor, using normalization (removing front determiners, stemming the head) and a preset threshold, so the system is not forced to select one from the available candidates.

For **common nouns with a definite article**, the algorithm first tries to exclude mentions that refer to unique objects inferable from common world knowledge, by searching a predefined list. Antecedent candidates are the proper names and common nouns in the preceding part of the paragraph of the anaphor, up to the VP containing it (Binding Theory excludes candidates dominated by the

main verb in the anaphor's VP.) Selection of the antecedent is done by identifying the closest candidate that has the same head, or the closest synonym or hypernym/hyponym, using Hungarian WordNet and the Leacock-Chodorow similarity formula [37].

The system also deals with **personal pronouns**, with the addition of *az* ("that") demonstrative pronoun in subject position and not referring to a subordinate relative clause. The antecedent candidates are collected from the 2 sentences before the anaphor's sentence (if they exists) plus the clauses prior to the clause containing the anaphor in its sentence. The candidates are filtered by checking person, number, 2 semantic features (*animate* and *human*) and by excluding candidates that have already been identified as antecedents of other NPs in the current clause (Binding Theory.) Multiple pronominal anaphors in a clause are processed in obliqueness order to rule out already bound candidates. Resolution for common nouns and proper names is performed before pronouns within a sentence to further help resolution of pronouns by eliminating some of the possible antecedents.

Identifying the antecedent of the pronoun or zero pronoun that is the subject in its VP follows research on Hungarian psycholinguistics [53], [54]. The algorithm assumes parallel grammatical functions across sentences, where the subject is preserved from the previous clause/sentence. This is overridden by the presence of the demonstrative pronoun *az* in subject position, indicating change of subject. In case of multiple non-subject NPs in the prior clause, the antecedent is selected using the obliqueness

hierarchy and by checking distance from the anaphor (NPs closer to the end of the sentence are preferred). Resolution of pronouns and zero pronouns with grammatical roles other than subject are based on the obliqueness hierarchy and closeness to the anaphor.

For the **evaluation** of the coreference resolution algorithm, I prepared a small hand-tagged corpus (10 text segments, total 99 sentences, 1240 words, 81 annotated NPs.) Average precision of coreference resolution was 68.92%, average recall was 62.96% on this corpus. For the most frequent types of anaphora, precision was between 71-80%, while recall was between 61-83%. The WordNet-based methods, using hypernym and synonym information showed a poor performance (0-33% F-measure), but since they were represented by only 6 instances in the corpus, the evaluation figures might not be realistic.

I also performed an evaluation of the error types produced by the algorithm, which showed that for pronouns (the most frequent type of anaphora in the corpus) nearly half of the mistakes were due to errors in the parser's output. Perfectly parsed input would increase overall precision to 75%, pronoun/zero pronoun resolution precision to 91%.

III.2. I proposed a rule-based method, similar to pronominal anaphora resolution for the identification of detached possessor-possession structures in Hungarian.

I relied on the assumptions that 1) the subject of the possession NP's dominating verbal phrase is the default possessor, 2) the possessor noun phrase matches in grammatical number and person to the possession NP's owner number and person, carried by morphological information in Hungarian. The second assumption can override the 1st, so when the subject of the possession's VP does not match in number/person, the previous clause's subject can be the possessor, if it's still in the same discourse segment.

My possessor identification algorithm is therefore implemented as follows: noun phrases, in up to the -2nd sentence before the clause of the possessor but not further than the 1st sentence in the containing paragraph, that are subjects in their clause and match in number and person to the possession are identified, and the one that is closest to the possessor is picked. If no sentence-level parse, therefore no grammatical role information is available in the parser's output, the rightmost NP before the possession with nominative case and matching number and person is selected.

The evaluation of the algorithm was carried out on the same corpus as the coreference resolution (38 detached possessive structures were annotated by hand.) Precision of possessor-possession identification was 76.47%, recall was 68.42% (F-measure 72.22%) on this corpus.

4. Applications of the Results

All of the work discussed in the dissertation was related to projects where practical applications of my results were carried out.

The methods proposed for the automatic construction of a **Hungarian WordNet** ontology were implemented and applied in the Hungarian WordNet project [6] (2005-2007), funded by the European Union ECOP program (GVOP-AKF-2004-3.1.1.) with the participation of several Hungarian academic and industrial partners (Research Institute for Linguistics of the Hungarian Academy of Sciences, Department of Informatics, University of Szeged, and MorphoLogic Ltd.) with the aim of producing a WordNet ontology for the Hungarian language. The project used Princeton WordNet 2.0 as a basis of the expand approach, and used my heuristics to automatically generate translations of noun and adjective synsets, which were edited and corrected by human annotators for the final ontology. The project ended with a Hungarian WordNet containing more than 40,000 synsets.

The resulting ontology was used in an **information extraction** project as well [6]. I developed a system for the frame-based extraction of information from short business news articles. 124 event frames based on verb frames, morphological and semantic constraints were prepared manually and were used by the IE system utilizing partial and full parses of the MetaMorpho parser [43], [44]. The semantic constraints were formulated by mapping semantic classes used in the event frames to hierarchies in the nominal Hungarian WordNet ontology.

The **word sense disambiguation** system described in the dissertation was designed specifically for MorphoLogic's MetaMorpho English-Hungarian machine translation system [43],

where manually created context-free grammar analysis and translation rules only code a limited amount of semantic information, therefore external help is needed from an “oracle” that can make a decision about the proper senses by looking at the available context. A WSD module using the methods described in the dissertation was integrated into the MetaMorpho engine, operating after a source language paragraph has been preprocessed (segmentation, tokenization, morphological analysis and word stemming). The WSD module specifies the value of a grammar feature that indicates the actual sense of a recognized ambiguous word. In the subsequent steps of the source-language analysis, the syntactic parser can rely on the value of this semantic feature. At the target language translation generation phase a branching algorithm uses the sense identifier feature in order to select the correct translation. The mapping between English senses and Hungarian translations is represented in the translation grammar rules, which allows for easy manual editing.

The Hungarian **coreference and possessor resolution** methods proposed in the dissertation were incorporated into the psychological content analysis system developed in the project *A Narrative Study of National and Ethnic Identity* [57], realized by a group of Hungarian institutions (Research Institute for Psychology of the Hungarian Academy of Sciences, Research Institute for Linguistics of the Hungarian Academy of Sciences, Department of Informatics, University of Szeged, MorphoLogic Ltd, and the University of Pécs) between 2006-2008, sponsored by the National Office for Research and Technology in Hungary (NKFP6 00074/2005, Jedlik Ányos

Program.) In the project, a corpus of history textbooks were annotated with syntactic, morphological and semantic information (phrases, grammatical roles, thematic roles and semantic types). The corpus served as a basis for special queries that examined the distributional properties of special patterns in the project's focus. Coreference and possessor identification was successfully applied to increase the coverage of the study by adding coreferring mentions of the entities used in the queries.

5. Acknowledgements

I would like to say thank you to my supervisor Gábor Prószéky. I am grateful to all my colleagues who contributed ideas and useful comments to my work: Csaba Hatvani, Judit Kuti, Csaba Merényi, Mátyás Naszódi, Gábor Pohl, Péter Schönhofen, György Szarvas, László Tihanyi, Péter Vajda, Károly Varasdi and many others. I would like to express my gratitude to the Doctoral School of the Faculty of Information Technology at Pázmány Péter Catholic University for providing me with the opportunity to conduct my research. I am indebted to Gábor Vásárhelyi, Éva Bankó and the other students at the Doctoral School who provided valuable pieces of information that helped the completion of my dissertation. And last but not least, I am especially thankful to all of my friends and family who supported me.

Work covered in this dissertation was supported partly by the GVOP-AKF-2004-3.1.1. and NKFP6 00074/2005 (Jedlik Ányos Program) projects.

6. List of Publications

The Author's Journal Publications

[1] **Miháltz Márton**: Tudásalapú koreferencia- és birtokosviszony-feloldás magyar szövegekben. To appear in: *Általános Nyelvészeti Tanulmányok*

[2] Prószéky, Gábor, **Miháltz Márton**: Magyar WordNet: az első magyar lexikális szemantikai adatbázis. In: *Magyar Terminológia* 1 (2008) 1, pp. 43-57.

[3] Németh, Dezső, Ivády Eszter Rozália, **Miháltz Márton**, Krajcsi Attila, Pléh Csaba: A verbális munkamemória és morfológiai komplexitás. In *Magyar Pszichológiai Szemle*. 61. évf., 2. szám, pp. 265-298.

The Author's Conference Publications

[4] **Miháltz Márton**: Információ-kivonatolás szabad szövegekből szabályalapú és gépi tanulásos módszerekkel. In: *VI. Magyar Számítógépes Nyelvészeti Konferencia* kiadványa, Szeged, pp.49-58, 2009.

[5] **Miháltz, Márton**: Knowledge-based Coreference Resolution for Hungarian. In: *Proceedings of The Sixth International Conference on Language Resources and Evaluation* (LREC 2008), Marrakesh, Morocco, 2008.

[6] **Miháltz, Márton**, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, Tamás Váradi: Methods and Results of the Hungarian WordNet Project. In: *Proceedings of The Fourth Global WordNet Conference*, Szeged, Hungary (2008), pp. 311–321.

[7] **Miháltz Márton**, Naszódi Mátyás, Vajda Péter, Varasdi Károly: NP-koreferenciák feloldása magyar szövegekben a Magyar WordNet ontológia segítségével. In: *V. Magyar Számítógépes Nyelvészeti Konferencia kiadványa*, Szeged (2007), pp. 138–146.

[8] Hatvani Csaba, Kocsor András, **Miháltz Márton**, Szarvas György, Szécsi Katalin: Főnevek a Magyar WordNetben. *IV. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, pp. 109-116.

[9] **Miháltz, Márton**, Gábor Pohl: Exploiting Parallel Corpora for Supervised Word-Sense Disambiguation in English-Hungarian Machine Translation. *Proceedings of the 5th Conference on Language Resources and Evaluation*, 1294–1297. Genoa, Italy (2006)

[10] Alexin, Zoltán, János Csirik, György Szarvas, András Kocsor, **Márton Miháltz**: Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. In *Proceedings of the Third*

International WordNet Conference (GWC 2006), Seogwipo, Jeju Island, Korea, January 22-26, 2006, pp. 291-292.

[11] **Miháltz Márton**, Pohl Gábor: Javaslat szemantikailag annotált többnyelvű tanítókorpuszok automatikus előállítására jelentés-egyértelműsítéshez párhuzamos korpuszokból. *III. Magyar számítógépes nyelvészeti konferencia*, Szeged, 2005. december 8-9, pp. 418-419.

[12] **Miháltz Márton**, 2005: Magyar EuroWordNet projekt: bemutatás és helyzetjelentés. *III. Magyar számítógépes nyelvészeti konferencia*, Szeged, 2005. december 8-9, pp.68-78.

[13] **Miháltz, Márton**, 2005: Towards A Hybrid Approach To Word-Sense Disambiguation In *Machine Translation. Workshop „Modern Approaches in Translation Technologies" at Recent Advances in Natural Language Processing (RANLP-2005) Conference*, Borovets, Bulgaria.

[14] Németh, Dezső, Iványi Eszter Rozália, **Miháltz Márton**, Pléh Csaba: "Phonological loop and morphological complexity" XIVth ESCOP - *Conference of European Society for Cognitive Psychology*, August 31 - September 3, 2005, Leiden

[15] **Miháltz Márton**, 2004: Angol-magyar gépi fordítórendszer támogatása jelentés-egyértelműsítő modullal. *Második Magyar Számítógépes Nyelvészeti Konferencia* (MSzNy-2004), Szeged, pp. 92-99.

[16] **Miháltz, Márton**, 2004: Word Sense Disambiguation Using Random Indexing. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

[17] **Miháltz, Márton**, Gábor Prószéky, 2004: Results and Evaluation of Hungarian Nominal WordNet v1.0. In *Proceedings of the Second International WordNet Conference* (GWC 2004), Brno, Czech Republic, pp. 175-180.

[18] **Miháltz, Márton**, 2003: Magyar főnévi WordNet létrehozása automatikus módszerekkel (Constructing a Hungarian WordNet Ontology with Automatic Methods). *Első Magyar Számítógépes Nyelvészeti Konferencia* (MSzNy-2003), Szeged, pp. 153-160.

[19] **Miháltz, Márton**, 2003: Constructing a Hungarian ontology using automatically acquired semantic information. In *Proceedings of the 5th International Workshop on Computational Semantics* (IWCS-5), Tilburg, The Netherlands, pp. 475-478.

[20] Prószéky, Gábor and **Márton Miháltz**, 2002: Automatism and User Interaction: Building a Hungarian WordNet. In *Proceedings of the*

Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spain, Vol 3, pp. 957-961.

[21] Prószéky, Gábor and **Márton Mihaltz**, 2002: Semi-Automatic Development of the Hungarian WordNet. In *Proceedings of the LREC 2002 Workshop on WordNet Structures And Standardization, And How These Affect WordNet Applications And Evaluation*, Las Palmas de Gran Canaria, Spain, pp. 42-46.

[22] Prószéky, Gábor, **Márton Mihaltz** and Dániel Nagy, 2001: Toward a Hungarian WordNet. In Proceedings of the NAACL 2001. *Proc. Workshop on WordNet and Other Lexical Resources*, Pittsburgh, USA, pp.174-176.

The Author's Other Publications

[23] **Mihaltz, Márton**: Development of the Hungarian WordNet Ontology and its Application to Information Extraction. Presentation at the *10th International Protégé Conference*, Budapest, Hungary (2007)

[24] **Mihaltz Márton**, Prószéky Gábor: Egy magyar WordNet felé. Előadás a *W3C Szemantikus Web Műhelykonferencián*, MTA SZTAKI W3C Magyar Iroda, Budapest, 2006. április 13.

[25] Németh, Dezső, Rozália Eszter Ivády, **Márton Mihaltz**, Attila Krajcsi, Csaba Pléh, 2005: Verbal Working Memory And Morphology. Poster at the *9th European Congress of Psychology*, Granada, Spain.

[26] Ivády Rozália Eszter, Németh Dezső, **Mihaltz Márton**, Pléh Csaba, 2004: Fonológiai hurok és morfológia komplexitás. Magyar Pszichológiai Társaság Biennális Nagygyűlése, Debrecen, 2004.

[27] Ivády R. E., **Mihaltz M.**, Németh D., Pléh Cs. (2004). A rövidtávú emlékezet és morfológiai komplexitás. In Németh D. (szerk.). *Szegedi Pszichológiai Tanulmányok*, JGYTF Kiadó, Szeged, pp. 21-32.

7. Works Cited

[28] Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller: Introduction to WordNet: an on-line lexical database. Int. J. of Lexicography 3 (1990) 235–244.

[29] Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998)

[30] Tufiş, D., Cristea, D., Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal of Information Science and Technology Special Issue, vol. 7, no. 1-2 (2004)

[31] Vossen, P. (ed.): EuroWordNet General Document, Version 3. University of Amsterdam (1999)

[32] Atserias, J., S., Climent, X., Farreres, G., Rigau, H., Rodríguez: Combining multiple methods for the automatic construction of multilingual WordNets. Proc. of Int. Conf. on Recent Advances in Natural Language Processing, Tzigov Chark (1997)

[33] Farreres, X., G., Rigau, H., Rodriguez: Using WordNet for building Wordnets. Proc. of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal (1998)

[34] Eduard Barbu, Verginica Barbu Mititelu, Automatic Building of Wordnets. In N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov (Eds.), Recent Advances in Natural Language Processing IV (RANLP-05), 2005.

[35] Kiefer Ferenc (2001). Jelentéstan. Corvina, Budapest.

[36] Juhász, J., I., Szőke, G. O. Nagy, M. Kovalovszky (eds.): Magyar Értelmező Kéziszótár. Akadémiai Kiadó, Budapest (1972)

[37] Leacock, C., Miller, G. A., Chodorow, M.: Using Corpus Statistics and WordNet Relations for Sense Identification. Computational Linguistics, Special Issue on Word Sense Disambiguation. (1998)

[38] Manning, C. D., Schütze, H: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)

[39] Mihalcea, Rada Word sense disambiguation with pattern learning and automatic featureselection. Journal of Natural Language Engineering (special issue on evaluating word sense disambiguation systems, 8 (4) 279-291 (2002)

[40] Mihalcea, R., Chklovski, T.: Building a Sense Tagged Corpus with Open Mind Word Expert. Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions (2002)

[41] Mihalcea, R., Chklovski, T. and Kilgarriff, A.: The Senseval-3 English Lexical Sample Task. In Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Barcelona, Spain (2004)

[42] Prószéky, G.: Humor: a Morphological System for Corpus Analysis. Language Resources and Language Technology, Tihany (1996)

[43] Prószéky, G., Tihanyi, L.: MetaMorpho: A Pattern-based Machine Translation Project. Proceedings of the 24th 'Translating and the Computer' Conference. London, UK, 19–24 (2002)

[44] Prószéky, Gábor; László Tihanyi; Gábor Ugray: Moose: a robust high-performance parser and generator. Proceedings of the 9th Workshop of

the European Association for Machine Translation, Foundation for International Studies, La Valletta, Malta, pp. 138–142 (2004)

[45] Diab, M. (2004): Relieving the data acquisition bottleneck for Word Sense Disambiguation. In Proceedings of ACL 2004.

[46] Giménez, J., L. Márquez: SVMTool (2004): A general POS tagger generator based on Support Vector Machines. In Proceedings of the 4th International Conference on Language Re-sources and Evaluation (LREC'04). Lisbon, Portugal.

[47] Országh, L., Magay, T. (2004): Angol-magyar nagyszótár. Budapest: Akadémiai Kiadó.

[48] Specia, L., M. G. Volpe Nunes, M. Stevenson (2005): Exploiting Parallel Texts to Produce a Multi-lingual Sense Tagged Corpus for Word Sense Disambiguation. In Proceedings of Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria

[49] Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón (2005): Parallel corpora for medium density languages. In Proceedings of Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria.

[50] Kenesei István: Az alárendelt mondatok szerkezete. In: Kiefer Ferenc (szerk.): Strukturális Magyar Nyelvtan, I. kötet, Mondattan. Akadémiai Kiadó, Budapest (1992)

[51] Mitkov, Ruslan: Anaphora Resolution: The State of The Art. Working Paper, University of Wolverhampton, 1999.

[52] Ng, Vincent: Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. Proceeding of the 43rd Annual Meeting of the Associaction for Computational Linguistics (2005)

[53] Pléh Csaba, Radics Katalin: „Hiányos mondat", pronominalizáció és a szöveg. In Általános Nyelvészeti Tanulmányok, XI, 261-277 (1976).

[54] Pléh Csaba: Mondatközi viszonyok feldolgozása: az anafora megértése a magyarban. In: Pléh Csaba: Mondatmegértés a magyar nyelvben. Osiris Kiadó, Budapest (1998)

[55] Uryupina, Olga: Evaluating Name-Matching for Coreference Resolution. In Proceedings of the 4th International Conference on Language Resources and Evaluation (2004)

[56] Varasdi Károly: Koreferenciák feloldása. MTA Nyelvtudományi Intézet (2005)

[57] Szalai Katalin, Ferenczhalmy Réka, Fülöp Éva, Vincze Orsolya, László János: Történelmi szövegek narratív pszichológiai vizsgálata a nemzeti identitás tükrében. In VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2009, pp. 259–271.

[58] Witten, I. H., E. Frank, *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2005.