DOI:10.15774/PPKE.ITK.2010.006

SEMANTIC RESOURCES AND THEIR APPLICATIONS IN HUNGARIAN NATURAL LANGUAGE PROCESSING

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

DOI:10.15774/PPKE.ITK.2010.006

I dedicate this work to my Grandfather

DOI:10.15774/PPKE.ITK.2010.006

Acknowledgments

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: Gábor Pohl, Csaba Merényi, Judit Kuti, Károly Varasdi, Csaba Hatvani, György Szarvas, Mátyás Naszódi, László Tihanyi 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.

DOI:10.15774/PPKE.ITK.2010.006

SEMANTIC RESOURCES AND THEIR APPLICATIONS IN HUNGARIAN NATURAL LANGUAGE PROCESSING

by Márton Miháltz

Abstract

This thesis is about the creation and application of semantic resources in Hungarian natural language processing. The first part of my work deals with applying automatic methods in order to generate a WordNet ontology – a hierarchical lexicon of word meanings – for Hungarian. I used several methods to generate automatic Hungarian translation for English WordNet synsets in what is called the expand model of building wordnets. I applied methods described in the literature and also developed new ones specific to Hungarian based on the available machine-readable dictionaries and other resources. The second part of my work focuses on word meaning in the context of polysemy and machine translation. I developed a word-sense disambiguation system to improve the lexical translation quality of an English-to-Hungarian rule-based machine translation system. My WSD system uses classifiers applying supervised machine learning, each trained with local and global features extracted from the training contexts. The classes are Hungarian translations of the ambiguous English lexical items, which improves disambiguation accuracy. I also showed a way to semi-automatically generate training instances for such classifiers using an aligned parallel corpus. In this approach, I have shown that it is essential to recognize idiomatic multi-word expressions formed with the target word in the corpus. In the third part of my work, I proposed a system for noun-phrase coreference- and possessor-relationship resolution in Hungarian texts. The system uses rules relying on several knowledge sources, among them Hungarian WordNet. I also present shortly applications of my results both in research & development and in industrial projects.

DOI:10.15774/PPKE.ITK.2010.006

TABLE OF CONTENTS

DOI:10.15774/PPKE.ITK.2010.006

DOI:10.15774/PPKE.ITK.2010.006

Chapter 1

INTRODUCTION

Natural language technology (natural language processing) is a branch of computer science 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, token and named entity 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 have 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 that model our knowledge about certain aspects of the real world. In the first part of my work, I have 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 [35], [36]. 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

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

DOI:10.15774/PPKE.ITK.2010.006

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 [36], [31], and wordnets for languages other than English have started to be constructed. Projects were launched that aimed to create interconnected semantic networks for various languages [30], [41], [43], [56].

1.1. Research Aims

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 [30], [31], [56]. **I have 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** [41] – 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 [67], where fine semantic distinctions make it challenging even for humans to define what actual word meanings are. I have 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 have 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**.

DOI:10.15774/PPKE.ITK.2010.006

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 have 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.

1.2. Methods of Investigation

In 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.

The remaining part of the dissertation is organized as follows: in the next 3 chapters, I will present the background, the experiments and the results for the topics of wordnet construction, word sense disambiguation and coreference resolution. In Chapter 5, I present the concise summary of my new scientific results in the form of theses. A brief description of the application of my results in real-life projects also follows.

DOI:10.15774/PPKE.ITK.2010.006

1.3. Abbreviations

The following abbreviations are used throughout the dissertation:

Abbreviation	**Resolution**
BCS	BalkaNet Base Concept (Set)
BILI	BalkaNet Inter-lingual Index
BN	BalkaNet
CBC	Common Base Concept (Set)
CR	Coreference resolution
EWN	EuroWordNet
HuWN	Hungarian WordNet
ILI	Inter-Lingual Index
MRD	Machine-readable dictionary
MT	Machine translation
NLP	Natural language processing
NP	Noun phrase
OMWE	Open Mind Word Expert
PoS	part-of-speech
PWN	Princeton WordNet
RI	Random Indexing
TC	Top Concept
TO	Top Ontology
VP	Verb phrase
WN	WordNet
WSD	Word sense disambiguation

DOI:10.15774/PPKE.ITK.2010.006

Chapter 2

METHODS FOR THE CONSTRUCTION OF HUNGARIAN WORDNET

2.1. Introduction

Ontologies are widely used in knowledge engineering, artificial intelligence and computer science, in applications related to knowledge management, natural language processing, e-commerce, bio-informatics etc. [28]. The word ontology is borrowed from philosophy, where it means a systematic explanation of being [28]. In the above-mentioned fields of information technology there are many definitions of what ontologies are. I would like to cite the following definition by [29]:

> *An ontology is a formal, explicit specification of a shared conceptualization. Conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable. Shared reflects the notion that an ontology captures consensual knowledge, that it is not private of some individual, but accepted by a group.*

The notion of ontologies is often not distinguished from the notion of taxonomies, which only include concepts, their hierarchical structure, the relationships between them and the properties that describe them. The knowledge engineering community therefore calls the latter *lightweight ontologies*, while the former *heavyweight ontologies,* differing in the property that they also add axioms and constraints to clarify the intended meaning of the collected terms [28]. Ontologies can be modeled with a variety of different tools (frames, first-order predicate logic, description logic etc.) and could be classified based on various criteria: richness of content (vocabularies, glossaries, thesauri, informal and formal is-a hierarchies etc.), or subject of conceptualization (knowledge representation ontologies, general/common ontologies, top-level/upper-level ontologies, domain ontologies, task ontologies etc.) [28].

Linguistic ontologies model the semantics of natural languages, not just the knowledge of a specific domain. They are bound to the grammatical units of natural languages (“words”, multiword lexemes etc.) and are used mostly in natural language processing. Some linguistic ontologies depend entirely on a single language (e.g., Princeton

DOI:10.15774/PPKE.ITK.2010.006

WordNet), while others are multilingual (EuroWordNet, Generalized Upper Model etc.). They can also differ in origins and motivations: lexical databases (e.g., wordnets), ontologies for machine translations (e.g., SENSUS) etc. [28].

In natural language processing, knowledge-based applications like word sense disambiguation, machine translation, information retrieval, coreference resolution etc. (or knowledge-based approaches to these) can benefit from ontologies [31]. There are a number of different ontologies available (GUM, CYC, ONTOS, MIKROKOSMOS, SENSUS etc. [28]) that differ in scope, coverage, domain, granularity, relations etc. [31] WordNet, however, has become a de facto standard [31], [56], possibly due both to its large coverage and its unrestricted availability[2].

2.1.1. Princeton WordNet

The *Princeton WordNet* (PWN) lexical semantic network was developed by George Miller and his colleagues at the Cognitive Science Laboratory of Princeton University as a model of the mental lexicon (more specifically, the conceptual relationships of the English language) following the results of psycholinguistic experiments [35], [36]. The common noun *wordnet* denotes linguistic databases following the organization of the original WordNet developed at Princeton University.

In wordnet the senses of content words (nouns, verbs, adjectives, adverbs) are called word meanings. Synonymous meanings – words are interchangeable in a given context without changing (denotational) meaning – constitute *synsets* (synonym sets), the basic building blocks of wordnet's conceptual network. A concept in wordnet can be thus represented by sets of equivalent word meanings, eg. {board, plank}, {board, table}, {run, scat, escape}, {run, go, operate} etc.

There are several different types of ontological and linguistic *relationships* among the synsets that organize these nodes into an acyclic directed graphs, a conceptual network. Among noun concepts (synsets), the most important is the *hypernym* relationship (its inverse is called *hyponym*), which is an overloaded relation representing hierarchical (transitive, asymmetric, irreflexive) connections like is-a, specific/generic, inherits/generalizes, e.g. {house}-{building}, {bush}-{plant} etc. A special type of hyperonmy is the *instance* relationship holding between individual entities referred to by proper names and more general class concepts, e.g. {Romania}-{Balkan state}. Another

[2] http://wordnet.princeton.edu/wordnet/license/

DOI:10.15774/PPKE.ITK.2010.006

important hierarchical relationship between noun synsets is the *meronym* relation (inverse: *holonym*), which denotes part-whole relationships, and has three subtypes: *member* ({tree}-{forest}), *substance* ({paper}-{cellulose}), and *part* ({bicycle}-{handlebar}). *Domain* relations hold between a concept (domain term) and a conceptual class (domain), and have 3 types: *category* (semantic domain), e.g. {tennis racket}-{tennis}, *region* (geographical location of language users), e.g. {ballup, balls up}-{United Kingdom, Great Britain} and *usage* (language register), e.g. {freaky}-{slang}.

There are relations between noun concepts and synsets in other parts-of-speech: the *attribute* relation between a property (noun) and its possible values (adjectives), e.g. {color}-{red}; *derivationally related* forms, e.g. {reader}-{read}.

The *antonym* relation is defined for nouns, adjectives and verbs, and expresses opposition within a fixed denotational domain, e.g. {man}-{woman}, {die}-{be born}, {hot}-{cold} etc. For verbs the hypernym (inverse: *troponym*) relation expresses hierarchical types like for nouns, eg. {walk}-{travel, move}. Special relation for verbs are *entailment*, e.g. {snore}-{sleep} and *causes,* e.g. {burn (cause to burn or combust)}-{burn (undergo combustion)}. Instances of the domain relation also exist among verbs. For a certain class of adjectives, relational adjectives, the *antonym* and *similarity* relations form bipolar cluster structures, which consist of pairs of marked opposing adjectives and their synonyms. Adverbial synsets only connect to synsets in other parts-of-speech (derivational morphology.)

DOI:10.15774/PPKE.ITK.2010.006

Figure 2.1: A sample of Princeton WordNet illustrating the most important semantic relations

Princeton WordNet version 2.0 (the version used in my work) contains 146.000 different words in 115.400 synsets (79.700 noun, 13.500 verb, 18.500 adjective and 3.700 adverb synsets.)

Besides the many applications in word sense disambiguation, machine translation, information retrieval etc. [36], [56], [31], a number of criticisms have been expressed regarding the usage of WordNet as an ontology. [30] mentions too fine-grained sense distinctions, the lack of relationships between different parts of speech, simplicity of the relational information etc. WordNet also does not distinguish between types of polysemy and homonymy, and does not represent productive semantic phenomena such as metonymy. Some of these and other problems have been addressed by the OntoWordNet project [66].

2.1.2. EuroWordNet

The EuroWordNet (EWN) project (1996-1999, sponsored by the European Community), extended the Princeton WordNet formalism into a multilingual framework [41], [42]. EWN provided a modular architecture, where the synsets of the various participating languages (Dutch, Italian, Spanish, English, German, Czech, French and Estonian) were connected via a common connecting tier, the so-called Inter-Lingual Index (ILI).

DOI:10.15774/PPKE.ITK.2010.006

EuroWordNet's ILI is made of the English synsets of Princeton WordNet version 1.5, without the semantic relations. The so-called equivalence relations connect non-English synsets to the ILI records and provide connections among equivalent concepts among different languages. Besides exact equivalence there are a number of other equivalence relations (total 15) providing flexible ways of mapping concepts across languages.

In order to have roughly the same coverage of conceptual domains across languages, the various language concept hierarchies were constructed top-down from the so-called Common Base Concepts (CBC). The CBC set (1310 synsets) was selected together by the 8 participants from synsets in PWN 1.5 as being most important and fundamental concepts. The English CBC concepts were implemented in all languages, and were extended by Local Base Concepts (essential concepts specific to the local languages), and the local wordnets were developed by extending these with hyponyms, while connecting them to the ILI records. This meant that the different wordnets were based on a common core but could develop language-specific conceptualizations at the same time.

Even though the ILI is an unstructured list of PWN 1.5 synsets, a new, language-independent hierarchical structure, the so-called Top Ontology (TO) was created and imposed over it. The TO is a hierarchy of 63 Top Concepts (TC), which reflect essential distinctions in contemporary semantic and ontological theories. The TO connects to the CBC as a set of features (a CBC node can connect to several TC features), and the TC features can be inherited to the language-specific concepts via the CBC's ILI records.

DOI:10.15774/PPKE.ITK.2010.006

Figure 2.2: *Illustration of the EuroWordNet architecture with an equivalent concept in the Interl-Lingual Index, the Dutch and Spanish wordnets*

In the EuroWordNet project, the following two methodologies were defined for the construction of local wordnets:

a) **Merge Model:** the local base concepts and their semantic relations were derived from existing structured semantic resources available for the language, and were afterwards mapped to the ILI.

b) **Expand Model:** the local base concepts were selected from PWN 1.5 and were then translated to local language, equivalent synsets. In this approach, the language-internal semantic relations were inherited from Princeton WordNet and were then revised, using available monolingual resources if possible.

Following the Merge Model leads to a wordnet independent of Princeton WordNet, preserving language-specific characteristics. The Expand Model results in a wordnet strongly determined by Princeton WordNet. In EWN, the approach used was mainly determined by the available linguistic resources.

DOI:10.15774/PPKE.ITK.2010.006

2.1.3. BalkaNet

The aim of the BalkaNet (BN) project (2001-2004) was to extend EuroWordNet with 5 additional, South-Eastern European languages (Bulgarian, Greek, Romanian, Serbian and Turkish) [43].

In the final version of BalkaNet, Princeton WordNet 2.0 played the role of Inter-Lingual Index. Above the BN ILI (BILI), a new, language-independent hierarchy was defined using the SUMO upper-level ontology [46] and the mapping between SUMO and PWN [47].

The common core of BalkaNet (BalkaNet Concept Set, BCS) consists of 8.516 PWN 2.0 synsets, which includes the EWN CBC and additional concepts selected together by participants of the BN project.

All the resources used and generated in the project were converted to a common XML platform, which enabled the application of the VisDic tool [44], developed for the BN project, which supports the simultaneous browsing and editing of several linguistic databases. For quality assurance, a number of validation methodologies were introduced to ensure the syntactic and semantic consistency of the wordnets, and the validity of the connections between the languages [45].

2.1.4. Hungarian WordNet

Research on methodologies for the development of a wordnet for Hungarian started in 2001 at MorphoLogic [22], [21], [20], [19], [18], [17], [58]. The 3-year Hungarian WordNet (HuWN) project was launched in 2005 with the participation of 3 Hungarian academic and industrial institutions and funding from the European Union ECOP program (GVOP-AKF-2004-3.1.1.) (see also Section 5.2.) [12], [10], [8], [6], [2].

The Hungarian WordNet project followed mainly the footprints of the BalkaNet project, which meant taking the BalkaNet Concept Set as a starting point, using Princeton WordNet 2.0 as ILI, and the application of the VisDic editor and its XML format [12].

The development of the HuWN mainly followed the expand model (see Section 2.1.2.), except for the case of verbs, where a mixture of the expand and merge approaches were used [12]. Following the expand model meant that the selected BCS synsets were translated from English to Hungarian, and their semantic relations were imported. In order to ensure that the results would reflect the specialties of the Hungarian

DOI:10.15774/PPKE.ITK.2010.006

lexicon, the translated synsets and the imported relations were checked and if necessary, edited by hand using the VisDic editor.

Figure 2.3: *Illustration of the Expand Model for building a Hungarian WordNet: translating the English synsets and inheriting their semantic relations*

As I will show in Section 2.2., this method was sustainable in the case of the nominal, adjectival and adverbial parts of HuWN, while some adjustments to the language-specific needs were allowed as well. In the case of verbs, however, some major modifications were necessary. Due to the typological differences between English and Hungarian, some of the linguistic information that Hungarian verbs express through prefixes, related to aspect and aktionsart called for an additional different representation method [49], [50], [52]. Some innovations were introduced for the adjectival part as well [51], [52].

The design principle of following mainly the expand model was justified by the lack of structured semantic resources for Hungarian, the lower costs of development, and the availability of automatic synset translation heuristics, which I developed [17], [18], [19]. These will be discussed in more detail in the following. Following the expand model 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.)

DOI:10.15774/PPKE.ITK.2010.006

2.1.5. Automatic Methods for WordNet Construction

There are many examples of acquiring knowledge from machine-readable dictionaries (MRDs) – reference texts that were originally written for human readers, but are available in electronic format and can be processed by NLP algorithms to extract structured pieces of information [59]. Of these, several sources deal with the construction of taxonomies/ontologies across different languages.

In the framework of the ACQUILEX project, Ann Copestake and colleagues describe experiments [53], [54] where a limited set of Spanish and Dutch nominal lexical entries were successfully linked automatically to a taxonomy extracted from the Longman Contemporary Dictionary of English (LDOCE) MRD 103.

[30] gives an overview of some attempts to automatically produce multilingual ontologies. [60] link taxonomic structures derived from the Spanish monolingual MRD DGILE and LDOCE by means of a bilingual dictionary. [61] focus on the construction of SENSUS, a large knowledge base for supporting the Pangloss MT system, merging ontologies (ONTOS and UpperModel) and WordNet with monolingual and bilingual dictionaries. [62] describe a semi-automatic method for associating a Japanese lexicon to an ontology using a Japanese-English bilingual dictionary. [63] links Spanish word senses to WordNet synsets using also a bilingual dictionary. [64] exploit several bilingual dictionaries for linking Spanish and French words to WN senses.

For wordnet construction in a non-English language, the researchers at the TALP research group, Universitat Politecnica Catalonia, Barcelona have proposed several methods. They participated in the EuroWordNet project, and successfully applied their methods to boost the production of the Spanish and Catalan wordnets [30], [31], [64].

Their main strategy was to map Spanish words to Princeton WordNet (version 1.5) synsets, thus creating a taxonomy. This approach assumed a close conceptual similarity between Spanish and English. They relied on methods that used information extracted from several MRDs: bilingual Spanish-English and English-Spanish dictionaries, a monolingual Spanish explanatory dictionary (DGILE) and Princeton WordNet itself. The results of the different methods underwent manual evaluation (using a 10% random sample) and were assigned confidence scores. They describe several methods that can be grouped into 3 groups.

DOI:10.15774/PPKE.ITK.2010.006

The first group of methods („class methods”) are based on only structural information in the bilingual dictionaries. 6 methods are based on monosemous and polysemous English words with respect to WordNet, and 1-to-1, 1-to-many, and many-to-many translation relations in the bilingual dictionary. The so-called „field” method uses semantic field codes in the bilingual MRD. The „variant” method links Spanish words to synsets if the synset contains two or more English words that are the only translations of the Spanish word.

The second group („structural methods”) contains heuristics that rely on the structural properties of PWN itself. For each entry in the bilingual dictionary, all possible combinations of English translations are produced, and 4 heuristics decide on which synsets the Spanish words should be attached to: the „intersection”criterion works when all English words share at least one common synset in PWN. The „brother”, „parent” and „distant hypernym” criteria are applied when one of these relationships hold between synsets of English translations.

The third group of methods (“Conceptual Distance Methods”) rely on the conceptual distance formula, first presented by [65], which models conceptual similarity based on the length of the shortest connecting path of the two concepts in PWN's hierarchy. The formula is used for 1) co-occurring Spanish terms in the monolingual MRD's definitions, 2) headword and genus pairs extracted from the monolingual MRD, and 3) entries in the bilingual MRD having 1-to-many translations.

The authors first selected methods that produced confidence scores of at least 85%, yielding a total number of 10,982 connections between Spanish words and PWN senses. Then, relying on the assumption that individual methods that were discarded for lower confidence scores, when combined, could produce higher confidence, tested the intersection of each pair of discarded methods. By adding combinations whose confidence exceeded the threshold, they were able to add 7,244 further connections, a 41% increase, while keeping the estimated total connection accuracy over 86%.

In a more recent work, [56] describe a method for automatically generating a “target language wordnet” aligned with a “source language wordnet”, which is PWN. The authors demonstrate the method in the automatic construction of a wordnet for Romanian, and evaluate their results against the already available Romanian WordNet, which was manually constructed in the BalkaNet project. The method consists of 4 heuristics, relying on a bilingual and a monolingual dictionary. The first heuristic relies

DOI:10.15774/PPKE.ITK.2010.006

on the assumption that synonymous source language words will have common translations in the target language. The second heuristic relies on the assumption that a concept and its hypernym will have common information, and generates target language synsets from the intersection of the translations of words in hypernym-hyponym synsets in the source language. The third heuristic uses the results of the WordNet Domains project [57], where each of PWN's synsets had been labeled with one or more of 200 domain category labels. The authors label Romanian words with domain labels by 1) manual document categorization and automatic feature selection, 2) using field codes available in the bilingual MRD. Target language synsets are generated by using only translations of source language words that are compatible with the domain labeling. Finally, the fourth heuristic uses PWN glosses and the glosses in the Romanian explanatory dictionary, and generates synsets by measuring similarity between them, using the bilingual dictionary. The authors combine the individual heuristics using manually constructed meta-rules. The final set comprises 9,610 automatically generated Romanian synsets with 91% estimated accuracy.

2.2. Experiments

My goal was to create methods that would automatically propose Hungarian translations (Hungarian literals) for English PWN synsets. In practice, this meant that I was interested in developing methods that would map Hungarian words – the entries in the Hungarian side of a bilingual English-Hungarian MRD – to English synsets in Princeton WordNet.

This task can be grasped as a special case of *word sense disambiguation* (see Chapter 3), since it involved overcoming two levels of semantic 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 (Figure 2.4). With the available resources it meant that $1 \leq n \leq 19$, on average n was 1.71, while $1 \leq m \leq 9$, m was on average 2.16, yielding n*m = 3.69 on average. I used an ensemble of various heuristics that would rely on structural and semantic information found in the available bilingual and monolingual MRDs in order to get necessary semantic context – synonyms, hypernyms etc. – needed for the disambiguation process.

DOI:10.15774/PPKE.ITK.2010.006

Figure 2.4: *Levels of ambiguity in the Hungarian words–PWN synsets mapping process [22]. Solid lines represent translation links in the bilingual dictionary and synset membership in PWN, dotted lines mark incorrect, while dashed lines mark correct Hungarian word–PWN synset mappings from the possible choices.*

The choice of disambiguation methods follows the research of the Spanish EWN developers [31], [32], since the available resources were similar. I also developed and applied new methods that utilize the special properties of Hungarian and the available MRDs. The methods are presented below grouped by the type of resources they rely on.

In the following Section, I present the available resources that determined the applicable methods, which are presented in Sections 2.2.3.-2.2.4.. The methods were applied to the nominal part of the input set, and evaluated on a manually annotated random sample, described in Section 2.2.5. In Section 2.2.6., all the methods that were found reliable in the latter experiment, plus some new variants are applied to all parts of speech (nouns, verbs, adjectives) and are evaluated against the final, human-approved Hungarian WordNet database.

2.2.1. Resources

The English-Hungarian bilingual dictionary plays an important role in the process: on the one hand, it provides the translation links, and on the other, the set of Hungarian headwords serves as the domain of the disambiguation methods.

I compiled an in-house bilingual MRD from several available bilingual sources:

- MorphoLogic's Basic (“Alap”) English-Hungarian dictionary

DOI:10.15774/PPKE.ITK.2010.006

- MorphoLogic's Students' (“Iskolai”) English-Hungarian dictionary
- MorphoLogic's “Web Dictionary” (IT terms) English-Hungarian dictionary
- The *Gazdasági Szókincstár* (Vocabulary of Economy) English-Hungarian dictionary
- The Országh-Magay comprehensive English-Hungarian dictionary [34].

The dictionaries were available in XML format. I processed them to extract only the part-of-speech information besides the source and target language equivalents. Some of the dictionaries were English-Hungarian, while some were Hungarian-English, so I reversed each direction, creating sets of English-Hungarian translation pairs. These were simply unified into one set, which produced the merged bilingual dictionary. I removed all but the noun, verb and adjective entries, and also omitted translation pairs where the English entry was not available in PWN 2.0. The figures of the final bilingual dictionary are shown in Table 2.1.

TABLE 2.1.THE BILINGUAL MRD USED

	Hungarian words	**English words**	**Translation links**
Nouns	112,093	70,407	202,308
Verbs	33,695	12,769	79,831
Adjectives	37,377	23,743	82,952
Total	183,165	106,919	365,091

Two monolingual Hungarian MRDs were at my disposal: an explanatory dictionary and a thesaurus.

I converted an electronic version of the Hungarian explanatory dictionary *Magyar Értelmező Kéziszótár* (EKSz) [33] to XML format. Figures for the nominal part of the EKSz monolingual dictionary are presented in Table 2.2.

DOI:10.15774/PPKE.ITK.2010.006

TABLE 2.2. FIGURES FOR THE NOMINAL ENTRIES OF THE EKSZ MONOLINGUAL

Headwords	42,942
Definitions	64,146
Definitions annotated with usage codes	31,023
Headwords with translations in WordNet (through the bilingual)	10,507
Monosemous entries	30,062
Average polysemy count (polysemous entries only)	2.65
Average definition length (number of words)	5.22

In order to aid the construction of the Hungarian WN, I acquired information from the monolingual dictionary. The explanatory dictionary's definitions follow patterns which can be recognized to gain structured semantic information pertaining to the headwords [37]. I developed programs to parse each dictionary definition and extract semantic knowledge. The definitions were pre-processed by a simple tokenizer and the HuMor Hungarian morphological analyzer [38], and the programs used simple hand-written extraction rules based on morphological information and word order (the extraction algorithm is presented in details in Appendix A1.) In 83% of all the definitions, genus words were identified, which can be accounted for as *hypernym* approximations of the corresponding headwords, as in the following example:

koala: *Ausztráliában honos, fán élő, medvére emlékeztető erszényes* ***emlős****.*
(**Koala**: ***Mammal*** *resembling bears and living on trees native in Australia*.)

In 13% of the definitions, I was able to identify a synonym of the headword. Either the gloss consisted of synonym(s), or it was marked by punctuation:

forrásmunka: *Forrásmű*.
(“**Source work”:** *“Source creation”)*

lélekelemzés: *A tudat alatti lelki jelenségek vizsgálata;* ***pszichoanalízis***.
(“**soul analysis”:** *Examination of subconscious phenomena;* ***psychoanalysis)***

DOI:10.15774/PPKE.ITK.2010.006

In about 1,700 cases, the identified genus word was either a group noun, or a word denoting "part" relationship. For example, consider the EKSz entries for *alphabet* and *face:*

Ábécé: *A valamely nyelv helyesírásában használt* ***betűk*** *meghatározott sorrendű* ***összessége****.*
(Alphabet: *The ordered* ***set*** *of* ***letters*** *used in the spelling of a language.)*

Arc: Fejünknek *az a* ***része****, amelyen a szem, az orr és a száj van.*
(Face: *The* ***part*** *of the* ***head*** *that holds the eyes, nose and the mouth.)*

Using morphosyntactic and structural information, the meronym or holonym word (in our example: *letter, head*) could be identified instead of a genus word. This method provided *holonym/meronym* word approximations for 2.7% of all the headwords (only distinguishing between "part" and "member" subtypes of holonymy, as opposed to the 3 types represented in PWN). Summary of the processing of the definitions can be seen in Table 2.3.

These simple methods provided me with hypernym, holonym and synonym words for 99.2% of all the senses of 98.9% of all the nominal dictionary entries. Such information extracted from machine-readable dictionaries can be used to build hierarchical lexical knowledge bases [54], or semantic taxonomies [32]. The extracted genus word approximations also provide a valuable resource for the construction of the nominal part of Hungarian WN.

TABLE 2.3. THE RESULTS OF PARSING THE EKSZ NOMINAL DEFINITIONS

Definitions processed	**64,146**	**100.00%**
Processing failed	470	0.73%
Genus (hypernym) identified	53,526	83.44%
Synonym identified	10,589	16.51%
Holonym identified	826	1.29%
Meronym identified	584	0.91%

DOI:10.15774/PPKE.ITK.2010.006

A Hungarian electronic thesaurus was also available. The *MorphoLogic Thesaurus* contains 12,981 entries, which are groups of synonyms and other semantically related terms, containing 10,826 different Hungarian nouns, verbs, adjectives and other parts of speech.

2.2.2. Methods Relying on the Monolingual Dictionary

I used the semantic information acquired from the nominal EKSz definitions (see Section 2.2.1.) for the attachment of Hungarian nouns to PWN synsets in the following way:

- SYNONYMS: the PWN synset is chosen from the ones available for all the translations of the headword which contains the greatest number of the synonyms' English translations. This method was inspired by the INTERSECTION method (see Section 2.2.3.), but uses synonyms instead of equivalent translations.
- HYPERNYMS: for those cases where both the headword and the corresponding acquired hypernym (genus) have English translations, the headword is disambiguated against WordNet using a modified version of the conceptual distance formula [65], shown in Figure 2.5.

$$dist'(w_1, w_2) = \min_{\substack{c_{1i} \in w1 \\ c_{2j} \in w2 \\ depth(c_{1i}) < depth(c_{2j})}} |path(c_{1i}, c_{2j})|$$

Figure 2.5: *The simplified conceptual distance formula is applied to the pairs of English translations of a Hungarian noun and its hypernym. The formula returns two concepts (WN synsets) representing the input words which are closest to each other in the WN hypernym hierarchy*

A **third heuristic** which I developed depends on the LATIN equivalents available for about 1,500 EKSz headwords, mostly covering animal or plant species, taxonomic groups, diseases etc. Since these Latin terms are unambiguous, and PWN also contains most of them in different synsets, they can be used to attach the EKSz headwords in a straightforward way.

2.2.3. Methods Relying on the Bilingual Dictionary

These heuristics rely on information found in the connections between Hungarian and English words in the bilingual dictionary, and between English headwords and

DOI:10.15774/PPKE.ITK.2010.006

corresponding synsets in PWN. The first three heuristics were developed by [31], [32] for the Spanish WordNet project:

- MONOSEMOUS METHOD: if an English headword is monosemous with respect to WN (belongs to only one synset), then the corresponding Hungarian translations are linked to the synset.
- VARIANT METHOD: if a WN synset contains two or more English words that each has only one translation to the same Hungarian word, it is linked to this synset.
- INTERSECTION METHOD: links a Hungarian headword to all synsets sharing at least two of its English translations.

I developed a **fourth heuristic**, which depends on morpho-semantic information found in the Hungarian side of the bilingual dictionary. A number of Hungarian headwords in the bilingual dictionary are endocentric (noun + noun) compounds, which have the property that the second segment of the compound defines the semantic domain of the whole word [67]. For example, the compound *hangversenyzongora* ('grand piano') can be analysed as *hangverseny+zongora* ('concert'+'piano'), where the second segment, *zongora* serves as the **DERIVATIONAL HYPERNYM** noun of the compound. This piece of semantic information can be used with the modified conceptual distance formula (Figure 2.5) in order to select a target synset from the candidates (Figure 2.6). I used the HuMor morphological analyzer [38] to identify morpheme boundaries inside the headwords of the Hungarian side of the bilingual dictionary in order to find the noun+noun compounds for this method.

Figure 2.6: *Application of the conceptual distance formula on a N+N compound and its derived hypernym*

Performance of all the methods introduced so far, applied to the nominal part of PWN is shown in Table 2.4.

DOI:10.15774/PPKE.ITK.2010.006

TABLE 2.4. PERFORMANCE OF EACH METHOD ON NOUNS: NUMBER OF HUNGARIAN NOUNS AND WN SYNSETS COVERED, AND NUMBER HUNGARIAN NOUN-WN SYNSET CONNECTIONS

Method	Hungarian nouns	WN synsets	Connections
Monosemous	8,387	5,369	9,917
Intersection	2,258	2,335	3,590
Variant	164	180	180
DerivHyp + CD	1,869	1,857	2,119
EKSz synonyms	927	707	995
EKSz hypernyms + CD	5,432	6,294	9,724
EKSz Latin equivalents	1,697	838	848

As Table 2.4 shows, the most productive methods were the *Monosemous* method and the *Conceptual Distance* formula with EKSz hypernyms. While both methods produced about the same amount of connections, the latter generated more polysemy, with 1.79 connections for Hungarian words on average, compared to 1.18 connections on average by the Monosemous methods. The *Intersection* method, which relies on the bilingual dictionary follows the latter two in terms of produced connections. It is followed by the *Conceptual Distance* formula applied to *derivational hypernyms*, which found its place in the middle field in the ranking based on productivity. The remaining EKSz-based methods (*synonyms*, *Latin equivalents*) produced about the same amount of connections, but the former used less Hungarian entries. The least productive heuristic proved to be the *Variant* method.

2.2.4. Methods for Increasing Coverage

About 7% of the hypernyms or synonyms identified in the EKSz definitions had no English translation equivalents in the bilingual dictionary. To overcome this bottleneck, I used two additional methods to gain a related hypernym word that has a translation and can thus be used for disambiguation with the modified conceptual distance formula.

The **first method** was to look for derivational hypernyms of the (endocentric compound) synonyms or hypernyms, using the method described above. Since hyperonymy is a transitive semantic relation, the hypernym of the headword's hypernym (or synonym) will also be a hypernym.

DOI:10.15774/PPKE.ITK.2010.006

The **second method** looks up the hypernym (or synonym) word as an EKSz entry, and if it corresponds to only one definition (eliminating the need for sense disambiguation), then the hypernym word identified there is used, if it is available (and has English equivalents).

These two methods provided a 9.2% increase in the coverage of the monolingual methods. Table 2.5 summarizes the results of all the automatic methods used on different sources in the automatic attachment procedure (for nouns only.)

TABLE 2.5. TOTAL FIGURES FOR THE DIFFERENT TYPES OF METHODS

Type of Methods	Hungarian nouns	WN synsets	Connections
Bilingual	10,003	7,611	13,554
Monolingual	7,643	7,380	10,901
Increasing coverage 1-2	700	819	1,284
Total	13,948	12,085	22,169

2.2.5. Validation and Combination of the Methods

In order to validate the performance of the automatic methods, I constructed an evaluation set consisting of 400 randomly selected Hungarian nouns from the bilingual dictionary, corresponding to 2,201 possible PWN synsets through all their possible English translations. Two annotators manually disambiguated these 400 words, which meant answering 2 201 yes-no questions asking whether a Hungarian word should be linked to a PWN synset or not. Inter-annotator agreement was 84.73%. In the cases where the two annotators disagreed, a third annotator made the final verdict.

I evaluated the different individual methods against this evaluation set. I measured precision as the ratio of correct connections generated by the method to all connections proposed by the method, and recall as the ratio of generated correct connections to all possible human-approved connections. The results are shown in Table 2.6.

DOI:10.15774/PPKE.ITK.2010.006

TABLE 2.6. PRECISION, RECALL AND BALANCED F-MEASURE ON THE EVALUATION SET FOR THE INDIVIDUAL ATTACHMENT METHODS, IN DESCENDING ORDER OF PRECISION. THE LATIN METHOD IS NOT INCLUDED, BECAUSE FOR THE MOST PART IT COVERS TERMINOLOGY NOT COVERED BY THE GENERAL VOCABULARY OF THE EVALUATION SET.

Method	Precision	Recall	F-measure
Variant	92.01%	50.00%	64.79%
Synonym	80.00%	39.44%	52.83%
DerivHyp	70.31%	69.09%	69.69%
Increasing Coverage 1.	67.65%	46.94%	55.42%
Monosemous	65.15%	55.49%	59.93%
Intersection	58.56%	35.33%	44.07%
Increasing Coverage 2.	58.06%	28.57%	38.30%
Hypernym + CD	48.55%	41.71%	44.87%

In comparison to the results of the Spanish WordNet, [30] reports 61-85% precision (using manual evaluation of a 10% sample) on the methods described in Table 2.6 (excluding my own DerivHyp and Increasing Coverage 1-2 methods.)

[30] describes a method of manually checking the intersections of results obtained from different sources. They determined a threshold (85%) that served as an indication of which results to include in their preliminary WN. Then drawing upon the intuition that information discarded in the previous step might be valuable if it was confirmed by several sources, they checked the intersections of all pairs of the discarded result sets. This way, they were able to further increase the coverage of their WN without decreasing the previously established confidence of the entire set.

I used a similar approach. I decided to set the threshold for the individual methods to 70%, leaving only the Variant, Synonym and Derivational Hypernym methods. I then evaluated all the possible combinations of the eliminated further 5 methods. Table 2.7 lists the combinations that exceeded the 70% threshold.

DOI:10.15774/PPKE.ITK.2010.006

TABLE 2.7. PRECISION AND RECALL OF INTERSECTIONS OF SETS NOT INCLUDED IN THE BASE SETS, EXCEEDING 70% PRECISION

Combinations of methods	Precision	Recall
Inc. cov. 2. & Hypernym	95.78%	50.00%
Inc. cov. 2. & Intersection	88.14%	90.00%
Inc. cov. 2. & Mono	87.50%	70.00%
Hypernym & Mono	71.91%	52.46%

On the nominal WordNet set, the 2,722 Hungarian word—PWN connections generated by the individual ≥70% methods could be extended by 8,579 connections provided by the combination methods, producing 9,635 unique connections. The evaluation of these connections against the evaluation set showed 75% accuracy [17].

2.2.6. Application and Evaluation in the Hungarian WordNet Project

In the Hungarian WordNet project (Section 2.1.4.), I applied all the methods and method combinations selected in the validation experiments (Section 2.2.5.) for noun, verb and adjective entries in the bilingual dictionary using all respective candidate synsets in Princeton WordNet 2.0. In addition, I also applied some additional variations of the above methods:

- **Synonyms** method using the **MorphoLogic Thesaurus**: I applied the Synonym method to the synonym groups extracted from the MorphoLogic Thesaurus (see 2.2.1.)
- **Derivational hypernyms** of **multiword** expressions: 76,385 Hungarian entries of the bilingual MRD were multiwords, i.e. the lexemes contained two or more space- or hyphen-separated tokens. Using the HuMor analyzer, I identified 34,155 of these where the last segment (assumed to be the head) was a noun. Like in the DerivHyp method, I took the last token as the derivational hypernym and applied the Conceptual Distance formula.
- **Polysemous** English entries with **unambiguous translation** links: following [30], in addition to monosemous English words (having only one sense in PWN) I also used polysemous words (more than 1 senses in PWN) and their Hungarian translations. However, I only attached Hungarian translations to these synsets if

DOI:10.15774/PPKE.ITK.2010.006

the translation relation between the English word and its Hungarian equivalent was unambiguous (1-to-1), assuming these cases to be most reliable.

After the completion of the Hungarian WordNet project, where 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 was interested in the precision and recall of automatic synset translation (Hungarian words to PWN synsets mapping) in the perspective of this final human-edited data set, containing 42,000 synsets .

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 (not eliminated) by the human 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 present (considering only the synsets the heuristics attempted to translate.)

These measures were calculated for all affected parts of speech in HuWN (nouns, verbs, adjectives). A summary of the results, in addition to other statistics of the automatic synset translation can be seen in Table 2.8.

TABLE 2.8. EVALUATION RESULTS OF AUTOMATIC SYNSET TRANSLATION AGAINST HUNGARIAN WORDNET

	All	**Nouns**	**Verbs**	**Adjectives**
Precision	24.61%	31.53%	13.89%	17.36%
Recall	64.81%	63.77%	64.46%	71.96%
% of synsets attempted (synsets with proposed links)	51.96%	53.30%	57.27%	40.41%
% of synsets with proposed and at least 1 approved links	39.22%	39.44%	40.99%	36.69%

Table 2.8 reveals a two-sided picture. On the one hand, for each part of speech, the precision of the automatically generated translation links was low (24.61% overall). However, on the other hand, recall was over 60% for all parts of speech (exceeding 70% in the case of adjectives.) This suggests that the translation heuristics had an obvious tendency to overgenerate: they proposed more Hungarian translations for each synset than it was approved by the human lexicographers. However, after deleting the superfluous synonyms, the ones remaining had high accuracy. This means that the

DOI:10.15774/PPKE.ITK.2010.006

automatic methods did actually succeed in supporting the process, since the lexicographers had to resort more to deleting than to adding new synonyms (which is a more time-consuming procedure).

51.95% of all synsets in the final Hungarian WordNet ontology were attempted by the automatic translation heuristics. This figure is highest for verbs (57.27%) and lowest for adjectives (40.41%). A significant amount (39.22%) of synsets in the final product contains at least one synonym that was automatically proposed.

In this round of validation, I also performed the individual evaluation of the 3 additional heuristics described in this section. The results are shown in Table 2.9.

TABLE 2.9. INDIVIDUAL EVALUATION OF THE 3 NEW METHODS DESCRIBED IN THIS SECTION, TOGETHER WITH THE DETAILED EVALUATION OF THE MONOSEMOUS METHOD ON DIFFERENT PARTITIONS OF THE BILINGUAL DICTIONARY

	Nouns		**Verbs**		**Adjectives**	
	Precision	Recall	Precision	Recall	Precision	Recall
ML-Thesaurus Synonyms	28.02%	15.74%	27.00%	47.68%	13.5%	29.3%
Multiwords DerivHyp	18.61%	2.89%	n.a.	n.a.	n.a.	n.a.
Polysemous 1-1	44.98%	1.23%	3.45%	0.15%	42.5%	0.6%

The method that used synonyms from MorphoLogic's Thesaurus showed a precision of 28.02% and recall of 15.74% (F1-score 20.16%) on nouns. This is in high contrast with the performance of this method when it was used on synonyms extracted from EKSz definitions and was evaluated on the manually disambiguated random sample (Table 2.6). In the latter case, precision reached 80% and recall was 39% (F1-score 52.83%). This significant difference implies – apart from the divergence between the two evaluation methodologies – important information on the quality of the two resources when used for the construction of HuWN. The synonyms extracted from the explanatory dictionary's definitions seem to be more valuable for this purpose than the terms obtained from the thesaurus. This may be explained by the fact that entries in the thesaurus usually employ a more slack notion of synonymy, covering a far broader range of relations than the more strict, denotational application of synonymy in the monolingual dictionary.

Performing a similar comparison between using derivational hypernyms obtained from multi-word lexemes (Table 2.9) and morphological analysis of single-word compounds (Table 2.6) with the *Conceptual Distance* formula reveals that the latter (69.69% F1-

DOI:10.15774/PPKE.ITK.2010.006

score on the 200-noun evaluation sample) outperforms the first (5% F1-score for nouns in the final HuWN). Since it starts off from more ambiguous information, it is not a surprise that the *Polysemous* method (precision 44.98%, recall 1.23%, Table 2.9) ranks lower when compared to the *Monosemous* method (precision 65.15%, recall 55.49%, Table 2.6).

DOI:10.15774/PPKE.ITK.2010.006

2.3. Summary

In this Chapter, I presented my experiments with the automatic generation of a Hungarian WordNet ontology. I applied the expand model of building wordnets, using heuristics for the automatic mapping of Hungarian lexical items to English WordNet synsets. I applied 4 heuristics (MONOSEMOUS, POLYSEMOUS, VARIANT, INTERSECTION) that use only information in a bilingual dictionary, and 2 methods that also use semantic information acquired from the glosses in a monolingual dictionary or found in a thesaurus (CONCEPTUAL DISTANCE on headword and hypernym, or using SYNONYM groups.) I developed some heuristics that are based either on the special characteristics of Hungarian: DERIVING HYPERNYMS of endocentric noun COMPOUNDS and MULTI-WORDS, or on the special properties of the monolingual dictionary: available LATIN headword equivalents. I also proposed two methods to extend the coverage of the automatic synset translation by utilizing the transitive nature of the hypernym relation (both derivational and acquired.)

I performed evaluation of all the methods on a manually disambiguated random sample of 400 Hungarian nouns (2,201 possible connections to PWN). I also evaluated the performance of methods that were selected using a threshold in the first evaluation round in the framework of the final Hungarian WordNet product, where the automatic methods were applied and the results were manually revised.

Related theses (see Section 5.1. for more details):

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

I.2. I proposed 4 new heuristics for the automatic construction of Hungarian synsets in the expand model (*using synonym groups, using derivational hypernyms of endocentric compounds and multiwords, using Latin headword equivalents, extending coverage by using derivational or acquired hypernyms of untranslatable hypernyms/synonyms***). The methods disambiguate Hungarian nouns against English synsets, and rely on the special properties of the Hungarian language and the available resources.**

DOI:10.15774/PPKE.ITK.2010.006

Related publications: [2], [3], [6], [8], [9], [10], [11], [12], [17], [18], [19], [20], [21] [22], [23]

DOI:10.15774/PPKE.ITK.2010.006

Chapter 3

WORD SENSE DISAMBIGUATION IN MACHINE TRANSLATION

3.1. Introduction

In natural language processing, the task of **word sense disambiguation** (WSD) is to determine which of the senses of a lexically ambiguous word is invoked in a particular use of the word by looking at the context of the word. The disambiguation of cross-part-of-speech ambiguities (e.g. verb or noun senses of *house*) does not fall in the domain of WSD, as these can be effectively tackled using n-gram models (part-of-speech tagging), WSD should deal only with ambiguities *within* a certain part of speech.

The above definition of WSD raises some problems: when do we call a word polysemous, and how do we define its possible senses?

Lexical ambiguity covers a whole range of phenomena and is an actively researched field in theoretical and computational linguistics [67], [68], [69]. An interesting question is the distinction between **homonymy** and **polysemy.** It is easy to see that finding a way to distinguish between semantically unrelated homonym senses will be more easy for a computer system than discriminating between the vaguely distinguishable senses of a polysemous lexical item (see below).

Homonymy can be defined as a phenomenon when there is no common element of meaning between two words that are represented by identical phonetic/orthographic signs in a language. Examples in Hungarian include words like *kar* ("choir, group of people*"* and "upper limb*"*), where the the common sound form is the result of linguistic changes. In the case of polysemous words, however, there is a common aspect of meaning, as in the example of the Hungarian nound *gép*: "structure to convert energy or to carry out a task" and "airplane." While in the latter case, one can speak about separate senses, in the case of the words like *teacher* (in English and in Hungarian), although it can mean both a male or a female person, one can assume one underspecified meaning [67], while in German, these would be expressed with separate lexical units: *Lehrer, Lehrerin.* [67] proposes to represent the three phenomena along a continuum, where the two extremes

DOI:10.15774/PPKE.ITK.2010.006

would be homonymy and semantic underspecification, while types of polysemy could be placed in between.

Polysemy can show patterns: certain word classes can behave ambiguously in a similar, predictable way. For nouns, one of the most common such phenomenon is metonymy. For example, in the sentence “I finished the book/song/house” the basic meaning of the object noun changes in a similar way. [70] analyzes similar productive lexical operations and shows the shortcomings of representing them with simple sense enumerations.

In the remaining part of the dissertation, if not otherwise stated, for the sake of simplicity, I will generally use the term “ambiguity” to denote all of the above-mentioned lexical phenomena.

In practical WSD implementations, the set of possible word senses are defined by the available items in a specific available lexical resource, such as a machine-readable dictionary or a lexical database like **WordNet**. This, however, presents problems. These dictionaries were originally written for comprehension by human readers, and the sense distinctions and definitions are often based on the subjective decisions of the lexicographers. (With the exception of novel corpus-based dictionaries such as the Macmillan English Dictionary [73], which was compiled by analyzing corpus data.) Extensive lexicons such as WordNet can make very fine-grained, arguable sense distinctions, which will make automatic sense discrimination difficult. [59] shows that even human annotators find such tasks problematic: agreement between two annotators could be as low as 65-70% when distinguishing between instances of words with many closely related senses. [59] proposes that inter-annotator agreement in a given WSD annotation task could be interpreted as a possible upper bound to the machine's performance, since we cannot expect the algorithm to perform better when human experts disagree on a sense assignment.

A different, more natural possibility is to characterize the different senses of ambiguous words based on the different translations the word may have in a second language [59]. This approach is especially convenient in machine translation environments, and is the approach adopted by the WSD system I will present in this Chapter.

DOI:10.15774/PPKE.ITK.2010.006

3.1.1. Polysemy in English

I conducted an experiment in order to examine the presence of polysemy in representative real-life English texts. I used the First Release of the American National Corpus (ANC) [71], which contains 10 million corpus tokens, all of which are part-of-speech tagged and the base forms are also annotated. This latter feature was important and ruled out the application of other available, larger, but not lemmatized corpora (for example, the 100-million word British National Corpus.)

I was interested in the distribution of polysemous content words in ANC, with polysemy information based on Princeton WordNet 2.0. I classified each noun, verb, adjective and adverb token occurrence in the corpus according to how many senses its base form had in PWN. I also performed this analysis for the token types (base form classes). The distribution of sense counts is shown in Figure 3.7 (the graph data is available in Appendix A2), which is in accordance with Zipf's Law [59]. Table 3.10 summarizes the number of monosemous (1 sense in PWN) and polysemous (more than 1 sense in PWN) content word types and tokens in the ANC corpus.

TABLE 3.10. NUMBER OF MONOSEMOUS (1 SENSE IN WORDNET) AND POLYSEMOUS (MORE THAN 1 SENSE IN WORDNET) CONTENT WORD TYPES AND TOKENS IN THE THE ANC

	Types	**Tokens**
Monosemous	19,550	655,807
Polysemous	16,689	4,255,234
Total	36,239	4,911,041

DOI:10.15774/PPKE.ITK.2010.006

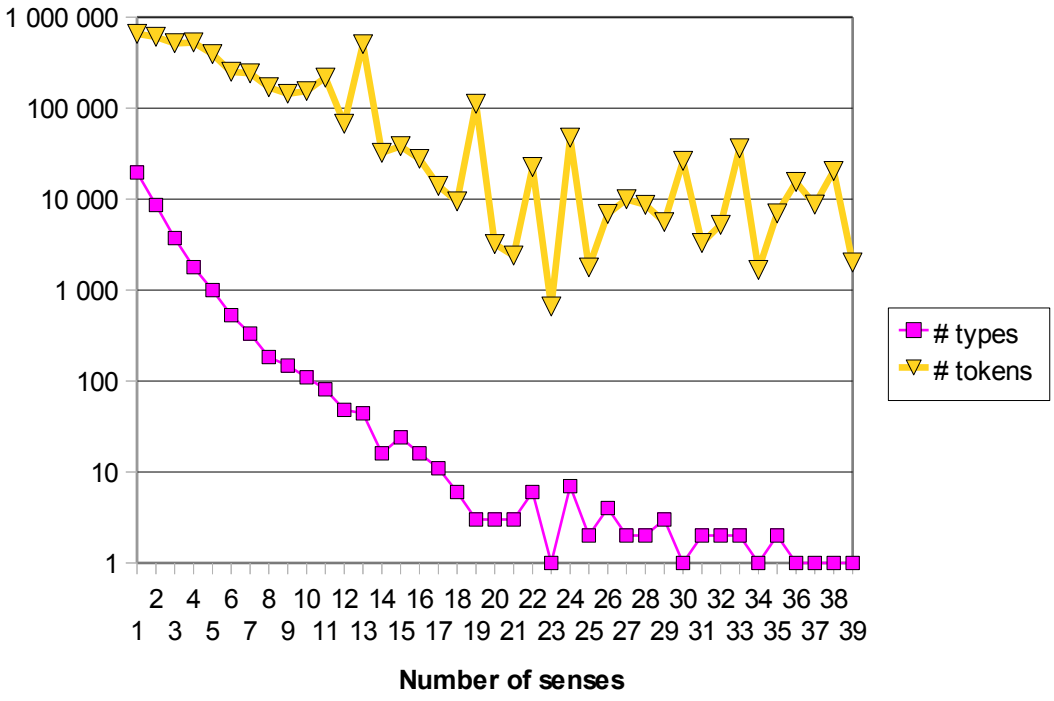

Figure 3.7: The distribution of polysemy in the 10-million word American National Corpus. The two graphs show the number of types/tokens as a function of the number of Princeton WordNet senses (X-axis).

Table 3.10 shows that polysemy with respect to WordNet is a fundamental phenomenon in the 10-million word American National Corpus. When looking at word types (the "lexicon"), 46% of all content words could have more than 1 meaning, and when looking at word tokens (the actual occurrence of the lexical items), 86% of the (closed-class) corpus tokens could be potentially interpreted with multiple meanings.

3.1.2. Approaches in WSD

There are many approaches to word sense disambiguation, applying methods ranging from hand-developed disambiguation rules, methods utilizing lexical resources like explanatory dictionaries and thesauri to methods using machine learning [59], [72].

One of the earliest, most well-known methods using an MRD – the definitions in an explanatory dictionary – is the Lesk Algorithm [72]. The algorithm uses the bag of words from the glosses of the words in the context of the ambiguous word, and selects the sense whose definition has the highest overlap, drawing on the intuition that surrounding words will be semantically related. Lesk reports 50-70% disambiguation accuracy.

DOI:10.15774/PPKE.ITK.2010.006

[74] uses the semantic categories available in *Roget's Thesaurus* for disambiguation. Ideally, the different senses of polysemous words can be labeled with different categories. Texts were collected to characterize the categories by selecting sentences from a 100-million word corpus that contained one or more of the terms in the categories. Relevant words for the categories were then extracted from the text collections using the mutual information-like formula, where *w* is a word and *Rcat* is a Roget's category:

(3.1) $$\frac{P(w \mid RCat)}{P(w)}$$

The formula was used to assign *salience weights* to each extracted word, indicating association strengths for each category.

Disambiguation was performed by examining a 100-word context window around the ambiguous word (50 words before, 50 words after.) For all Roget's categories, the salience weight for all of the the categories were computed from the context words, and the category with the highest score was selected, determining the selected sense of the ambiguous word. This method was applied with a fairly high precision (91-99%) for certain words. However, [74] points out that *topic-independent* words will show lower performance: for example, while senses of the word *crane* correlate well with the ANIMAL and MACHINERY categories, the meanings of *interest* could appear in any semantic context.

[76] describes an approach relying on the information in WordNet's semantic network. The authors define the so-called *conceptual density* formula to characterize connections between word senses as the ratio of the size of a set of word senses to the size of the hyponym subtree in WN containing these senses. During disambiguation, the algorithm chooses the sense that is contained in the most dense WordNet subtree together with the senses of the context words. Combining this method with several other (unsupervised) methods, [76] reports on average 68% precision in the SensEval-1 competition [77].

[78] uses supervised machine learning for word sense disambiguation. The authors use the simple but well-known Naive Bayes algorithm, which is widely used in machine learning due to its efficiency and its ability to combine evidence from a large number of features [59]. The Naive Bayes method chooses the sense s_i that is most likely given the observed values for contextual features $c_1,\ldots,c_n$, or formally, the sense with the maximum value for conditional probability $P(s_i \mid c_1,\ldots,c_n)$. These can be calculated using *Bayes' Theorem*:

DOI:10.15774/PPKE.ITK.2010.006

(3.2) $$P(s_i \mid c_1,\ldots,c_n) = \frac{P(c_1,\ldots,c_n \mid s_i)P(s_i)}{P(c_1,\ldots,c_n)}$$

$P(s_i)$ is the *a priori* probability of sense s_i, which can be given by its relative frequency in the training corpus. However, there is usually not enough training data to calculate $P(c_1, \ldots,c_n \mid s_i)$. To overcome this, the method uses the assumption that got its name: the contextual features are assumed to be conditionally independent. While this assumption is obviously inappropriate for most linguistic data, there is still a surprisingly large number of cases the algorithm does well [59]. The independence assumption also makes it possible to estimate $P(c_1,\ldots,c_n \mid s_i)$.with the product of the conditional probabilities of the individual features:

(3.3) $$P(c_1,\ldots,c_n \mid s_i) \approx \prod_{j=1}^{n} P(c_j \mid s_i)$$

The value of $P(c_j \mid s_i)$ can be computed with the relative frequencies observed in the training corpus (F denotes frequency), perhaps applying smoothing [59]:

(3.4) $$P(c_j \mid s_i) = \frac{F(c_j,s_i)}{F(s_i)}$$

[78] used two groups of features: a) Global (topical features characteristic of the general semantic domain): content words in the 3 neighboring sentences, and b) Local (features describing syntactic dependencies): content words in a 3+3 window, function words and part-of-speech tags in a 2+2 window around the ambiguous word. The authors use the average of the Good-Turing formula and the values observed in the corpus for smoothing. The system was tested using a manually disambiguated training corpus for the noun *line*, the verb *serve* and the adjective *hard*, obtaining an overall precision rate of 83%. Using only global features decreased precision to 78%, while local features alone produced 67% precision. Precision increased with the number of the training instances, best performance was reached by using enough training material that provided 200 training instances even for the least frequent sense.

[75] uses a supervised decision list algorithm. The central idea was to select the strongest available contextual features and make decisions based on them. For this, the feature-value pairs observed in the corpus were ranked according to the log-likelihood ratio:

DOI:10.15774/PPKE.ITK.2010.006

(3.5)
$$abs(\log\left(\frac{P(s_{i1} \mid c_j)}{P(s_{i2} \mid c_j)}\right))$$

$P(s_{ik}|c_j)$ is the probability of sense s_{ik} given feature-value pair c_j. Disambiguation is done by choosing the first/strongest feature-value pair observed in the context and selecting the appropriate sense. If none of the features in the list could be used, the most frequent sense is selected. A claimed advantage of this method as opposed to the Naive Bayes algorithm is that there is no need for the assumption of feature independence by combining the features. [75] uses word forms and collocations as features, and with a few hundred training examples reports over 90% disambiguation accuracy for homonym pairs.

[79] uses memory-based learning for WSD. This "lazy" supervised machine learning method does not generalize over the training instances but simply stores them all. [79] claims that the advantage of this approach is that it accounts for all the original training instances and there is no risk of ignoring exceptions, which are typical in natural language systems. Disambiguation consists of extracting the values of the contextual features and comparing them to all the memorized training instances. The algorithm selects the sense that is most prominent among the *k* most similar memorized instances (*k-nearest neighbor algorithm*.)

The features are represented in feature vectors (which can be compared using vector-space similarity measures), and are weighted by their *information gain* scores. In addition, [79] also employs a feature selection algorithm, justified by the claim that in certain cases it is more effective to disregard certain features altogether instead of giving them lower weights. The feature selection automatically generates the optimal set of features for each ambiguous word. Starting with an empty feature set, the algorithm adds the single best feature from the pool of all 18 possible features that provides best precision with 10-fold cross validation. This is repeated until the disambiguation precision no longer improves or there are no more features to add. The original feature pool contains all features known to have worked in WSD: words, n-grams, PoS-tags, parse information etc. [79] achieved the highest score in the SensEval-2 competition: in the lexical sample task for nouns with fine-grained sense distinctions it produced 69.5% precision in average, with coarse-grained sense distinctions 76.6%. On the three parts of speech (nouns, verbs, adjectives) it performed 63.8% and 71.2% on average.

DOI:10.15774/PPKE.ITK.2010.006

[91] describes a method for word sense disambiguation using a second language monolingual corpus and a bilingual dictionary. The idea is that different senses of a word in language *A* will have different translations in language *B*, and that the translations of the collocations formed with the different senses are also different. In order to disambiguate an instance of the ambiguous word in language *A,* the phrase it occurs in is identified and a corpus of language *B* is searched for the translation. If the phrase occurs with only one of the translations of the ambiguous word in language *B*, then the corresponding sense can be assigned to the original instance in language *A*.

I conducted experiments to explore the possibilities of applying a novel representation method called Random Indexing to WSD [87], [16]. Random Indexing (RI) uses high-dimensional sparse vectors with random patterns modeling neural activation in the brain to represent linguistic information. This representation is similar to the Hyperspace Analogue to Language (HAL) [89]and Latent Semantic Analysis (LSA) [90] but requires less computational resources [88]. I used sense-tagged corpora for the English nouns *line* and *party,* available form the OMWE project (see Section 3.2.1.) to generate an accumulated representation of the training examples. I experimented with various contextual features (content words in the entire context, content words, function words and PoS-tags in a window around the ambiguous word), window sizes, weight functions and absolute position markers. While the results exceeded the baseline scores, two problems prevented further investigations of RI in WSD: the lack of a way to combine information from the different features, and the instability of the representation. I showed that as much as 18% difference in precision can be observed when using the same set of parameters, just from the random factor in the representation method.

3.1.3. WSD in Machine Translation

As defined at the beginning of this Chapter, in word sense disambiguation (WSD), the machine has to select the correct sense of an ambiguous word in its context. In the remaining parts of this Chapter, I will present an application of WSD in machine translation (MT), where the system has to select the correct translation equivalent in the target language of an ambiguous item in the source language. For example, the polysemous English noun *party* would translate to two different Hungarian words (*párt* for the political organization sense, or *parti* for the social event sense) in the following two sentences:

DOI:10.15774/PPKE.ITK.2010.006

(3.6) a. The *party* that won the elections four years ago did not make it into Parliament this time.

b. The *party* yesterday celebrated her birthday at one of the finest restaurants in town.

While statistical machine translation architectures are able to handle such phenomena by design, in a rule-based machine translation system, making such distinctions is a challenge. The MetaMorpho English-Hungarian machine translation system [39] contains more than 100.000 manually created context-free analysis and generation rule pairs (translation patterns). Some source language disambiguation is performed within the grammar by selectional restrictions using simple semantic features:

(3.7) a. They *fired* the furniture[-ANIM]. *Eltüzelték* a bútort.

b. He *fired* the employee[+ANIM]. *Kirúgta* az alkalmazottat.

In the above examples, the nouns *furniture* and *employee* are encoded with the ANIM (animate yes/no) semantic feature in the lexicon. In the source language analysis phase, the actual value of the ANIM feature in the object NP position determines which of two verb patterns will be selected that encode different translations of the verb in the target language.

Such grammar-based disambiguation is however limited only to a small number of verb frames, while for the majority of verbs and none of the nouns are not disambiguated during translation in MetaMorpho. For these there is only a single sense (translation) encoded in the pattern database, one that the rule experts decided was the most frequent sense. This obviously presents problems:

(3.8) a. We moved to another *state*. Egy másik *államba* költöztünk.

b. Her *state* was satisfactory. Az **állama* kielégítő volt.

Idiomatic multiwords (collocations, non-compositional units) on the other hand have a special treatment in MetaMorpho, they have their own translation patterns, which override the patterns that fire for the elements constituting the phrase:

(3.9) The *state of affairs* is intensifying. A *helyzet* fokozódik.

However, for all the compositional cases, external help is needed from an outside "oracle" that can hint the proper sense by looking at the available semantic context and relies on knowledge acquired from real-life data. I will propose a solution for this problem in the following sections.

DOI:10.15774/PPKE.ITK.2010.006

3.2. Experiments

I was interested in developing a method to perform the automatic disambiguation of source language (English) nouns to their target language (Hungarian) translations, with possible implementation in the MetaMorpho rule-based MT system.

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. Supervised machine learning methods have shown huge success in WSD [59], and there are a number of openly available training corpora for English, which will be described in Section 3.2.1. I will discuss the selection of the learning features, the learning algorithm and experiments to optimize the feature representation in Section 3.2.2. I present evaluation of the system using the training corpus in Section 3.2.3. The details of implementation inside the MetaMorpho MT engine and an experiment to evaluate is presented in Section 3.2.4.

Section 3.2.5. deals with an experiment that explores the possibility of generating training instances automatically from parallel corpora.

3.2.1. Training Data

The WSD classifier described in the following sections uses manually sense-tagged training corpora in the source language (English), since no tagged training material was available for the target language (Hungarian). I used the following openly available corpora to obtain training instances for the classifiers for my experiments:

- Open Mind Word Expert (OMWE) data [80]
- SensEval-1 and 2 English Lexical Sample Task data [81]
- A sense-tagged corpus for the noun *line* [78]

The Open Mind Word Expert project used online volunteer work to sense-tag occurrences of ambiguous English words in texts form the *Penn TreeBank*, the *Los Angeles Times* and the *Open Mind Common Sense* project, which collected common sense assertions from volunteers. The original dataset consisted of annotations for 285 English nouns. However, looking at the sense frequencies revealed that for many items the sense distributions were obviously unrealistic, or the data was too sparse for some senses (e.g. the noun *brother* has 4 senses with 96, 3, 1 and 1 examples for each), so I did

DOI:10.15774/PPKE.ITK.2010.006

not use items where the number of training examples for more than 1 senses was below 5. This gave training data for 22 nouns (Table 3.11).

The SensEval corpora were used to compare systems in the English Lexical Sample task in the SenseEval 1 and 2 WSD competitions. I used both the training and test sets for the 29 available nouns. 6 nouns were also covered by the OMWE corpus. I performed some tests which lead me to use the OMWE for these, ending up with data for 23 nouns from this source (Table 3.11).

I also used the annotated corpus available for the noun *line,* containing texts form the *Wall Street Journal* and various works of fiction, developed for investigations by [78]. The OMWE corpus also contained this noun, but I decided to use this dataset instead because it had more instances and it provided a way for comparison with the results by [78].

I converted the set of all training instances first to a common XML format, then preprocessed them in the following steps: segmentation into paragraphs, sentences and words, morphological analysis [38], disambiguation (using MorphoLogic's transformation-based PoS-tagger), and obtaining word stems. I manually identified idiomatic multi-word lexemes formed by the ambiguous words among the sense tags, and compiled a list of these to be coded later as separate translation patterns in the MT system, since these usually have a single sense that can be translated without the aid of WSD (Section 3.1.3.)

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. During translation, when it was possible I intentionally tried to find target language equivalents that would be polysemous, subsuming as many of the original senses as possible. In machine translation, it is not a problem if lexical ambiguity is preserved in the translation process, only if an inappropriate translation is chosen:

(3.10) a. The jar had a wide mouth$_1$. Az üvegnek széles volt a *szája*.

b. He stuffed his mouth$_2$ with candy. Cukrot tömött a *szájába.*

c. New York is at the mouth$_3$ of the Hudson. New York a Hudson **szájánál* van.

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

DOI:10.15774/PPKE.ITK.2010.006

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 disambiguation for these, so 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 (see Table 3.11).

DOI:10.15774/PPKE.ITK.2010.006

TABLE 3.11. THE ORIGINAL LEXICAL ITEMS IN THE TRAINING DATA SET. THE ITEMS SET IN BOLD WERE NOT USED SINCE ALL THEIR ENGLISH SENSES COULD BE MAPPED TO A SINGLE HUNGARIAN TRANSLATION.

Noun	Source	English senses	Hungarian translations	Training instances
arm	OMWE	5	4	787
art	OMWE	4	2	108
authority	SEVAL	3	3	257
bank	OMWE	4	2	398
bar	SEVAL	7	4	337
bum	SEVAL	5	2	118
chair	SEVAL	8	3	191
chance	OMWE	6	4	615
channel	**SEVAL**	**5**	**1**	**132**
chapter	OMWE	3	2	137
child	SEVAL	7	2	180
church	SEVAL	3	2	183
circuit	OMWE	6	4	184
day	OMWE	2	2	192
degree	OMWE	4	2	485
detention	**SEVAL**	**2**	**1**	**72**
dyke	SEVAL	4	2	86
facility	SEVAL	3	2	37
fatigue	SEVAL	4	2	104
feeling	SEVAL	3	2	149
grip	OMWE	5	2	218
hearth	SEVAL	3	2	96
holiday	SEVAL	4	2	83
image	OMWE	7	2	512
lady	SEVAL	4	2	134
letter	OMWE	3	2	927
line	LINE	6	5	4,157
material	**OMWE**	**4**	**1**	**192**
mouth	SEVAL	2	2	169
nation	**SEVAL**	**3**	**1**	**113**
nature	**SEVAL**	**4**	**1**	**125**
operator	OMWE	2	2	119
party	OMWE	2	3	623
performance	OMWE	2	2	353
plane	OMWE	4	3	474
post	SEVAL	3	3	141
process	OMWE	2	2	302
report	OMWE	3	3	335
restraint	OMWE	6	4	89
sense	SEVAL	4	3	136
spade	SEVAL	5	3	89
stress	SEVAL	3	2	115
term	OMWE	5	3	125
unit	**OMWE**	**7**	**1**	**229**
yew	**SEVAL**	**2**	**1**	**81**

DOI:10.15774/PPKE.ITK.2010.006

3.2.2. Contextual Features and Learning Algorithm

To represent contextual information for training the classifiers, I used features identified from the context of the ambiguous words based on [78] and [79], that can be grouped into two categories. Features in the first type ("local") are 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 certain content words from a 3+3 window. The other group of features ("global") represents the semantic domain, or topic of the entire available context (usually the paragraph containing the ambiguous word). This information is represented by coding the presence of certain frequent content words in the global context.

I conducted a simple experiment to select the best **machine learning algorithm** for the problem. I tested several of the various supervised learning algorithms available in the Weka Data Mining Toolkit (Version 3.3) [82]. I tested 2 Bayesian (AODE, Naive Bayes) and 3 lazy algorithms (IB1 (nearest neighbor), IBk (k-nearest neighbor, with k=2,3,4) and K-star.) (For a detailed description of the algorithms please see [82], [83] and [84].) For this experiment, I used only closed-class words in a 2+2 window around the ambiguous word, because I didn't want to bias the learning algorithms with unoptimized parameters for the more complex features (see below). I used 2 well-represented nouns from the training corpus (applying English sense tags): *party* (3 senses, 623 instances) and *line* (6 senses, 4157 instances.) I evaluated classification precision both on the training set itself *and* using 10-fold stratified cross-validation on the training data (Table 3.12). The results led me to choose the Naive Bayes algorithm (which was also the choice for its simplicity and its previously reported good performance in WSD, see Section 3.1.2.)

DOI:10.15774/PPKE.ITK.2010.006

TABLE 3.12. EVALUATION OF MACHINE LEARNING ALGORITHMS FOR THE WSD TASK (PRECISION)

	word: *party*		word: *line*	
Learning Algorithm	**Training set**	**10-fold cross val.**	**Training set**	**10-fold cross val.**
AODE	81.46%	61.65%	72.14%	64.95%
Naive Bayes	72.58%	**61.97%**	68.41%	**66.33%**
K*	93.5%	59.43%	70.89%	61.61%
IB1	91.44%	55.78%	44.57%	53.47%
IBk, k=2	74.01%	60.22%	69.93%	64.42%
IBk, k=3	71.16%	61.97%	68.1%	64.13%
IBk, k=4	69.73%	61.81%	67.02%	64.01%

There were additional parameters to be **optimized**: the frequency threshold for determining the set of content words to be used, both in the local and global contexts.

To optimize the threshold for selecting content words from the local context window, I experimented with local content words whose frequencies in the training corpus was equal to or more than *f,* where *f* was 1, 2 or 3. Only this set of features was used. Experiments with *party* and *line* lead me to choose f=3 (Table 3.13.)

TABLE 3.13. EVALUATION OF LOCAL CONTENT WORD FREQUENCY THRESHOLD (PRECISION)

	word: *party*		word: *line*	
frequency >=	**Training set**	**10-fold cross val.**	**Training set**	**10-fold cross val.**
1	83.84%	59.75%	85.29%	71.01%
2	86.53%	70.11%	79.24%	73.90%
3	92.56%	**77.27%**	82.88%	**74.79%**
4	86.69%	72.27%	83.22%	74.05%
5	85.74%	74.80%	81.07%	72.55%

To optimize the frequency threshold for global content words, I experimented with selecting the top *n* content words in the training instances , where n was 1000, 500, 300, 200, or 100. N=300 was chosen after investigations with *party* and *line,* using only these features (Table 3.12).

DOI:10.15774/PPKE.ITK.2010.006

TABLE 3.14. EVALUATION OF GLOBAL CONTENT WORD FREQUENCY THRESHOLD (PRECISION)

	word: *party*		word: *line*	
Top n	**Training set**	**10-fold cross val.**	**Training set**	**10-fold cross val.**
1000	93.82%	68.94%	93.65%	88.70%
500	87.16%	68.94%	93.65%	88.48%
300	92.23%	**70.05%**	96.45%	**92.47%**
200	90.97%	68.30%	96.99%	91.82%
100	88.27%	67.04%	93.65%	88.59%

Finally, I was interested in what would be the optimal representation for the "content words in the global context" feature. The top 300 content words extracted from the training instances can be collected into a vector G. The value g_i in the vector could either be a) c_i = the actual count of the ith word in the training instance, or b) 1 if $c_i > 0$ or 0 otherwise. In other words, the value set is either numeric (positive integers and 0) or binary (0, 1). The Naive Bayes classifier in Weka uses kernel density estimators [85] to model numeric features. While this works better than assuming simple normal distributions, it is a question whether the classifier really needs this knowledge. I therefore evaluated the precision of the classifiers for all 38 nouns using 10-fold cross-validation using both numeric and binary representation for the values in the feature vector (Table 3.15). English sense classes were used in this experiment. The average precision with numeric values was 76.54%, with binary features it was 77.99%. A one-sided t-test showed that the improvement was significant (test value=1.53, critical value=1.51 at alpha=.07).

DOI:10.15774/PPKE.ITK.2010.006

TABLE 3.15. EVALUATION OF REPRESENTATION SCHEME FOR GLOBAL CONTENT WORDS (PRECISION)

Word	Numeric	Binary
arm	90.34%	90.22%
art	67.59%	75.00%
authority	59.53%	71.60%
bank	96.73%	96.48%
bar	60.24%	58.46%
bum	81.36%	83.05%
chair	87.43%	88.48%
chance	77.24%	81.46%
chapter	83.21%	79.56%
child	66.67%	71.67%
church	77.60%	74.32%
circuit	71.74%	72.83%
day	58.33%	61.98%
degree	86.60%	86.80%
dyke	87.21%	86.05%
facility	94.59%	94.59%
fatigue	93.27%	89.42%
feeling	51.01%	64.43%
grip	72.48%	76.61%
hearth	59.38%	64.58%
holiday	96.39%	96.39%
image	79.10%	78.91%
lady	82.84%	82.09%
letter	92.13%	92.23%
line	83.11%	80.08%
mouth	59.76%	62.13%
operator	78.15%	76.47%
party	75.12%	72.39%
performance	65.44%	59.49%
plane	97.26%	96.41%
post	79.43%	68.09%
process	76.82%	77.15%
report	81.79%	78.51%
restraint	71.91%	74.16%
sense	48.53%	69.85%
spade	85.39%	84.27%
stress	52.17%	54.78%
term	80.80%	92.80%

3.2.3. Evaluation

In order to evaluate the performance of all the classifiers, I trained them using the training sets described in 3.2.1. and then performed 10-fold stratified cross-validation on these sets (precision and recall were identical since the classifiers provided sense assignments for all input instances.) I was interested in the results of WSD using both the original PWN sense tags and the manually mapped Hungarian translations. For baseline

DOI:10.15774/PPKE.ITK.2010.006

value I used the relative frequency of the most frequent sense in each case. The results are shown in Table 3.16.

Table 3.16. Evaluation of WSD precision using 10-fold cross-validation using both English and Hungarian sense labels. "Delta" is the difference between precision and baseline for Hungarian (zero or negative instances set in bold)

	English		Hungarian		
Noun	**Baseline**	**Precision**	**Baseline**	**Precision**	**delta**
arm	55.15%	90.22%	57.05%	91.99%	34.94%
art	38.89%	75.00%	97.22%	98.15%	0.93%
authority	38.91%	71.60%	54.09%	82.49%	28.40%
bank	96.48%	96.48%	98.24%	98.49%	0.25%
bar	54.01%	58.46%	54.01%	58.46%	4.45%
bum	83.05%	83.05%	83.05%	83.05%	**0.00%**
chair	87.96%	88.48%	87.96%	88.48%	0.52%
chance	65.37%	81.46%	65.37%	81.46%	16.10%
chapter	67.15%	79.56%	67.15%	81.75%	14.60%
child	62.78%	71.67%	63.33%	75.00%	11.67%
church	58.47%	74.32%	58.47%	74.32%	15.85%
circuit	32.07%	72.83%	43.48%	79.89%	36.41%
day	34.90%	61.98%	65.10%	72.92%	7.81%
degree	74.43%	86.80%	74.43%	94.64%	20.21%
dyke	84.88%	86.05%	84.88%	86.05%	1.16%
facility	94.59%	94.59%	94.59%	94.59%	**0.00%**
fatigue	89.42%	89.42%	89.42%	89.42%	**0.00%**
feeling	54.36%	64.43%	92.62%	92.62%	**0.00%**
grip	48.62%	76.61%	92.20%	93.58%	1.38%
hearth	64.58%	64.58%	82.29%	82.29%	**0.00%**
holiday	96.39%	96.39%	96.39%	96.39%	**0.00%**
image	43.95%	78.91%	57.23%	85.55%	28.32%
lady	82.09%	82.09%	91.79%	91.04%	**-0.75%**
letter	84.90%	92.23%	84.90%	92.34%	7.44%
line	53.43%	80.08%	53.43%	81.62%	28.19%
mouth	49.11%	62.13%	94.67%	94.67%	**0.00%**
operator	73.95%	76.47%	73.95%	76.47%	2.52%
party	42.05%	72.39%	42.05%	83.79%	41.73%
performance	43.34%	59.49%	62.89%	86.97%	24.08%
plane	96.41%	96.41%	96.41%	96.41%	**0.00%**
post	63.12%	68.09%	63.12%	68.79%	5.67%
process	76.82%	77.15%	76.82%	78.15%	1.32%
report	67.76%	78.51%	67.76%	78.51%	10.75%
restraint	44.94%	74.16%	44.94%	77.53%	32.58%
sense	37.50%	69.85%	50.74%	77.21%	26.47%
spade	71.91%	84.27%	71.91%	84.27%	12.36%
stress	53.91%	54.78%	87.83%	87.83%	**0.00%**
term	70.40%	92.80%	70.40%	92.80%	22.40%
Average:	*64.16%*	*77.99%*	*73.48%*	*85.00%*	*11.52%*

DOI:10.15774/PPKE.ITK.2010.006

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, a 11.52% improvement on the average baseline. In Hungarian, all but 10 of the 38 classifiers performed above the baseline, and in only 1 case did the precision fall below the baseline, for the noun *lady* (delta=-0.75%).

The noun *lady* has 3 different English WordNet senses in the corpus, which were mapped to 2 different Hungarian translations, as shown in Table 3.17.

TABLE 3.17. ENGLISH AND HUNGARIAN SENSES AND NUMBER OF TRAINING EXAMPLES FOR THE NOUN *LADY*

PWN sense key	**PWN sense gloss**	**Number of instances in corpus**	**Hungarian translation**
lady_1:18:00::	(a woman of aristocratic family)	11	*lady*
lady_1:18:01::	(a woman of refinement)	13	*hölgy*
lady_1:18:02::	(a polite name for any woman)	110	*hölgy*
	Total:	*134*	

In English, the majority sense is *lady_1:18:02::*, the baseline score therefore is its relative frequency, 110/134=82.09%. In Hungarian, the sense *hölgy* subsumes English senses *lady_1:18:01::* and *lady_1:18:02::*, becoming the baseline sense with a higher probability of 123/134=91,79%.

In order to investigate the possible reason for the disambiguation precision falling below the baseline value in Hungarian, I examined the confusion matrices of the English and Hungarian classifiers after 10-fold cross-validation on the training corpus (see Table 3.18 and Table 3.19). In a confusion matrix M, the value $m_{i,j}$ shows the number of instances that have the ith correct sense label and were assigned to the jth class by the classifier (the main diagonal therefore contains the numbers of correctly classified instances for each sense class.)

TABLE 3.18 CONFUSION MATRIX FOR THE DISAMBIGUATION OF *LADY* OVER ENGLISH SENSE LABELS

	Classified as *lady_1:18:00::*	**Classified as *lady_1:18:01::***	**Classified as *lady_1:18:02::***
Sense *lady_1:18:00::*	11	0	0
Sense *lady_1:18:01::*	1	0	12
Sense *lady_1:18:02::*	0	0	110

DOI:10.15774/PPKE.ITK.2010.006

TABLE 3.19 CONFUSION MATRIX FOR THE DISAMBIGUATION OF LADY OVER HUNGARIAN SENSE LABELS

	Classified as *hölgy*	**Classified as *lady***
Sense *hölgy* (lady_1:18:01:: or lady_1:18:02::)	122	1
Sense *lady* (lady_1:18:00::)	11	0

All the instances in class *lady_1:18:02::* are correctly disambiguated in both Hungarian and English. In English, this is the baseline class. In Hungarian, the baseline class also includes the instances from class *lady_1:18:01::*. The confusion matrices reveal that there is an instance in the corpus which is misclassified in both languages (in English, it is assigned the *sense lady_1:18:01::* instead of the correct sense *lady_1:18:00::*). This means that in Hungarian, the disambiguation score falls below the baseline value, while in English, it is equal to the baseline.

To try to investigate the reason why precision does not exceed the baseline in Hungarian for 9 items, I looked at the number of training instances and the number of instances available for the least frequent sense for each word (Table 3.20). There is a weak correlation between exceeding the baseline and the number of training instances. For 8 of the 10 words (80%) not surpassing the baseline, the total number of training instances was not more than 150 and the number of training instances for the least frequent sense was below 20. Among the 28 nouns for which WSD precision exceeded the baseline, we only find 7 (25%) with such figures for the training data.

Mapping the English senses to Hungarian translations improved precision of the classifiers 7.01% on average. In 11 cases out of 38, the precision was identical for both sense inventories. Interestingly, this does not correlate with the cases where the number of senses was identical in both inventories (Table 3.20).

DOI:10.15774/PPKE.ITK.2010.006

TABLE 3.20. COMPARISON OF ENGLISH AND HUNGARIAN WSD PRECISION AND SENSE INVENTORY SIZE. $|S_{MIN}|$ IS THE NUMBER OF INSTANCES AVAILABLE FOR THE LEAST FREQUENT SENSE IN HUNGARIAN. ITEMS SET IN BOLD HAVE IDENTICAL PRECISION VALUES IN BOTH LANGUAGES.

		English		Hungarian		
Noun	**Instances**	**Senses**	**Precision**	**$\|S_{min}\|$**	**Senses**	**Precision**
arm	787	5	90.22%	16	4	91.99%
art	108	4	75.00%	3	2	98.15%
authority	257	3	71.60%	18	**3**	82.49%
bank	398	4	96.48%	7	2	98.49%
bar	337	7	58.46%	7	4	**58.46%**
bum	118	5	83.05%	20	2	83.05%
chair	191	8	88.48%	11	3	88.48%
chance	615	6	81.46%	21	4	81.46%
chapter	137	3	79.56%	45	2	81.75%
child	180	7	71.67%	66	2	75.00%
church	183	3	74.32%	76	2	**74.32%**
circuit	184	6	72.83%	25	4	79.89%
day	192	2	61.98%	67	**2**	72.92%
degree	485	4	86.80%	124	2	94.64%
dyke	86	4	86.05%	13	2	**86.05%**
facility	37	3	94.59%	2	2	**94.59%**
fatigue	104	4	89.42%	11	2	**89.42%**
feeling	149	3	64.43%	11	2	92.62%
grip	218	5	76.61%	17	2	93.58%
hearth	96	3	64.58%	17	2	82.29%
holiday	83	4	96.39%	3	2	**96.39%**
image	512	7	78.91%	219	2	85.55%
lady	134	4	82.09%	11	2	91.04%
letter	927	3	92.23%	140	2	92.34%
line	4157	6	80.08%	374	5	81.62%
mouth	169	2	62.13%	9	**2**	94.67%
operator	119	2	76.47%	31	**2**	**76.47%**
party	623	2	72.39%	108	3	83.79%
performance	353	2	59.49%	131	2	86.97%
plane	474	4	96.41%	2	3	**96.41%**
post	141	3	68.09%	18	**3**	68.79%
process	302	2	77.15%	70	**2**	78.15%
report	335	3	78.51%	42	**3**	**78.51%**
restraint	89	6	74.16%	2	4	77.53%
sense	136	4	69.85%	16	3	77.21%
spade	89	5	84.27%	4	3	**84.27%**
stress	115	3	54.78%	14	2	87.83%
term	125	5	92.80%	15	3	**92.80%**

In comparison to previous work, [78] reports 84% disambiguation precision for the noun *line* using a Naive Bayes classifier trained with 200 examples for the least frequent sense (about 57% of all available tagged instances), relying on similar contextual features with the addition of part-of-speech tags in the local context. My classifier for *line*, using

DOI:10.15774/PPKE.ITK.2010.006

the same corpus with English sense tags produced 78.28% precision averaged over 3 runs with random 57%/43% train/test splits.

3.2.4. Evaluation in Machine Translation

I integrated the WSD system described here into the MetaMorpho English-Hungarian machine translation system [39]. The task of the WSD module in this MT system is to specify the value of a special feature in the preprocessed source language translation units, which signifies the actual meanings of the ambiguous nouns. At this point, this sense feature is assigned one of the original PWN sense labels. After this, source language syntactic analysis is performed, and the grammatical analysis rules may use the specified values of the sense features. Along with the construction of the parse trees, the target language generation rules also prepare the translation structure. The mapping between English PWN senses and Hungarian lexical equivalents is defined in the target language generation rules of the ambiguous nouns. The system chooses the appropriate Hungarian translation as defined by the WSD module in the sense features. This solution has two advantages: on the one hand, the Hungarian translations are not "hard-wired" into the disambiguation engine, so the mapping from fine-grained English senses to Hungarian translations can be easily maintained. On the other hand, by replacing the lexical equivalents in the generation rules, it is possible to create machine translation from English to another target language, using word sense disambiguation.

I performed an evaluation of the WSD module operating in the MetaMorpho MT system with the aid of the *Bleu* evaluation methodology [86], which measures the quality of machine-translated text against human translations. The 3 English reference texts (total 4,500 words) contain 22 sentences with 10 of the 38 known ambiguous nouns. Human translators provided Hungarian translations for these texts. At the time of this experiment, the Bleu-index of the MetaMorpho system was 0.3513 (human-to-human translator Bleu scores range from 0.3972 to 0.4294) without using WSD (always selecting translations of the most frequent senses for the polysemous nouns). With the help of the WSD module, the Bleu-index changed to 0.3514. Because the number of treated ambiguous items and the number of test instances is very low, we can only maintain that the operation of the WSD module does not impair general translation quality, but rather presents a small increase.

DOI:10.15774/PPKE.ITK.2010.006

I also introduced the possibility in the MetaMorpho system to manually create disambiguation rules. A classifier for a previously unknown ambiguous item in the MT system can be set up relatively fast by manually analyzing occurrences of the word in corpora, then entering a few collocations, or other types of contextual information (using the available features) that can be used as evidence for either of the senses. An extension to the input format makes it possible to manually set the prior sense distributions for the Naive Bayes classifier, since the sense distribution in the manually crafted training rules usually does not represent real life proportions.

3.2.5. Obtaining Training Instances From a Parallel Corpus

Since semantically annotated training corpora are available only in limited quantity, I needed a solution for scaling up the system with new lexical items. 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 [92], [94].

The Hunglish Corpus [95] 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 (Table 3.21). I processed the English texts in the corpus with a PoS-tagger [93], and used the Humor morphological analyzer [38] 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.

TABLE 3.21.THE HUNGLISH CORPUS

Sections of Hunglish		Sentence pairs	Hungarian words	English words
[film]	movie subtitles	324,174	1,357,430	1,719,670
[law]	EU law	951,491	14,041,482	17,483,884
[lit]	literature	652,142	7,721,359	9,497,310
[mag]	magazines	10,276	58,855	67,238
[swdoc]	software documentation	135,472	594,030	673,648
Total		2,073,555	23,773,156	29,441,750

DOI:10.15774/PPKE.ITK.2010.006

I experimented with the polysemous English noun *state* to explore the problems that would arise when producing automatically tagged training corpora for supervised WSD inside an English-to-Hungarian MT system [9], [11].

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 [34], Princeton WordNet version 2.1, and the lexical translation pattern database of the MetaMorpho MT system. 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 (Table 3.21).

TABLE 3.22. COLLOCATIONS COLLECTED FROM THE DIFFERENT SOURCES

Source	Collocations
Lexical rules from the MetaMorpho MT system	131
English-Hungarian dictionary (*Országh: Angol-magyar nagyszótár*)	64
Princeton WordNet 2.1	218
Automatic terminology extraction from the Hunglish corpus + manual filtering	22
Total (duplicates removed)	348

With the help of the bilingual dictionary, I also compiled a list of all the possible Hungarian translations of *state* in its single-noun usage, gaining 19 different translations. I also added all the different adjectival and adverbial derivations of these (e.g., noun *állapot* – adjectives/adverbs: *állapoti*, *állapotú*, *állapotos*, *állapotbeli*, *állapotszerű* etc.), because observations in the Hunglish corpus revealed that the Hungarian side often contained such derived forms corresponding to *state* in the English side.

I created a sub-corpus of Hunglish by selecting sentence pairs where the English sentence contained the noun *state* (92,500 sentence pairs). I then classified these sentence pairs into 3 groups: a) sentence pairs that contained one or more of the known collocations, b) sentence pairs that contained one or more of the known collocations in

DOI:10.15774/PPKE.ITK.2010.006

addition to other occurrences of *state*, and c) sentence pairs that contained only unknown occurrences (none of the known collocations). The results are shown in Table 3.23.

TABLE 3.23. TYPES OF SENTENCES CONTAINING *STATE* IN THE ENGLISH SIDE IN HUNGLISH

sentences with *state*	**film**	**law**	**lit**	**mag**	**swdoc**	**Total**
Only collocational	155	84,880	645	93	41	85,814
Collocational + non-collocational	0	2,562	8	5	4	2,579
Only non-collocational	85	2,861	874	44	138	4,002
Total	240	90,303	1,527	142	183	92,395

In categories b) and c) I looked for zero, one or more occurrences of any of *state*'s known Hungarian equivalents in the Hungarian side. Sentence pairs containing exactly 1 translation equivalent in the Hungarian side, without any additional collocational occurrences constituted 2,473 training examples (the most frequent sense represented by 1,296 examples), see Table 3.24. Appendix A3. lists the identified Hungarian equivalents in these sentences with their frequencies.

The result that over 92% of all occurrences of *state* are collocational is explained by the fact that the majority of these occurs in the law subcorpus, which is a collection of European Union legal documents, where the collocation *member state* was inherently frequent.

TABLE 3.24. SENTENCES IN HUNGLISH CONTAINING ONLY NON-COLLOCATIONAL OCCURRENCES OF *STATE* IN THE ENGLISH SIDE

0 translation found	1 211
exactly 1 translation found	**2 473**
2 or more possible translations	318

[78] remarks that their supervised WSD classifier's accuracy levelled off when trained with enough examples to represent 200 instances for the least frequent sense. My experiments with supervised WSD showed a similar result (Section 3.2.3.) The automatically annotated examples for *state* – perhaps after a cut for some top *n* senses – can be readily used to train a classifier for English-Hungarian word sense disambiguation. The same method could be used to generate training data for other source language lexical items automatically, only requiring input from human annotators

DOI:10.15774/PPKE.ITK.2010.006

when compiling the lexicon of target language translations and source language collocations.

3.3. Summary

In this Chapter, I described my experiments with word sense disambiguation in machine translation. I experimented with supervised WSD classifiers to improve the lexical translation accuracy of an existing, rule-based English-Hungarian machine translation system. The classifiers were trained with openly available sense-tagged data, where I mapped the existing English WordNet sense tags to Hungarian translations. I experimented with 45 English nouns. After the English-Hungarian sense mapping, the average number of different senses dropped from 3.97 to 2.49.

I performed several experiments to choose the optimal learning algorithm and feature representation parameters for the classifiers. I used used the Naive Bayes algorithm and local and global contextual features. I evaluated the classifiers on the training data using 10-fold stratified cross-validation, using both English and Hungarian sense tags. The baseline algorithm in both cases was selecting the most frequent sense. The average precision of the classifiers for English sense tags exceeded the average baseline by 13.83%, for Hungarian by 11.52%. Using Hungarian sense tags improved average sense disambiguation precision by 7.01%.

I conducted experiments to explore the possibilities of automatically creating sense-tagged training examples for the MT WSD classifiers using a large sentence-aligned parallel corpus. I experimented with the English noun *state* using the Hunglish English-Hungarian parallel corpus. After preparing lists of translation equivalents and collocations formed by *state*, I searched the corpus for occurrences. Filtering out collocational and ambiguous instances, the method provided 2,473 instances where the Hungarian translation equivalents can be readily used as sense tags for training.

Related theses (see Section 5.1. for more details):

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.

DOI:10.15774/PPKE.ITK.2010.006

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.

II.3. I showed that annotated training examples for word sense disambiguation in rule-based 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.

Related publications: [9], [11], [13], [15], [16]

DOI:10.15774/PPKE.ITK.2010.006

Chapter 4

COREFERENCE RESOLUTION AND POSSESSOR IDENTIFICATION

4.1. Introduction

In the automatic processing of natural language texts, the discovery of **relationships** – **coreference**, **possession** etc. – between mentioned entities is an important procedure. The solution of this task is a valuable help for NLP applications such as machine translation, information extraction, automatic text summarization, opinion mining and others [106].

Coreference resolution (CR) means the identification of noun phrases (NPs) that appear at various points in a given document, but refer to the same real world entity. Anaphoric elements are expressions that refer back to previous, coreferring mentions of the same entity, called antecedents. The coreference relation forms chains of NPs in the document, starting from the first mention of an entity and including every coreferring anaphoric expression. There is a great range of linguistic phenomena that express anaphoric relationships: pronouns, various lexical semantic relationships (synonyms, hypernyms, hyponyms, meronyms, holonyms etc.), grammatical constructions (copulas, appositions etc.), various forms of names and many others (some of these will be illustrated in Section 4.2.).

The challenge in possession relationships identification is to recognize NP pairs in the document (possessor and possession) that are several words, phrases or even sentences apart. As I will show, part of this task in Hungarian is similar to pronominal coreference resolution, and the advantages of developing a solution to this problem are also important for many text-processing applications.

Coreference resolution has been researched for a long time in the history of natural language processing [114], [106], [72]. The earliest solutions to coreference/anaphora resolution employed knowledge-rich methods: rule-based, algorithmic solutions which relied on heuristics observed in anaphoric phenomena. In one of the very first such works, [99] describes a naive algorithm that uses only syntactic information available from the output of a parser. Antecedent candidates were identified in the parse tree, and

DOI:10.15774/PPKE.ITK.2010.006

were matched with features of person, number, gender and Binding Theory (BT) (see Section 4.2.1. for more details on BT). The algorithm was evaluated on 300 examples of personal pronoun *(he, she, it, they)* occurrences in three different texts. The algorithm resolved 88.3% of the cases correctly (81.8% correct when there were multiple antecedent candidates.)

Discourse-based methods, on the other hand, rely on theories for intrasentential anaphora in theoretical linguistics, mainly Centering Theory by [98], which models the attentional salience of discourse entities, and relates it to referential continuity [114]. The Brennan-Friedman-Pollard (BFP) algorithm [96] is the most well-known such approach. It uses syntactic and morphological criteria like number and gender agreement to eliminate, and centering principles to rank antecedent candidates [114]. [115] presents a modification of the CT-based approach called the Left-Right Centering (LRC) approach. Based on evidence from psycholinguistic research, the LRC works by first trying to find an antecedent in the current utterance, and if it fails, then antecedents in previous utterances are considered, going from left-to-right within an utterance [114].

The work of [102] integrated several sources of knowledge (syntax, semantics, morphology and discourse phenomena). Their algorithm filtered antecedent candidates based on person, number and gender agreement and Binding Theory rules, then calculated salience using several criteria (recency, distance and various structural configurations.)

[103] augmented [99]'s algorithm by statistical information obtained from corpora (a coreference-annotated subsection of the Penn Treebank.) They used the training data to learn a probabilistic model for information like distance between anaphora and antecedent, syntactic constraints, person, number and gender information, and mention count. An evaluation on 2477 personal pronouns showed 82.9% accuracy.

The development of machine learning methods and the profileration of knowledge-poor, corpus-based approaches in NLP has also brought a change in coreference resolution approaches [114]. In the supervised machine learning paradigm for CR, binary classifiers are trained from annotated data, and are applied to decide if there is coreference relationship between an anaphoric element and each NP preceding it. Then, a separate clustering mechanism coordinates the possibly contradictory pairwise coreference classification decisions and constructs a partitioning on the given set of NPs, with clusters corresponding to coreference chains [108]. One of the first such approaches

DOI:10.15774/PPKE.ITK.2010.006

was introduced by [109], whose performance was comparable to earlier, knowledge-based solutions. Their classifier used the C5.0 decision tree algorithm, which was trained by feature vectors of 12 different features (the type of NP, distance between anaphoric element and antecedent candidate, agreement of person, number, gender and semantic class, proper names and appositions etc.) They evaluated their system on the MUC-6 and MUC-7 [110] coreference datasets, and measured an F-measure of 62.6% (precision 67.3%, recall 58.6%) and 60.4% (precision 65.5%, recall 56.1%). Their system performed significantly better than some MUC-6 and MUC-7 participant systems, while its performance was not behind the others (all other systems employed non-learning approaches.)

[107] extended [109] by using a total of 53 features representing lexical, semantic, grammatic and other knowledge-based features. Surprisingly, using all of these features resulted in decreased precision, so in a modified approach, they used hand-selected features to increase precision on the worst-performing NP types (common nouns). They reported F-measure of 70.4% on the MUC-6 dataset (as opposed to 62.6% F-score by [109].)

[108] improves [107] by focusing on the clustering mechanism operating after the CR-relation classifier. As opposed to using ad-hoc greedy clustering algorithms, [108] proposes a ranking model to select the optimal candidate partitioning from a space of .various CR system parameters (learning algorithm, features, instance creation and clustering algorithm). Using the ACE coreference corpus [111] for evaluation, [108] reports a 5-10% improvement in F-score over [107] (depending on the evaluation metrics used.)

In a more recent work, [117] uses 351 different, linguistically motivated learning features, relying on different sources of knowledge (syntactic, semantic and discourse phenomena, plus character-based methods), comparing 5 different machine learning algorithms. Using the MUC-7 dataset for evaluation, the best F-measure was obtained using the C4.5 algorithm (64.6%, while a reimplementation of [109] produced 60.4%) and best precision with a maximum entropy learner (72.2%, versus 65.5% by reimplementation of [109]).

In the only prior attempt to coreference resolution in Hungarian, [105] applied the BFP algorithm to resolve pronouns, zero pronouns and adverbial anaphora (pronouns like *ott* („there") etc.) The algorithm generated anaphora-antecedent pairs, filtered out obviously

DOI:10.15774/PPKE.ITK.2010.006

unsuitable candidates and scored the remaining using the ranking of transitions between the sentences. A small part of the Szeged Treebank [97] was annotated with coreference to test the algorithm, which produced 39.6% precision and 21% recall.

4.2. Experiments

My goal was to develop a coreference resolution system for NPs in Hungarian texts. My proposed system deals with the following types of NP-coreference:

TABLE 4.25. TYPES OF NP COREFERENCE HANDLED BY THE PROPOSED SYSTEM

Type	Example (Hungarian, English)
Repetition	*Tegnap találkoztam **egy ismerősömmel**. **Az ismerősöm** nagyon sietett.* *"*I met **an acquaintance** today**. My acquaintance** was in a hurry."
Proper Name Variant	***Kovács Jakab, az ABC Kft. igazgatója** tegnap sajtótájékoztatót tartott. Az eseményen **Kovács úr** bejelentette az új termékeket.* **"Jakab Kovács, chairman of ABC Ltd.** held a press conference today. **Mr. Kovács** announced the new products."
Synonym	*Tamás kapott **egy biciklit**. Én is láttam **a kerékpárt**.* "Tamás got **a new bicycle**. I saw **the bike**, too."
Hypernym	*Bejött **egy kutya**. **Az állat** fáradtnak tűnt.* **"A dog** just came inside. **The animal** seemed tired."
Pronoun	*Beszéltem **Julival**. Megadtam **neki** a számodat.* "I talked to **Juli**. I gave **her** your phone number."
Zero Pronoun	***Viktor** ismeri **Ferit**, de (**ő**) nem kedveli (**őt**) túlságosan.* **"Viktor$_1$** knows **Feri$_2$**, but **he$_1$** doesn't like **him$_2$** very much."

The case of zero pronouns is a phenomenon in Hungarian when the pronominal arguments of the main verb are phonologically empty – the suffixes on the verb carry information about the number and person of the arguments –, but otherwise require the same treatment as regular, phonologically not empty personal pronouns.

My proposed system does not deal with cataphora (the antecedent is preceded by the anaphora). It also does not handle components of complex noun phrases (possessive structures, coordination, appositions, deverbal nouns with their arguments etc.), only simple, maximal NPs corresponding to the arguments of the main verb or its nominal

DOI:10.15774/PPKE.ITK.2010.006

modifiers. At this stage, the system only handles personal pronouns (and a certain type of demonstrative pronoun, see later), but no other types of pronouns.

At the time of my research, there was no hand-tagged corpus of coreference-annotated examples available for Hungarian that would be necessary for experiments with supervised machine learning methods. For this reason, I had to commit myself to a rule-based approach in the design of the CR system, utilizing as many as possible of the available state-of-the-art knowledge sources.

The rest of this Chapter is organized as follows: Section 4.2.1. describes the underlying linguistic and other theories and the resources that provide the knowledge for my methods. Section 4.2.2. explains the details of the proposed coreference resolution algorithm, and Section 4.2.3. presents the results of its evaluation against a small, manually annotated corpus. My proposed solution for a task closely related to CR in Hungarian – identification of certain types of long-distance possessor-possession relationships – is presented in Section 4.2.4., and its evaluation in 4.2.5.

4.2.1. Knowledge Sources

The proposed rule-based coreference resolution algorithm for Hungarian relies on information from several knowledge sources. The most important input is the morphological, syntactic and semantic information available from the output of the MetaMorpho MT system's deep parser [40]. Rules based on Binding Theory in syntax [100] and the results of psycholinguistic research on Hungarian sentence understanding [113], [112] (see below) operate on these structures. Rules based on semantic relationships and world knowledge utilize information available from Hungarian WordNet [2], [6]. For the matching of proper name variants, I employ character-based heuristics, similar to some of those described by [116].

Coreference resolution starts with the linguistic analysis of the entire input document, using the MetaMorpho parser. The text is segmented into paragraphs and sentences, and for each sentence I assign a simplified version of the original parse tree. These trees contain only nodes for the clauses of the sentence (main, coordinate and subordinate clauses) and the maximal verb phrases (VPs) and noun phrases (NPs). The parser is often – especially in the case of long, complex sentence – not able to produce parse trees that cover all the tokens of the input sentence, in these cases I use the trees available for the partial analyzes of segments of the sentence (VPs, NPs and nominal adverbial phrases

DOI:10.15774/PPKE.ITK.2010.006

(ADPs)). In the identified noun phrases – the input units for coreference resolution – 25 different features are used to describe lexical, morphological and semantic properties:

- *lexical*: head of the NP (base form), morphological information (case, person, number, owner number and person, post-positional modifiers), type of the NP (common noun, proper name, pronon or zero pronon), type of determiner (definite, indefinite, quantifier or none), type of pronoun if applicable (personal, reflexive, demonstrative etc.)
- *syntactic*: grammatical role of the NP (either an argument of a VP with a specific function (SUBJ, OBJ, COMPL) or a free modifier (MOD) or UNKNOWN (in incomplete parses), thematic role of an argument NP assigned by the VP, NP is coordinated or not, possessive status (NP is a complete possessive structure (2 types), or single possession or possessor), position relative to the main verb (before, after), syntactic class of the NP in the parse tree
- *semantic*: based on the head, two binary syntactic-semantic features (a 3rd, underspecified NIL value is also possible): ANIMATE, HUMAN, plus a feature for the semantic class of the head (abstract, bodypart, currency etc.) using lexical information coded into the parser's lexicon.

Features on VPs, taken from the parser, specify: head information (base form, morphological features), prefix modifier information, pointers to its governed NPs, pointers to arguments that refer to subordinate clauses of the VP.

Government and Binding Theory is one of the major results of transformational grammar [101] dealing with the properties of pronominal and anaphoric elements. It attempts to provide a universal [100] structural account of the distribution and coreference conditions of pronouns, reflexives and R-expressions (referential expressions: common nouns, proper names etc.). Its 3 basic principles are formulated as the following [100]:

- *Condition A*: A reflexive must be bound in its governing category (i.e., the reflexive pronoun and its antecedent can't be far away.)
- *Condition B*: A pronoun must be free in its governing category (i.e., a non-reflexive pronoun and its antecedent may not occur in the same clause, and the antecedent of a pronoun must precede or command[3] the pronoun)

[3] Node X commands node Y if and only if the node directly dominating node X also dominates node Y, and X does not dominate Y [100].

DOI:10.15774/PPKE.ITK.2010.006

- *Condition C*: An R-expression must be free (i.e., cannot have an antecedent).

Currently, various modifications of BT as well as additional alternative theories of binding exist in formal linguistic theories [114]. Its basic principles, however, present themselves in syntactic components in various approaches to coreference resolution.

4.2.2. Coreference Resolution Methods

Coreference resolution for a given NP in the input document is based on satisfying constraints, in order to eliminate as much as possible from the antecedent candidates, and evaluating preferences in order the select the most likely candidate ("Constrains and Preferences" approach, [106]). The various parameters of the algorithm – method for generating the list of antecedent candidates, filtering the list and finally selecting the winning candidate – are specific to the type of the anaphoric NP (proper name, definite common noun, pronoun/zero pronoun). The general algorithm for processing an input document is the following:

1) *Pre-filtering*: NPs assumed to be anaphoric are identified for further processing. The system attempts to recognize and exclude NPs that are not treated (non-personal pronouns etc.), and also formally anaphoric expressions that refer to entities outside of the texts (exophoric expressions) [118]. At this point I also included heuristics that attempt to recognize NPs that were likely analyzed incorrectly by the parser and would only introduce further errors in CR and therefore should be excluded from further processing. The system uses the following criteria to identify such NPs:
   a. The grammatical role of the NP is UNKNOWN (NP is not governed by the main VP).
   b. The head of the governing VP is *van* (copular verb), and the whole VP does not cover more than 2 tokens.
   c. The head of the NP is *az* (demonstrative pronoun), and he whole VP does not cover more than 2 tokens.
   d. The parse tree containing the NP is partial (does not cover all tokens in the sentence) and the head of the governing VP is *van* but it is phonologically empty (nominal predicate with 3rd person subjects).

DOI:10.15774/PPKE.ITK.2010.006

e. The parse tree containing the NP is partial (does not cover all tokens in the sentence), and there is another, not zero pronoun NP in the sentence that is not under the same VP and whose case is the same as this NP's case (the main verb's argument exists in the sentence, but was not recognized under the VP).

2) *Generating the list of antecedent candidates:* in this step, with method depending on the type of the anaphor, the system goes back up to a given distance in the document and lists the NPs that are compatible with the anaphor and may be potential antecedents. In accordance with Binding Theory, not even the closest antecedent candidate can fall under the VP of the anaphor (since the system does not handle reflexive pronouns).

3) *Filtering of the candidates*: the system attempts to exclude as many as possible from the candidates (method specific to the type of anaphor), and also applies the incorrect parse recognition heuristics listed for step 1.

4) *Selecting the antecedent:* an antecedent is selected from the remaining candidates, with method depending on the type of the anaphor. Certain types force the system to choose one of the candidates, while certain types allow one or zero candidate to be selected.

In the following, I will describe the specific algorithm parameters for the various types of anaphora in detail.

For **proper names**, the list of antecedent candidates consists of all the proper names prior to the anaphor in the entire document. At present, no filtering is applied to these candidates. The winning antecedent candidate is the one having smallest Minimum Edit Distance (MED) with the anaphor, normalized by the length of the longer string. Both antecedent and anaphor are normalized before the string matching: determiners are removed from the beginnings of the names, and the head word is lemmatized. The rule selects an antecedent only in case the MED for the closest candidate falls below a preset threshold (I used a value of 0.7). This way, the system is not forced to select one from the available candidates in each case (it is possible that the NP has no antecedent in the text.)

DOI:10.15774/PPKE.ITK.2010.006

For definite **common nouns**, the system first tries to exclude mentions that refer to unique objects inferable from common world knowledge (e.g. “the president of the United States”). At present, this is done by searching a predefined list of such NPs.

The antecedent candidates are the proper names and common nouns (any type of determiner) 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.)

Selecting the antecedent is done by identifying the closest candidate that has the same head (repetition), or the closest synonym or hypernym/hyponym. Synonymity is checked using Hungarian WordNet: if there is a synset that contains both anaphor and candidate, they are considered synonyms. Since there is no word sense disambiguation, lexical ambiguities probably do add a level of noise to the algorithm.

I use the Leacock-Chodorow similarity formula [104] to measure semantic relatedness via the hypernym/hyponym paths connecting all the possible senses of the anaphor and the candidate in HuWN (lexical forms of the heads are used.) The closest candidate that falls below a preset threshold is considered the winning antecedent, but only if no identical (repeated) or synonymous candidate was found before. The threshold was configured to accept candidates available in WN not further than 2 edges away in the hypernym tree (allowing longer paths seems to generate too many unwanted unrelated connections.) Hypernym and hyponym candidates are only selected from the sentence preceding the anaphor's sentence in order to rule out further unwanted incorrect connections.

In the case of **pronouns**, the system first excludes every pronoun from processing that is not a personal pronoun, zero pronoun, or the special *az* demonstrative pronoun provided that it is in subject position and does not refer to a subordinate sentence. The system also excludes first and second person (single) deixic pronouns and zero pronouns, referring to entities in the context of the discourse, not inside it.

The antecedent candidates are collected from up to 2 sentences before the anaphor’s sentence, plus the clauses prior to the clause containing the anaphor in its sentence. All types of NPs in this scope are considered.

The antecedent candidates are filtered by checking person, number and 2 semantic features specified by the parser (ANIMATE and HUMAN.) The semantic features can have underspecified values in the case of zero pronouns and lexically ambiguous nouns, these

DOI:10.15774/PPKE.ITK.2010.006

are compatible with all other values. The filtering process also excludes candidates that have already been identified as antecedents of other NPs in the current clause (in accordance with Binding Theory.)

If there is more than one pronominal anaphor in the current clause, the system always processes the one with subject role first. This way, by the exclusion of already bound antecedents, instances of non-subject pronominal anaphora can be resolved by simple exclusion.

Identifying the antecedent of the pronoun or zero pronoun that is the subject in its VP follows the results of research on sentence understanding in Hungarian psycholinguistics [113]. The heuristic first 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, which indicates change of subject:

(4.11) *a.* ***Hugó**$_j$ felhívta **Amáliát**$_k$. **(Ő$_j$)** Elmondta **neki**$_k$ a történetet.*
("**Hugo**$_j$ called **Amália**$_k$. **He**$_j$ told **her**$_k$ what happened.")

b. ***Hugó**$_j$ felhívta **Amáliát**$_k$. **Az**$_k$ elmondta **neki**$_j$ a történetet.*
("**Hugo**$_j$ called **Amália**$_k$. **She**$_k$ told **him**$_j$ what happened.")

[113] describes other indicators of subject change (such as semantic preference of arguments by predicates), but at the present stage, the system does not deal with these phenomena. If the preceding clause does not contain a subject-role NP after filtering, the algorithm moves on to the subject of the previous clause, but going not further than the 1st sentence of the current paragraph. This reflects the heuristic that personal pronoun anaphora will usually not refer back further than a single discourse segment (a paragraph.)

In case there are more than one non-subject NPs in the prior clause, the antecedent is selected using the following criteria, based on observations by [113]:

1. *Accessibility*: The NP higher in the obliqueness hierarchy (object argument < other arguments < free modifiers) is selected.
2. *Distance*: The NP that is closer to the anaphor is preferred (among items on the same level in the obliqueness hierarchy.)

DOI:10.15774/PPKE.ITK.2010.006

Resolution of pronouns and zero pronouns with grammatical roles other than subject is also based on the above two criteria.

The system performs CR for common nouns and proper names before resolving pronouns within a sentence. This is done in order to further help the resolution of pronouns by using the above-mentioned filtering conditions.

4.2.3. Evaluation of Coreference Resolution

In order to assess the performance of my CR system, I compiled a small corpus of 10 excerpts from history textbooks (one of the focus areas of the psycholinguistic text processing project that utilizes the CR system (Section 5.2.)) The texts in the corpus were processed with MetaMorpho to annotate structural boundaries and NPs. The NPs identified by the parser were manually annotated by the ids of their closest antecedents in their coreference chains.

It should be noted that since the parser did not recognize all possible NPs in the text (and some were recognized incorrectly: wrong boundaries, wrong type etc.), and since only the correctly recognized NPs were annotated (and only if their closest antecedents were also present in the analysis), the corpus is not suitable for the evaluation of CR recall. Table 4.26 shows the statistics of the evaluation corpus.

TABLE 4.26. STATISTICS OF THE CR EVALUATION CORPUS

Texts	10
Paragraphs	31
Sentences	99
NPs	488
NPs annotated with their closest antecedents	111

I used 14 different types of NP coreference in the manual annotation, which covered 111 NPs in the corpus. 5 of these types are handled by the CR system, which gives 81 annotated NPs for testing. Table 4.27 shows the distribution of the various types of NP coreference annotated in the corpus.

DOI:10.15774/PPKE.ITK.2010.006

TABLE 4.27. DISTRIBUTION OF NP COREFERENCE TYPES MANUALLY ANNOTATED IN THE CORPUS (IN BOLD TYPES HANDLED BY THE CR SYSTEM)

Coreference Type	Number of occurrences
Personal pronoun, zero pronoun	**47**
Repeated NP	**15**
Proper name variant	**14**
Demonstrative pronoun	8
Frame	7
"that"-clause	6
Hypernym	**3**
Relative pronoun for relative clause	5
Synonym	**2**
Apposition	1
Copula	1
Hyponym	**1**
Meronym	1
Holonym	**0**
Total:	*111*

I performed coreference resolution for the texts in the corpus with the implementation of the proposed system and compared the results to the manual annotation. I regarded automatically tagged references correct that were not identical to the annotated reference but belonged to the same coreference chain, i.e. referred to the same entity. I calculated precision (the ratio of correctly resolved NPs to the number of NPs tagged by the system) for each type of coreference handled, shown in Table 4.28:

DOI:10.15774/PPKE.ITK.2010.006

TABLE 4.28. PRECISION OF THE VARIOUS CR METHODS

Anaphora type	**Manually annotated NPs**	**Automatically annotated NPs**	**Automatically, correctly annotated NPs**	**Precision**
Proper name	14	15	12	80.00%
Pronoun, zero pronoun	46	35	25	71.43%
Repetition	15	18	13	72.22%
Synonym	2	4	1	25.00%
Hypernym	4	2	0	0.00%
Total/Average:	*81*	*74*	*45*	*68.92%*

A first look at the results confirms that the system performs fairly well (precision 71-80%, recall 61-83%) for the most frequent types of anaphora currently handled in the corpus (proper name variant matching, repeated forms of common nouns and pronouns, zero pronouns.) On the other hand, the performance of the synonym and hypernym heuristics was poor, but since the evaluation corpus contained only a small number of such instances, this figure might not reflect realistic evaluation.

I was interested in how the performance of the parser affected coreference resolution in this architecture, therefore I also conducted an examination of the types of errors produced by the system. I examined each automatically assigned coreference link and assigned it to one of four categories:

- *OK*: antecedent marked by system is identical to manual annotation (correct).
- *OK_equ*: antecedent marked by system is not identical to manual annotation, but refers to the same entity (in the coreference chain), so it was also regarded correct.
- *KO_parser*: antecedent marked by system is different from manual annotation (ie. incorrect), and the error is due to incorrect analysis by the parser (if the parser would have provided correct results, the automatic coreference assignment would have been correct.)
- *KO_cr*: incorrec result; the antecedent was present in the text and the parsing was correct, the mistake was committed by the CR algorithm.

DOI:10.15774/PPKE.ITK.2010.006

As it can be seen from Table 4.29, about half of all the mistakes committed by the CR system are results of errors in parsing (incorrect NP boundaries, incorrectly assigned zero pronouns etc.) Having perfectly parsed input would increase overall precision to 75%, pronoun/zero pronoun resolution precision to 91%. This indicates that the proposed method is rather sensitive to parsing accuracy in the input.

TABLE 4.29. TYPES OF ERROR COMMITED BY THE SYSTEM

Anaphor type	**OK**	**OK_equ**	**KO_parser**	**KO_cr**
Pronoun, zero pronoun	19	6	7	3
Repetition	13	0	4	1
Proper name	12	0	0	3
Hypernym	0	0	0	2
Synonym	1	0	0	3
Hypernym	0	0	0	0
Total:	45	6	11	12

Finding not exactly matching, but referentially equivalent antecedents (marked *KO_equ*) is a phenomenon only observed in the case of pronouns/zero pronouns. Tracing back to the beginning of coreference chains in order to label the first mention of each entity as antecedent would result in lower precision, due both to parsing errors and CR errors.

I also experimented with a second round of evaluation in order to compare my results to a previous work on Hungarian anaphora resolution by [105], which uses an implementation of the BFP algorithm [96]. The only coreference type handled by both [96] and my system and therefore a basis for comparison is zero pronouns in subject position. I selected 3 news articles from the Szeged Treebank [97], which has accurate, manually created syntactic annotation. There were 15 anaphora occurrences in the selected articles, which were first manually labeled with their antecedents and then compared to the results of applying [96] and my system. Due to the small size of the evaluation corpus, both systems had very low coverage (4 anaphors attempted by [96] and 3 by my system), of which 3 were correct (75% and 100% precision respectively).

DOI:10.15774/PPKE.ITK.2010.006

4.2.4. Possessor Identification

I was also interested in researching the problem of identifying expressions in Hungarian texts that correspond to entities having the of relationship of possession and its respective possessor. More specifically, I was interested in dealing with cases where words, phrases or even sentence boundaries would appear between possessor and possession.

In Hungarian, there are three different such detached possessive structures (possessor and possession in bold):

a) Special possession predicate:
Jánosnak *van egy nagy, sárga* ***esőkabátja****.*
*(János+DAT has a big, yellow raincoat+POSS – "***János*** has a big yellow* ***raincoat****")*

b) Detached dative case possessor:
Jánosnak *ellopták a* ***könyvét.***
*(János+DAT stolen book+POSS – „***János*** had* ***his book*** stolen*")*

c) Zero pronoun possessor:
János *tegnap itt hagyta az* ***(ő) esernyőjét****.*
*(János yesterday here left the (his) umberella+POSS – „***János*** yesterday left his umberella* ***here****")*

In type a) sentences, the copula expresses a special possession relationship between the subject (possession) and the dative-case complement (possessor). In type b) sentences the possession structure is originally a complex NP, which is detached (e.g. topicalization etc.) Type c) sentences are more of a discourse phenomena, where the hearer chooses the possessor from entities introduced earlier.

The MetaMorpho deep parser component of my proposed coreference and possession relationship identification system is able to handle type a) and b) phenomena using its grammar (provided that the parse was complete). For these, I directly used the available possession relationship pointers available in the output of the parser.

DOI:10.15774/PPKE.ITK.2010.006

In type c) sentences, I proposed a method to resolve the detached possessor similar to that proposed earlier for pronominal coreference resolution. I assume that

1. The subject of the VP dominating the possession NP is the default possessor.
2. The possession agrees in number and person with the owner-number and owner-person morphological features (specifying the person and number of the possessor) on the possession.

The second assumption may override the first, so if the subject of the possession's VP does not agree in person and number, the algorithm may go back to a prior clause or sentence (while it stays in the same discourse segment (paragraph):

(4.12) ***János*** *elutazott nyaralni. Én vigyázok a* ***lakására****.*
*("****János*** *went on vacation. I am taking care of* ***his flat****."*

The algorithm also has to take into account that in Hungarian, the owner-number and owner-person morphemes are allomorphs for 3rd person in singular and in plural: *Ádám almája* (*Adam apple+OWN3* – „Adam's apple"), *A lányok almája (The girls apple+OWN3* – "The girls' apple").
Considering the above, the algorithm for identifying possessors for type c) sentences is the following:

1) The system identifies subject role NPs in the clauses before the clause containing the possession, up to two clauses away, but not further than the beginning of the current paragraph.

2) The system selects the NP as possessor that is closest to the possession (rightmost NP) from the candidates that agree in person and number with the possessor's owner-person and owner-number (allowing both singular and plural for owner-person=P3).

If there is no complete sentence parse available (no information about the grammatical role of the NPs) in the parser's output, the system selects the rightmost, nominative case NP before the possession NP that matches in person and number.

DOI:10.15774/PPKE.ITK.2010.006

4.2.5. Evaluation of Possessor Identification

For the evaluation of possessor identification I used the same hand-tagged corpus that I used in the evaluation of coreference resolution (Section 4.2.3.). 38 out of the 488 NPs present in the 10 texts had detached possessor NPs at earlier points in the text. The ids of the possessor NPs were manually annotated on the possession NPs.

I applied the possessor identification system to this corpus, which annotated the ids of assumed possessor NPs on possession NPs, and also indicated the source of the decision: a) MetaMorpho, using possession predicate rules (type a) sentences), b) MetaMorpho, detached dative possessor rules (type b) sentence), c) heuristics (zero pronoun possessor). The following values were computed:

- True positives: number of NPs that contained both manual and automatic possessor annotation whose values were identical.
- Mistakes: number of NPs that contained both manual and automatic possessor annotation whose values were not identical.
- False positives: number of NPs that contained only automatically assigned possessor annotation.
- False negatives: number of NPs that contained only manually assigned possessor annotation.

Using these, I defined precision, recall and F-measure using the following formulas:

precision = |true positives| / (|true positives| + |mistakes| + |false positives|)
recall = |true positives| / (|true positives| + |mistakes| + |false negatives|)
F-measure = (2 * precision * recall) / (precision + recall)

I calculated precision, recall and F-measure for overall performance. For each possessor identification strategy I also calculated precision separately, but recall could not be computed since the evaluation corpus did not contain manual annotation for this information. The results are shown in Table 4.30.

DOI:10.15774/PPKE.ITK.2010.006

TABLE 4.30. EVALUATION OF POSSESSOR IDENTIFICATION FOR THE DIFFERENT DETACHED POSSESSION TYPES

	Type a) (Parser)	**Type b) (Parser)**	**Type c) (Heuristics)**	**Overall**
True positives	0	6	20	26
Mistakes	0	0	7	7
False positives	0	0	1	1
False negatives	-	-	-	5
Precision	0	100.00%	71.43%	76.47%
Recall	-	-	-	68.42%
F-measure	-	-	-	72.22%

Table 4.30 shows that the evaluation corpus did not contain any instances of type a) that the system tried to resolve. The parser resolved all of the type b) instances correctly that it attempted. The precision of type c) possession identification, using my heuristics was 71.43%. The overall performance of the system (parser-based and heuristic strategies combined) evaluated to 72.22% F-measure (76.47% precision and 68.42% recall).

While possessor identification, like coreference resolution, relies heavily on the parser and is therefore sensitive to parsing errors, there are situations when the above algorithm is not sufficient. The following excerpt (about fighting tactics of Hungarian light cavalrymen) presents an example:

> *Az első, ellenséggel való összecsapás után menekülést színlelve megfordultak, és futásnak eredtek. **Az ellenfél** ekkor üldözőbe vette őket. Ez lett a **vesztük**. A harcosok a vágtató ló hátán „kengyelbe állva", hátrafordulva lenyilazták üldözőiket.*
>
> *("After the first clash with the enemy, they turned around pretending to flee. **The enemy** started to pursue them. This became **their** fate. The warriors shot their pursuers shooting their arrows turning backwards on their galloping horses.")*

To interpret the possession relationship between the entities set in bold in the above example we need additional world knowledge. Since the possessor and the possession

DOI:10.15774/PPKE.ITK.2010.006

don't match in number, we need to rely on the knowledge that the possessor NP (which is single in number) denotes a group of people and can therefore be referenced with an anaphor plural in number. Such problems could be solved by incorporating this kind of information in the supporting ontology.

DOI:10.15774/PPKE.ITK.2010.006

4.3. Summary

In this Chapter, I discussed my experiments with rule-based algorithms for NP-coreference and possessor identification in Hungarian texts.

The coreference resolution algorithm relies on syntactic, semantic and morphological information from the MetaMorpho deep parser and semantic information in Hungarian WordNet. The algorithm is based on the constraints and preferences approach. The candidate generating, filtering and ranking rules vary for each type of anaphor. For proper names, I used minimum edit distance and string normalization. For definite common nouns, I examined the head of previous mentions for repetition or semantic relatedness via the hypernym hierarchy in Hungarian WordNet. For pronouns and zero pronouns, I used rules based on Binding Theory, morphological and semantic feature agreement and the results from Hungarian sentence understanding experiments.

The identification of detached possessor relationships that are not handled by the parser is based on morphological agreement and subject preference constraints. Both coreference and possessor identification algorithms are equipped with a set of heuristics to recognize common parser errors in order to improve precision.

I evaluated both algorithms on a small, hand-tagged corpus. Overall precision of coreference resolution was 68.92%, average recall was 62.96%. Examination of error types showed that the CR algorithm is sensitive to parser errors. Precision of my possessor-possession identification was 71.43%.

Related theses (see Section 5.1. for more details):

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.

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

Related publications: [1], [5], [7]

DOI:10.15774/PPKE.ITK.2010.006

Chapter 5

SUMMARY

5.1. 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 ontology for Hungarian nouns.

The first group of heuristics for automatic synset translation were proposed by [30], [31] 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 [31] 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 [65] was then applied to 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) [33]. 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 [38] 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 [30], [31] on Hungarian data. In the first evaluation 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 for each heuristic using this set. The methods from [30], [31] in my implementation ranged in precision 49-92%, while [30] reports 61-85% on the manual evaluation of a 10% sample. Following [30], I also experimented with different

DOI:10.15774/PPKE.ITK.2010.006

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 [30] 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

DOI:10.15774/PPKE.ITK.2010.006

WordNet project [2], [6]. 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 5.31.

TABLE 5.31.: EVALUATION RESULTS OF SYNSET TRANSLATION METHODS AGAINST HUNGARIAN WORDNET

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

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

DOI:10.15774/PPKE.ITK.2010.006

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 rule-based 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 bilingual MRDs, 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 also available that contain a collocation and exactly one translation.)

DOI:10.15774/PPKE.ITK.2010.006

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

DOI:10.15774/PPKE.ITK.2010.006

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 [112], [113]. 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, 111 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

DOI:10.15774/PPKE.ITK.2010.006

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.

5.2. Applications

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 [23]. 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

DOI:10.15774/PPKE.ITK.2010.006

full parses of the MetaMorpho parser [40]. The semantic constraints were formulated by mapping semantic classes used in the event frames to hierarchies in the nominal Hungarian WordNet ontology (Figure 5.8).

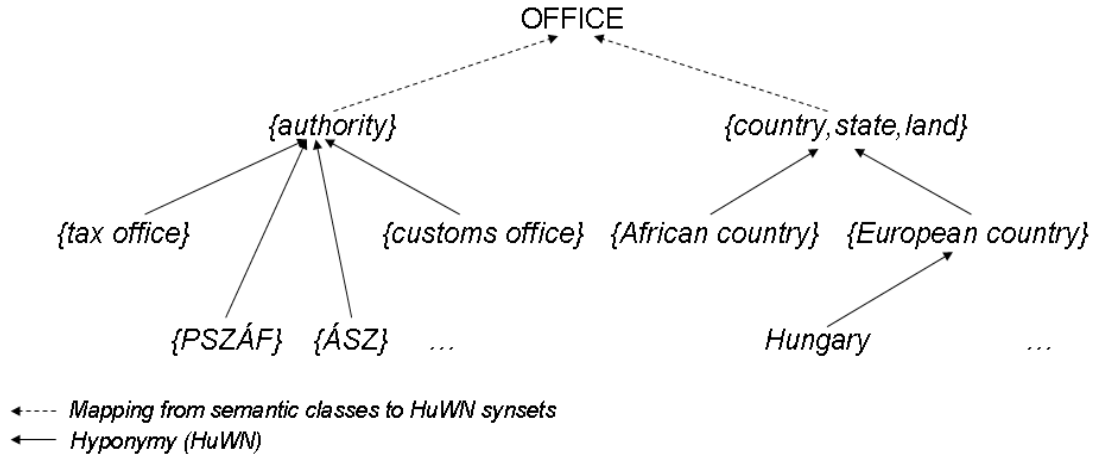

Figure 5.8: Mapping semantic classes used by the information extraction engine to concepts in the Hungarian WordNet ontology

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* [119], [120] realized by a

DOI:10.15774/PPKE.ITK.2010.006

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 automatically 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. Figure 5.9 demonstrates how my coreference and possession identification methods helped to discover relationships between entities in texts used in the project.

DOI:10.15774/PPKE.ITK.2010.006

Figure 5.9: *Aiding text analysis with coreference resolution and possessor identification. The above structure represents syntactic and semantic relationships in the following text segment: "*A magyarok szinte minden csatában győztek. Harci sikereiket az erős törzsszövetségnek és könnyűlovas harci taktikájuknak köszönhették.*"*

DOI:10.15774/PPKE.ITK.2010.006

Chapter 6

APPENDIX

A1. Extracting Semantic Information from EKSz Definitions

The algorithm for extracting semantic relations from each EKSz definition is composed of 4 main steps:

1. Pre-processing: omitting non-processable parts of definitions; processing synonyms-only definitions.
2. Extraction of genus (hypernym) using patterns.
3. Extraction of holonym or meronym when the genus was one of special words.
4. Extraction of coordinated target words.

The details for each step, formulated as functions are described below in Python-like pseudo-code. The input for each function is an EKSz definition *d*, which is a list of tokens that have properties *stem* (base form of word), *surface* (original form of word), *pos* (part-of-speech of word), and morphological features such as *case*, *number* etc.) Each function extends global set *output* which is composed of <relation *target,* relation *type,* word *position* in the definition> triplets.

```
function preprocess(d):
    if d is empty or d.text==”Rövidítésként.”:
        return
    if i exists such that d[i]==“;”:
        for all j such that: j>i and d[j].pos==”noun”
        and d[j].case==”nominative”:
            output.add( d[j].stem, “synonym”, j)
        delete all words inside d starting from position i+1
    if i exists such that d[i]==“pl.” and for all w words after i in d: w.pos==”noun”:
        delete all words inside d starting from position i
    return
```

DOI:10.15774/PPKE.ITK.2010.006

```
function extract_genus(d):
    if d consists of only 1 word:
        if d[0].pos==”noun”:
            output.add( d[0].stem, “synonym”, 0)
        else:
            output.add( d[0].surface, “synonym”, 0)
    else:
        if d[0].pos==”noun” and d[0].case==”nominative” and d[1].surf=”:”:
            output.add( d[0].stem, “hypernym”, 0)
            return
        if i exists such that d[i].pos==”noun” and d[i+1].surface==”,”
        and d[i+2].surface is one of
        {“aki”,”ami”,”ahol”,”amikor”,”ahova”,”amely”,”ahogy”,”ahonnan”}:
            output.add( d[i].stem, “hypernym”, i)
            return
        if d[0].stem==”aki”:
            output.add( “person”, “hypernym”, -1)
            return
        if i exists such that (d[i].stem==”az” and d[i+1].surface==”,”
        and d[i+2].stem==”aki”) or (d[i].surface==”:” and d[i+1].stem==”aki”):
            output.add( “person”, “hypernym”)
            return
        for i in range(d.length-1, 0, -1):
            if d[i].pos==”noun”:
                output.add( d[i].stem, “hypernym”)
                return
```

DOI:10.15774/PPKE.ITK.2010.006

```
function extract_holo-mero(d):
    if output[output.length-1].target==”rész”:
        for i in range(0, d.length, 1):
            if d[i].pos==”noun” and d[i].case==”dative”:
                output.add( d[i].stem, “part-holonym”)
                return
        for i in range(0, d.length, 1):
            if d[i].pos==”noun” and d[i].stem!=”rész”:
                output.add( d[i].stem, “part-holonym”)
                return
        return
    if output[output.length-1].target==”összesség”:
        for i in range(0, d.length, 1):
            if d[i].pos==”noun” and d[i].case==”dative”:
                output.add( d[i].stem, “member-meronym”)
                return
        for i in range(0, d.length, 1):
            if d[i].pos==”noun” and d[i].number==”plural”
            and d[i].stem!=”összesség”:
                output.add( d[i].stem, “member-meronym”)
                return
        return
    return

function extract_coordinated(d):
    c = d[output[output.length-1].position].case
    for i in range(output[output.length-1].position-1, 1, -1):
        if d[i].surface is one of {“,”, ”illetve”} and d[i-1].pos==”noun”
        and d[i-1].case==c:
            output.add( d[i-1].stem, output[output.length-1].type, i-1)
```

DOI:10.15774/PPKE.ITK.2010.006

A2. Distribution of Polysemy in the American National Corpus

# Senses	# Types	# Tokens	(Word)
1	19,550	655,807	
2	8,581	606,731	
3	3,727	515,670	
4	1,785	525,767	
5	998	392,830	
6	530	250,420	
7	331	241,279	
8	183	169,422	
9	148	143,750	
10	110	153,618	
11	81	216,391	
12	48	67,254	
13	44	504,566	
14	16	32,210	
15	24	38,054	
16	16	27,665	
17	11	13,974	
18	6	9,409	
19	3	111,518	
20	3	3,207	
21	3	2,385	
22	6	22,513	
23	1	658	(beat.v)
24	7	47,224	
25	2	1,770	
26	4	6,880	
27	2	9,929	
28	2	8,657	
29	3	5,578	
30	1	26,280	(go.v)
32	2	3,283	
35	2	5,205	
36	2	36,005	
39	1	1,661	(carry.v)
41	2	6,967	
42	1	15,400	(take.v)
44	1	8,685	(give.v)
49	1	20,419	(make.v)
59	1	2,000	(break.v)

DOI:10.15774/PPKE.ITK.2010.006

A3. Hungarian Equivalents of *state* in the Hunglish Corpus

Stem	Frequency
állam	1,296
állapot	648
ország	169
állami	162
helyzet	58
állapotú	34
állás	21
izgalom	12
rend	11
fény	9
körülmény	9
osztály	6
dísz	5
pompa	5
rang	5
pompás	4
országbeli	3
aggodalom	2
nyugtalanság	2
országos	2
állású	2
díszes	1
fényes	1
fényű	1
helyzetű	1
méltóság	1
országú	1
rangú	1
rendes	1

DOI:10.15774/PPKE.ITK.2010.006

Chapter 7

REFERENCES

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

DOI:10.15774/PPKE.ITK.2010.006

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

DOI:10.15774/PPKE.ITK.2010.006

[24] **Miháltz 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 Miháltz**, 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, **Miháltz Márton**, Pléh Csaba, 2004: Fonológiai hurok és morfológia komplexitás. Magyar Pszichológiai Társaság Biennális Nagygyűlése, Debrecen, 2004.

[27] Ivády R. E., **Miháltz 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.

Works Cited

[28] Gómez-Pérez, A. – Fernández-López, M. –Corcho, O. 2006. Ontological Engineering. London: Springer-Verlag.

[29] Studer R, Benjamins VR, Fensel D (1998): Knowledge Engineering: Principles and Methods. In IEEE Transactions on Data and Knowledge Engineering 25(1-2):161-197.

[30] 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)

[31] 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)

[32] Rigau, G., H. Rodriguez and E. Agirre, 1998. Building Accurate Semantic Taxonomies from Monolingual MRDs. In Proceedings of COLING-ACL '98. Montréal, Canada.

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

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

[35] 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.

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

[37] Barnbrook, Geoff: Defining Language: A local grammar of definition sentences. Studies inCorpus Linguistics. Amsterdam: John Benjamins (2002).

[38] Prószéky, Gábor: Humor: a Morphological System for Corpus Analysis. Language Resources and Language Technology, Tihany (1996) 149–158

[39] 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)

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

DOI:10.15774/PPKE.ITK.2010.006

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

[41] Vossen, P. (eds): EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht (1998)

[42] Vossen, P. (ed.): EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document Deliverable D032D033/2D014, (1999).

[43] 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)

[44] Horak, A., P. Smrz: New Features of Wordnet Editor VisDic. In Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1-2) (2004)

[45] Smrz, P.: Quality Control and Checking for Wordnets Development: A Case Study of BalkaNet. In Romanian Journal of Information Science and Technology Special Issue (volume 7, No. 1-2) (2004)

[46] Niles, I., Pease, A. (2001): Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001.

[47] Niles, I. Pease, A. (2003): Linking Lexicons and Ontologies : Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE ’03), Las Vegas, Nevada, June 23-26, 2003.

[48] Váradi, T.: The Hungarian National Corpus. In Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas, pp 385-389 (2002)

[49] Kuti Judit, Vajda Péter, Varasdi Károly (2005): Javaslat a magyar igei WordNet kialakítására. In III. Magyar Számítógépes Nyelvészeti Konferencia Kiadványa, pp. 79-87.

[50] Kuti Judit, Varasdi Károly, Cziczelszky Judit, Gyarmati Ágnes, Nagy Anikó, Tóth Marianna, Vajda Péter (2006): Igei wordnet és igei eseményszerkezet ábrázolása. In IV. Magyar Számítógépes Nyelvészeti Konferencia Kiadványa, pp. 97-108.

[51] Gyarmati Ágnes, Almási Attila, Szauter Dóra (2006): A melléknevek beillesztése a Magyar WordNetbe. In IV. Magyar Számítógépes Nyelvészeti Konferencia Kiadványa, pp. 117-128.

[52] Judit Kuti, Károly Varasdi, Ágnes Gyarmati, Péter Vajda (2008): Language Independent and Language Dependent Innovations in the Hungarian WordNet. In *Proceedings of The Fourth Global WordNet Conference*, Szeged, Hungary (2008), pp. 254–268.

[53] Copestake, A., T. Briscoe, P. Vossen, A. Ageno, I. Castellon, F. Ribas, G. Rigau, H. Rodriguez, A. Samiotou, 1994. Acquisition of Lexical Translation Relations from MRDs. In Journal of Machine Translation, 3.

[54] Copestake, A., 1990. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. In Proceedings of the First International Workshop on Inheritance in Natural Language Processing.

DOI:10.15774/PPKE.ITK.2010.006

[55] Longman Contemporary Dictionary of English. Longman, London, 1978.

[56] 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.

[57] B. Magnini and G. Cavaglia: Integrating subject field codes into WordNet. In Proceedings of LREC-2000, Athens, Greece, 2000.

[58] Nagy, D., 2001. Computer Aided Methods for Lexical Database Compilation (Hungarian Nominal WordNet). Master's Thesis, Budapest University of Technology and Economics.

[59] Chris Manning, Hinrich Schütze: Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

[60] A. Ageno, I. Castellón, F. Ribas, G. Rigau, H. Rodriguez, A. Samiotou: TGE: Tlink Generation Environment. In Proceedings of the 15th International Conference on Computational Linguistics (Coling'94), Kyoto, Japan, 1994.

[61] K. Knight, S. Luk: Building a large-scale knowledge base for machine translation. In Proceedings of the American Association for Artificial Intelligence, 1994.

[62] A. Okumura, E. Hovy: Building Japanese-English Dictionary based on Ontology for Machine Translation. In Proceedings of Arpa Conference on Human Language Technology, Princeton,1994.

[63] G. Rigau, H. Rodriguez, J. Turmo: Automatically extracting Translation Links using a wide coverage semantic taxonomy. In Proceedings of The Fifteenth International Conference AI'95, Language Engineering '95, Montpellier, France, 1995.

[64] G. Rigau, E. Agirre: Disambiguating Bilingual Nominal Entries against WordNet. In Proceedings of The Computational Lexicon Workshop, Seventh European Summer School on Logic, Language and Information, pp. 71-82, Barcelona, Spain, 1995 .

[65] E. Agirre, X. Arregi, X. Artola, A. Díaz de Illarraza, K. Sarasola: Conceptual Distance and Automatic Spelling Correction. In Proceedings of the Workshop on Computational Linguistics for Speech and Handwriting Recognition, Leeds, UK, 1994.

[66] A. Gangemi, R. Navigli, P. Velardi. The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet, In Proc. of International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE 2003), Catania, Sicily (Italy), 2003, pp. 820-838.

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

[68] Kálmán László, Trón Viktor és Varasdi Károly (szerk.) (2002). Lexikalista elméletek a nyelvészetben. Tinta Könykiadó, Budapest.

[69] Ravin, Yael, Claudia Leacock (2000). Polysemy. Theoretical and Computational Approaches. Oxford University Press, Oxford.

[70] Pustejovsky, James (1995). The Generative Lexicon. MIT Press, Cambridge, MA.

[71] Ide, N., Suderman, K. (2004). The American National Corpus First Release. Proceedings of the Fourth Language Resources and Evaluation Conference (LREC), Lisbon, 1681-84.

DOI:10.15774/PPKE.ITK.2010.006

[72] Jurafsky, Daniel, James H. Martin (2000): Speech and Natural Language Processing. Prentice Hall, New Jersey.

[73] Macmillan English Dictionary. Heinemann Secondary Education, 2008.

[74] Yarowsky, David (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In Proceedings of COLING 14.

[75] Yarowsky, David (1994). Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In Proceedings of ACL-94, Las Cruces, NM

[76] Agirre, Eneko, German Rigau (2000). Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. In Computer And The Humanities, 34.

[77] Edmonds, Philip (2002). Introduction to Senseval. In ELRA Newsletter, October 2002.

[78] Leacock, Claudia, George A. Miller, Martin Chodorow (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. In Computational Linguistics, special issue on Word Sense Disambiguation.

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

[80] R. Mihalcea, T. Chklovski, 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.

[81] R. Mihalcea, T. Chklovski, T. and A. Kilgarriff, The Senseval-3 English Lexical Sample Task. Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems

[82] I. H. Witten, E. Frank, Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco, 2000.

[83] Richard Duda, Peter Hart, David G. Stork (2000). Pattern Classification (2nd Edition). Wiley, New York.

[84] Alpaydin, Ethem (2004): Introduction to Machine Learning. The MIT Press, Cambridge, MA

[85] John, H. G., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo (1995)

[86] Vancsa, L.: A „BLEU" automatikus kiértékelési eljárás alkalmazása angol-magyar fordítóprogram gyakori, folyamatos minősítésére. Magyar Számítógépes Nyelvészet Konfeencia, Szeged (2003)

[87] Miháltz Márton: Szemantikai hasonlóság és számítógépes jelentés-egyértelműsítés: a Random Indexing reprezentációs módszer vizsgálata (Semantic Similarity and Word Sense Disambiguation: an Examination of the Random Indexing Representation Method). Master's Thesis, University of Szeged, 2003

[88] Sahlgren. M. (2001). Vector-Based Semantic Analysis: Representing Word Meanings Based on Random Labels. Semantic Knowledge Acquisition and Categorisation Workshop. ESSLLI '01. Helsinki. Finland

DOI:10.15774/PPKE.ITK.2010.006

[89] Lund, K., Burgess, C., Atchley, R. (1995). Semantic and associative priming in high dimensional semantic space. In Proceedings of the 17th Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

[90] Landauer. T. K., Dumais. S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisiton. Induction and Representation of Knowledge. Psychological Review. 104(2). 210—240.

[91] Dagan, Ido, Alan Itai: Word sense disambiguation using a second language monolingual corpus. In Computational Linguistics, 20:563-596.

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

[93] 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.

[94] 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

[95] 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.

[96] Brennan, Susan E., Marilyn W. Friedman, Carl J. Pollard. A centering approach to pronouns. In Proceedings of the 25th Meeting of the Association for Computational Linguistics (1987), pp. 155-162.

[97] Csendes D., Alexin Z., Csirik J., Kocsor A.: A Szeged Korpusz és Treebank verzióinak története. III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005) kiadványa, Szeged, december 8-9., pp. 409-412 (2005)

[98] Grosz, Barbara, Joshi, Aravind, Weinstein, Scott: Centering: A framework for modelling the local coherence of discourse. Computational Linguistics, Volume 21, Number 2: 203-226 (1995)

[99] Hobbs, Jerry: Resolving pronoun references, in Readings in Natural Language Processing, Grasz, Jones and Webber, eds., Morgan Kaufman Publishers, Inc. Los Altos, California, USA (1977): 339 - 352

[100] Kenesei István: Az alárendelt mondatok szerkezete. In: Kiefer Ferenc (szerk.): Strukturális Magyar Nyelvtan, I. kötet, Mondattan. Akadémiai Kiadó, Budapest (1992), pp. 529–715.

[101] Chomsky, Noam: Lectures on Government and Binding. Dordrecht: Foris Publications (1981).

[102] Lappin, Shalom, Leass, Herbert, 1994, An algorithm for pronominal anaphora resolution, Computational Linguistics, Volume 20, Number 4: 535-562

[103] Niyu Ge, John Hale, Eugene Charniak: A statistical approach to anaphora resolution. In Proceedings of the Sixth Workshop on Very Large Corpora (1998).

[104] Leacock, C., M. Chodorow: Combining Local Context and WordNet Similarity for Word Sense Identification. In C. Fellbaum (ed.): WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA (1998), pp. 265–285

DOI:10.15774/PPKE.ITK.2010.006

[105] Lejtovicz Katalin, Kardkovács Zsolt: Anaforafeloldás magyar nyelvű szövegekben. IV. Magyar Számítógépes Nyelvészet Konferencia, Szeged (2006), pp. 362–364.

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

[107] Ng, Vincent and Cardie, Claire. Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution in Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002), 2002.

[108] 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), pp. 157–164.

[109] Soon, Ng, Lim: A Machine Learning Approach to Coreference Resolution of Noun Phrases. In Computational Linguistics, Volume 27, Number 4, 2001.

[110] Ralph Grishman, Beth Sundheim: Message Understanding Conference - 6: A Brief History. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING), Kopenhagen, 1996, pp. 466–471.

[111] Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, Ralph Weischedel: The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In Proceeding of the LREC 2004 Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal.

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

[113] Pléh Csaba: Mondatközi viszonyok feldolgozása: az anafora megértése a magyarban. In: Pléh Csaba: Mondatmegértés a magyar nyelvben. Osiris Kiadó, Budapest (1998), pp. 164–195.

[114] Tejaswini, Deoskar: Techniques for Anaphora Resolution: A Survey. Cornell University (2004). http://www.cs.cornell.edu/courses/cs674/2005sp/ projects/tejaswini-deoskar.doc.

[115] Tetreault, Joel R.: A Corpus Based Evaluation of Centering and Pronoun Resolution. In Computational Linguistics, Volume 27, Number 4, 2001.

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

[117] Uryupina, Olga: Coreference Resolution with and without Linguistic Knowledge. In Proceedings of the 6th International Conference on Language Resources and Evaluation (2006).

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

[119] 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.

[120] Vincze Orsolya, Gábor Kata, László János: Technológiai fejlesztések a NOOJ pszichológiai alkalmazásában. In VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2009, pp. 285–294.

[121] Ehmann Bea, Balázs László, Fülöp Éva, Hargitai Rita, László János: A NooJ alapú narratív pszichológiai tartalomelemzés alkalmazása pszichológiai

DOI:10.15774/PPKE.ITK.2010.006

állapotváltozások monitorozására űranalóg szimulációs kísérletben. In VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2009, pp. 295–304.