MŰEGYETEM 1782

Approaches to Hungarian Named Entity Recognition

PhD Thesis Booklet

Eszter Simon

A Thesis submitted for the degree of Doctor of Philosophy
PhD School in Psychology – Cognitive Science
Budapest University of Technology and Economics

Budapest, 2013

Supervisor:
András Kornai

Introduction

Computational Linguistics (CL) is an interdisciplinary field of computer science and linguistics concerned with the computational aspects of human language faculty. It belongs to the cognitive sciences and overlaps with the field of artificial intelligence, a branch of computer science aiming at computational models of human cognition. The theoretical aim of CL is to build formal theories and models about the linguistic knowledge that a human needs for generating and understanding language. However, CL has an applied component as well, which is often called Human Language Technology (HLT), and is used to develop software systems designed to process or produce different forms of human language.

Information Extraction (IE) is one of the main subtasks of CL, aiming at automatically extracting structured information from unstructured or semi-structured machine-readable documents. It covers a wide range of subtasks from finding all the company names in a text to finding all the actors of an event, for example to know who killed whom, or who sold their shares to whom. Such capabilities are increasingly important for sifting through the enormous volumes of online text to find pieces of relevant information the user wants.

Named Entity Recognition (NER), the task of automatic identification of selected types of Named Entities (NEs), is one of the most intensively studied tasks of IE. Presentations of language analysis typically begin by looking words up in a dictionary and identifying them as nouns, verbs, adjectives, etc. But most texts include lots of names, and if a system cannot find them in the dictionary, it cannot identify them, making it hard to produce a linguistic analysis of the text. Thus, NER is of key importance in many Natural Language Processing (NLP) tasks, such as Information Retrieval (IR) or Machine Translation (MT).

The Definition of Named Entities

The NER task, which is often called as Named Entity Recognition and Classification in the literature, has two substeps: first, locating the NEs in unstructured texts, and second, classifying them into pre-defined categories.

A key issue is how to define NEs. This issue interconnects with the issue of selection of classes and the annotation schemes applied in the field of NER. The NER task was introduced with the 6th Message Understanding Conference (MUC) in 1995 [Grishman and Sundheim, 1996], consisting of three subtasks: recognizing entity names, temporal and numerical expressions. Although there is a general agreement in the NER community about the inclusion of temporal expressions and some numerical expressions, the most studied types are names of persons, locations and organizations. The fourth type, called `Miscellaneous`, was introduced in the NER tasks of the Conference on Computational Natural Language Learning (CoNLL) in 2002 [Tjong Kim Sang, 2002] and 2003 [Tjong Kim Sang and De Meulder, 2003], and includes proper names falling outside the three classic types. Since then, MUC and CoNLL datasets and annotation schemes have been the major standards applied in the field of NER.

The annotation guidelines of these shared tasks are based on examples and counterexamples of what to annotate as a NE, rather than an exact, theoretically well-founded definition of NEs. The next description is from the MUC-7 Named Entity Task Definition [Chinchor, 1998]:

> "This subtask is limited to proper names, acronyms, and perhaps miscellaneous other unique identifiers, which are categorized via the TYPE attribute as follows:
> ORGANIZATION: named corporate, governmental, or other organizational entity
> PERSON: named person or family
> LOCATION: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)"

Besides this description negative examples (non-entities) are also provided. For annotating texts with NE labels, this kind of definition is not really helpful. In addition, the annotation guidelines mentioned above contain instructions only for English entities and non-entities. But in other languages, e.g. in Hungarian, there are concepts which would be annotated as NEs according to these guidelines, but they are not proper names, and thus are not considered as NEs. During the work of writing annotation guidelines for Hungarian [11] based on the widely used guidelines, their weak points became evident. From these experiences we conclude that a stronger definition is needed for annotation of NEs.

For this purpose, we studied Kripke's theory [Kripke, 2000] about the proper names as rigid designators. Kripke broke up with Frege's [Frege, 2000] and Russell's [Russell, 2000] description theory of proper names. In Chapter 2, we give an overview of the philosophic and linguistic background of the theory of proper names. After discussing the theoretical background, we try to map our findings to the NER task.

Thesis 1 *After investigating several theories of proper names, we can conclude that for getting a usable definition of NEs, the classic Aristotelian view on classification, which states that there must be a differentia specifica which allows something to be the member of a group, and excludes others, is not applicable. For our purposes, the prototype theory seems more plausible, where proper names form a continuum ranging from prototypical (person and place names) to non-prototypical categories (product and language names). Finally, the goal of the NER application will further restrict the range of linguistic units to be taken into account.*

The author's contribution. The author participated in the work which aimed at building a large, heterogeneous, manually NE annotated Hungarian corpus called the HunNer corpus. The author prepared the annotation scheme, and wrote the guidelines. For reasons outside the author's control, the HunNer corpus is still not entirely complete, but the guidelines have been used for other projects, e.g. for building the Criminal NE corpus[1]. These results are partly described in [10] and [11] and in the annotation guidelines, which is accessible on the web through the URL `http://krusovice.mokk.bme.hu/~eszter/utmutato.pdf`.

Handling Metonymic Named Entities

In metonymy, the name of one thing is substituted for that of another related to it [Lakoff and Johnson, 1980]. Besides common nouns, many proper names are widely used metonymically, as it can be seen in Examples 1 and 2. (Examples of metonymic NEs are not intuitively created by us, but they are accurate linguistic samples from the datasets provided by the organizers of SemEval-2007 metonymy resolution shared task [Markert and Nissim, 2007b], various articles and the web. In examples, throughout the dissertation, the relevant parts are italicized, or, if tags are important, they are in square brackets, with the tags in subscript.)

(1) Denise drank *the bottle*.

(2) Ted played *Bach*.

None of the two sentences is literally true. In Example 1, Denise did not drink the bottle made of plastic or glass, but the liquid in the bottle. In Example 2, Ted did not play the person whose name is Bach, but music composed by Bach [Fass, 1988].

[1] http://www.inf.u-szeged.hu/rgai/nlp?lang=en&page=corpus_ne

This type of reference shift is very systematic, in that it can occur with any person name, as long as the discourse participants are aware of that he/she is an artist, and they can associate an artwork with him/her. Linguistic studies ([Lakoff and Johnson, 1980, Fass, 1988]) postulate conventional metonymies that operate on semantic classes (here: person, location, and organization names). A few examples of such conventional metonymies follow (the standard name of metonymies are indicated with small capitals after the example sentences, in parentheses):

(3) *Spain* won its third straight major soccer title Sunday. (PLACE FOR PEOPLE)

(4) The broadcast covered *Vietnam*. (PLACE FOR EVENT)

(5) *Apple* announced new iPads and Mac computers. (ORGANIZATION FOR MEMBERS)

(6) It was the largest *Fiat* anyone had ever seen. (ORGANIZATION FOR PRODUCT)

Besides such regular shifts, metonymies can also be created on the fly: in Example 7, 'seat 19' refers to the person occupying seat 19. Markert and Nissim [Markert and Nissim, 2007a] call such occurrences unconventional metonymies.

(7) Ask *seat 19* whether he wants to swap.

Apart from being regular and productive, metonymic usage of NEs is frequent in natural language. State-of-the-art NER sytems usually do not distinguish between literal and metonymic usage of names, even though it would be helpful for most applications. Resolving metonymic usage of proper names would therefore directly benefit NER and indirectly all NLP tasks that require NER. The importance of resolving metonymies has been shown for a variety of NLP tasks, e.g. MT [Kamei and Wakao, 1992], question answering [Stallard, 1993], and anaphora resolution [Harabagiu, 1998, Markert and Hahn, 2002].

Distinguishing literal and metonymic usage, then identifying the intended referent can be seen as a classification task. Markert and Nissim [Markert and Nissim, 2002] postulate the metonymy resolution task as comparable to the Word Sense Disambiguation (WSD) task, so that metonymies can be recognized automatically with similar methods. On this assumption, Markert and Nissim [Markert and Nissim, 2007b] organized a shared task of the 2007 evaluation forum of the Semantic Evaluation series (SemEval-2007), which aimed at recognition and categorization literal, mixed, and metonymic usage of location and organization names. We built a maximum entropy based system [5], which achieved the best overall results in the competition. In Chapter 3, we give an overview of conventional and unconventional metonymies, and present the system description.

Thesis 2 *Since conceptual mappings between the related referents of metonymic words are not linked to particular linguistic forms, recognizing metonymic NEs is quite difficult. However, using some surface and syntactic information, and applying several semantic generalization methods lead to improvement in resolving metonymies. We present a supervised system, which achieved the best overall results in the SemEval-2007 metonymy resolution task. As our results show, the main dividing line does not lie between conventional and unconventional metonymies, rather between literal and metonymic usage.*

The author's contribution. Building the metonymy resolution system was a joint effort with the co-authors, namely Richárd Farkas, György Szarvas, and Dániel Varga. The author is responsible for investigating the related work, and providing the theoretical background. In addition, the author is responsible for some semantic generalization features, in particular for using Levin's verb classes and collecting the trigger words. The author also participated in feature engineering to find out whether each feature has the requisite discriminative power, the evaluation of results, and the drawing of conclusions. These findings are described in [5] and partly in [10].

Gold and Silver Standard Corpora for Named Entity Recognition

The supervised statistical approach requires a large amount of texts to boost performance quality. Such a large and structured set of texts is called a *corpus*. Corpora can be classified according to different criteria: they can be general or domain-specific, monolingual or multilingual, tagged or untagged. To be a gold standard corpus, a dataset has to meet several requirements, for example to be exhaustive or aiming for representativeness; to be large enough for training and testing supervised systems on it; and to contain accurate linguistic annotation added by hand.

The gold standard corpora in the field of NER are highly domain-specific, containing mostly newswire, and are restricted in size. Researchers attempting to merge these datasets to get a bigger training corpus are faced with the problem of combining different tagsets and annotation schemes. Manually annotating large amounts of text with linguistic information is a time-consuming, highly skilled and delicate job, but large, accurately annotated corpora are essential for building robust supervised machine learning NER systems. Therefore, reducing the annotation cost is a key challenge.

There are more ways to reach this goal. One approach is to use semi-supervised or unsupervised methods, which do not require large amount of labelled data. Another approach is to generate the resources automatically, or at least applying NLP tools that are accurate enough to allow automatic annotation. Yet another approach is to use collaborative annotation and/or collaboratively constructed resources, such as Wikipedia or DBpedia. Here we present a method which combines these approaches by automatically generating freely available NE tagged corpora from Wikipedia.

An automatically generated or silver standard corpus provides an alternative solution which is intended to serve as an approximation to a gold standard corpus. Such corpora are very useful for improving NER in several ways.

In Chapter 4, first, we give an overview of corpus building in general (Section 4.1). Section 4.2 introduces the gold standard corpora used in NER. In Section 4.3, we present our method to create automatically NE tagged English and Hungarian corpora built from Wikipedia.

Thesis 3 *We present a new method with which we can get closer to one of the main goals of current NER research, i.e. reducing the annotation labour of corpus building. We built automatically generated NE tagged corpora from Wikipedia for English and Hungarian. The one presented here is the first automatically NE annotated corpus for Hungarian which is freely available. As for English, there are no such automatically built corpora freely available, except for the Semantically Annotated Snapshot of the English Wikipedia, but their method cannot be applied for less resourced languages. As our method is mainly language-independent, it can be applied for other Wikipedia languages as well.*

Thesis 4 *We showed that automatically generated silver standard corpora are very useful for improving NER in several ways: (a) for less resourced languages, they can serve as training corpora in lieu of gold standard datasets; (b) they can serve as supplementary or independent training sets for domains differing from newswire; (c) they can be sources of huge entity lists, and (d) feature extraction.*

The author's contribution. The author participated in several corpus building projects.

Within the Hungarian Diachronic Generative Syntax project, the author is responsible for building a corpus which contains all text sources from the Old Hungarian period and a balanced selection from the Middle Hungarian period. The corpus is available via an online search engine: `http://rmk.nytud.hu/`. Related publications: [14, 13, 8].

Within the ABSTRACT project, which was a multi-site, EU funded research project that investigated how abstract linguistic concepts are learned and represented by the human mind, the author is responsible for building a corpus containing annotation of metaphorical expressions. Several methods were investigated for automatic identification of metaphors. The findings and the corpus itself are published in [2, 1, 4].

Within the HunNer corpus project, the author is responsible for preparing the annotation scheme and writing the guidelines. The corpus is described in [11].

Building the silver standard corpora for English and Hungarian was a joint effort with the co-author, Dávid Nemeskey. The author is responsible for investigating the related work, and providing the linguistic background. In addition, the author contributed to the construction of mapping between DBpedia ontology classes and gold standard tagsets, handling several problematic cases of NE labelling, and analysing and evaluating the error types of our method. Experiments for evaluating the newly generated datasets are the author's work. The method and the corpora themselves are published in [12, 6].

Approaches to Named Entity Recognition

The NER task, similarly to other NLP tasks, can be approached in two main ways: by applying hand-crafted rules, or by statistical machine learning techniques. This dichotomy is typical in the entire field of NLP, which dated back to end of the 1950s, when Chomsky published his influential review of Skinner's *Verbal Behavior* [Chomsky, 1959]. Finite state and probabilistic models, which were widely used before, had lost popularity in this period, and NLP split very cleanly into two paradigms, the theory-oriented or rule-based, and the data-driven or stochastic paradigms. In the early 1990s, the success of statistical methods in speech spread to other areas of NLP. This period has been called as the "return of empiricism". Due to the philosophical background of the paradigms they have also been called rationalist and empiricist approaches. Section 5.1 gives an overview of the philosophical background and the history of the two camps, until recent years, when the field comes together, and researchers try to build hybrid systems reaping the benefits of both approaches.

A rule-based NER application requires patterns which describe the internal structure of names and context-sensitive rules which give clues for classification. In Section 5.2, we give an enumeration of several kinds of internal end external evidence of NER, and describe a rule-based system using such patterns to extract NEs from Hungarian encyclopedic texts. We point to the disadvantages of rule-based systems, and conclude that applying machine learning algorithms is more useful for NER.

Statistical machine learning algorithms can be classified according to the type of input data they need. Unsupervised learning means that we do not have linguistically annotated data, thus the challenge is finding hidden structure in unlabelled data. Semi-supervised learning combines both labelled and unlabelled examples to generate an appropriate classifier. NLP tasks can also be solved by using labelled corpora and supervised learning methods that induce rules by discovering patterns in the manually annotated source text.

For building a supervised NER system, first we need a manually annotated gold standard corpus, which contains linguistic information. Typically, the algorithm itself learns its parameters from the corpus, and the evaluation of the system is through comparing its output to an other part of the corpus. So the corpus is divided into two parts: a training and a test set. When building a supervised learning system, a major step is feature extraction, that is collecting information from the data that can be relevant for the task. These features are the input of the learning algorithm that builds a model based on the regularities found in the data. After that the test set is tagged with the most probable labels, then they are compared to the gold standard labels. The evaluation means here to quantify the similarity between the two labellings. The whole process from training to evaluating a supervised NER system is described in details in Subsection 5.3.1.

For major languages, hundreds of papers were published on NER systems based on several supervised machine learning techniques. There are not too many language-dependent components of these, yet for Hungarian, we are aware only of one quantitative study of a NER system which is based on machine learning methods [Szarvas et al., 2006]. Our statistical NE tagger, the `hunner` system overperforms that system, achieving the best F-measure for Hungarian. In Subsection 5.3.2, we give a detailed system description.

Thesis 5 *The NER task, similarly to other NLP tasks, can be resolved by applying hand-crafted rules or machine learning techniques. We present a rule-based system developed for recognizing NEs in Hungarian encyclopedic texts and a supervised machine learning NER system which achieved the best performance for Hungarian. As our results show, applying statistical algorithms results in a more robust system and in higher performance on Hungarian NER.*

The author's contribution. The author contributed to several works concerned with rationalist and empiricist approaches to language acquisition as well as to NLP tasks.

The author participated in the 'Analogical generalisation processes in language acquisition' project, which had the aim of modelling the mechanisms of child language acquisition, specifically the process of learning argument structures from the input available to young children. We applied several statistical models for the automatic acquisition of subcategorization frames, and we concluded that data frequency and the size of the input corpus are important factors in both psycholinguistics and machine learning. These findings are published in [9, 15, 3].

Within the Hungarian Diachronic Generative Syntax project, the author participated in the development of a semi-automatic text normalization system applied for Old Hungarian texts. Most of the work on text normalization of historical documents is centered around a manually crafted set of correspondence rules. In contrast, we used the noisy channel paradigm to build an automatic normalization system. The human labour has been shifted to building training data for the transliteration model, for which the author is responsible. By the means of automatic normalization, the manual annotation process can be reduced to a selection of the right solution from the list of candidates provided by the system. The methodology and the results are presented in [8, 7].

The rule-based system developed for recognizing NEs in a Hungarian encyclopedia, Magyar Nagylexikon, remained unpublished, because it was treated with confidentiality. The system development was a joint effort with the colleagues, György Gyepesi, Lajos Incze, Zsolt Czinkos and Árpád Kiss. The author is responsible for creating the NE affixing rules and the transcribing rules for 20 languages, constructing and manually checking the gazetteer lists, and writing regular expression patterns providing information about the NEs' internal and external evidence.

The development of the original `hunner` system was a joint effort with the co-author, Dániel Varga. The author is responsible for feature engineering, data collection and evaluation. The author did not participate in the system's reimplementation, but is responsible for implementing and testing new features and collecting new gazetteers. The original system is published in [19, 20].

Feature Engineering

Features are descriptors or characteristic attributes of datapoints in a text. In token-based classification tasks of NLP, feature vectors are assigned to every token, where the feature vector contains one or more features. Generally, Boolean- or string-valued features are applied in NER. For example, if a word is capitalized, it gets an `iscap=1` feature. Feature vector representation is a kind of abstraction over text. The task of the machine learning algorithm is then to find regularities in this large amount of information that are relevant for NER.

Defining features for a supervised system is a manual work, similarly to coding patterns for a rule-based system. In the statistical methodology, however, the linguist does not tell anything about the power of the features, but it is found out from the corpus. The human cognition tends to realize only salient phenomena, thus declare features as important ones which are then found out not to be important based on corpus data, and vice versa. For this reason, the power of every feature has to be measured on real data before inclusion into the system. This is called feature engineering.

To measure the strength of features, we virtually built NER systems for Hungarian and English by adding new features to them one by one. For this purpose, we used the reimplemented version of the `hunner` system. In Chapter 6, we describe the features generally used for NER, and provide results about their power. We organize the features along the dimension of what kind of properties they provide: surface properties, digit patterns, morphological or syntactic information, or gazetteer list inclusion. As for the last kind of features, we also study the effects of gazetteer list size on the performance of NER systems.

Thesis 6 *We present a way of feature engineering in which the features most often used in NER are measured for getting the knowledge about their discriminative power. We conclude that for a supervised NER system the string-valued features related to the character makeup of words are the strongest features. Quite counterintuitively, features indicating casing information and sentence starting position do not improve the performance. Features based on external language processing tools such as morphological analysers and chunkers are also not necessary for finding NEs in texts.*

Thesis 7 *We compare the performance of a maximum entropy NER system under widely different entity list size conditions, ranging from a couple of hundred to several million entries, and conclude that for statistical NER systems entity list size has only a very moderate impact. If large entity lists are available, we can use them, but their lack does not cause invincible difficulties in the development of NER systems.*

The author's contribution. Defining most features presented in Chapter 6, measuring and evaluating them is the author's own work. Pre-processing of the Hungarian and English data and enriching them with linguistic information so serving as an appropriate input corpus for NER is also the author's own work. (Except for mapping the chunk tags of the Szeged Treebank to the Szeged NER corpus, which is the work of Attila Zséder and Judit Ács.) Collecting and designing the gazetteers used in the experiments is also the author's own work.

The author contributed to the development of the Hungarian `morphdb`, a lexical database and morphological grammar, which was used for the morphological analysis of the input corpora used for NER. It is published in [17, 16, 18].

The author contributed to the work of designing a system for recognizing metaphorical expressions by the means of different kinds of lists. The author is responsible for designing the lists, developing the software environment, and building the corpora on which the methods were evaluated. One of the important findings of this work is that using accurately compiled lists by hand is the most successful method for recognizing the relevant elements in a text. These findings are published in [2, 1, 4].

The author contributed to several works on feature engineering of state-of-the-art NER systems: to building a system for recognizing metonymic NEs in English texts (cf. Chapter 3) and to building the original `hunner` system (cf. Chapter 5). In both of them, the author is responsible for defining new features and measuring their strength. These findings are published in [5, 19, 20].

Publications Related to the Theses

[1] Anna Babarczy, Ildikó Bencze, István Fekete, and Eszter Simon. A metaforikus nyelvhasználat egy korpuszalapú elemzése. In Attila Tanács and Veronika Vincze, editors, *VII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 145–156, Szeged, 2010.

[2] Anna Babarczy, Ildikó Bencze, István Fekete, and Eszter Simon. The Automatic Identification of Conceptual Metaphors in Hungarian Texts: A Corpus-based Analysis. In *Proceedings of the LREC 2010 Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods*, pages 31–36, Malta, 2010.

[3] Anna Babarczy, András Serény, and Eszter Simon. Magyar igei vonzatkeretek gépi tanulása. In Attila Tanács, Dávid Szauter, and Veronika Vincze, editors, *VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2009)*, pages 333–342, Szeged, 2009. SZTE.

[4] Anna Babarczy and Eszter Simon. A fogalmi metaforák és a szövegstatisztika szerepe a metaforák felismerésében. In Gábor Prószéky and Tamás Váradi, editors, *Általános Nyelvészeti Tanulmányok XXIV. Nyelvtechnológiai kutatások*, pages 223–241. Akadémiai Kiadó, Budapest, 2012.

[5] Richárd Farkas, Eszter Simon, György Szarvas, and Dániel Varga. GYDER: maxent metonymy resolution. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 161–164, Prague, June 2007. Association for Computational Linguistics.

[6] Dávid Márk Nemeskey and Eszter Simon. Automatikus korpuszépítés tulajdonnév-felismerés céljára. In Attila Tanács and Veronika Vincze, editors, *IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013)*, pages 106–117, Szeged, 2012.

[7] Csaba Oravecz, Bálint Sass, and Eszter Simon. Gépi tanulási módszerek ómagyar kori szövegek normalizálására. In Attila Tanács,

Dávid Szauter, and Veronika Vincze, editors, *VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2009)*, pages 317–324, Szeged, 2009. SZTE.

[8] Csaba Oravecz, Bálint Sass, and Eszter Simon. Semi-automatic Normalization of Old Hungarian Codices. In *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, pages 55–60, Lisbon, Portugal, 2010. Faculty of Science, University of Lisbon.

[9] András Serény, Eszter Simon, and Anna Babarczy. Automatic acquisition of Hungarian subcategorization frames. In *Proceedings of the 9th International Symposium of Hungarian Researchers on Computational Intelligence*, 2009.

[10] Eszter Simon. Nyelvészeti problémák a tulajdonnév-felismerés területén. In Balázs Sinkovics, editor, *LingDok 7. Nyelvész-doktoranduszok dolgozatai*, pages 181–196. Szegedi Tudományegyetem Nyelvtudományi Doktori Iskola, Szeged, 2008.

[11] Eszter Simon, Richárd Farkas, Péter Halácsy, Bálint Sass, György Szarvas, and Dániel Varga. A HunNER korpusz. In Zoltán Alexin and Dóra Csendes, editors, *IV. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, 2006.

[12] Eszter Simon and Dávid Márk Nemeskey. Automatically generated NE tagged corpora for English and Hungarian. In *Proceedings of the 4th Named Entity Workshop (NEWS) 2012*, pages 38–46, Jeju, Korea, July 2012. Association for Computational Linguistics.

[13] Eszter Simon and Bálint Sass. Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből. In Gábor Prószéky and Tamás Váradi, editors, *Általános Nyelvészeti Tanulmányok XXIV. Nyelvtechnológiai kutatások*, pages 243–264. Akadémiai Kiadó, Budapest, 2012.

[14] Eszter Simon, Bálint Sass, and Iván Mittelholcz. Korpuszépítés ómagyar kódexekből. In Attila Tanács and Veronika Vincze, editors, *VIII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 81–89, Szeged, 2011. SZTE.

[15] Eszter Simon, András Serény, and Anna Babarczy. Automatic Acquisition of Hungarian Subcategorization Frames. In *Proceedings of*

the LREC 2010 Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods, pages 7–11, Malta, 2010.

[16] Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Eszter Simon, and Péter Vajda. morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005)*, pages 169–179, Szeged, December 2005.

[17] Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1670–1673, 2006.

[18] Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon. Morphdb.hu: Hungarian lexical database and morphological grammar. In Katalin S. Nagy and István Szakadát, editors, *Média és társadalom. Válogatás a Szociológia és Kommunikáció Tanszék Média Oktató és Kutató Központ munkatársainak legújabb munkáiból*, pages 283–290. Műegyetemi Kiadó, 2006.

[19] Dániel Varga and Eszter Simon. Magyar nyelvű tulajdonnév-felismerés maximum entrópia módszerrel. In Zoltán Alexin and Dóra Csendes, editors, *IV. Magyar Számítógépes Nyelvészeti Konferencia*, pages 32–38, Szeged, 2006.

[20] Dániel Varga and Eszter Simon. Hungarian named entity recognition with a maximum entropy approach. *Acta Cybernetica*, 18:293–301, 2007.

References

[Chinchor, 1998] Chinchor, N. (1998). MUC-7 Named Entity Task Definition Version 3.5. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

[Chomsky, 1959] Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behavior. *Language*, 35(1):26–58.

[Fass, 1988] Fass, D. (1988). Metonymy and Metaphor: What's the Difference? In *Proceedings of the 12th Conference on Computational linguistics – Volume 1*, COLING '88, pages 177–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Frege, 2000] Frege, G. (2000). Ueber Sinn und Bedeutung (On Sense and Reference). In Stainton, R. J., editor, *Perspectives in the Philosophy of Language – A concise anthology*. Broadview Press.

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. (1996). Message Understanding Conference – 6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 466–471, Kopenhagen.

[Harabagiu, 1998] Harabagiu, S. (1998). Deriving Metonymic Coercions from WordNet. In *Workshop on the Usage of WordNet in Natural Language Processing Systems, COLING ACL*, pages 142–148.

[Kamei and Wakao, 1992] Kamei, S. and Wakao, T. (1992). Metonymy: Reassessment, survey of acceptability and its treatment in machine translation systems. In *Proceedings of ACL*, pages 309–311.

[Kripke, 2000] Kripke, S. (2000). Naming and Necessity. In Stainton, R. J., editor, *Perspectives in the Philosophy of Language – A concise anthology*. Broadview Press.

[Lakoff and Johnson, 1980] Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. Chicago University Press, London.

[Markert and Hahn, 2002] Markert, K. and Hahn, U. (2002). Understanding Metonymies in Discourse. *Artificial Intelligence*, 135(1/2):145–198.

[Markert and Nissim, 2002] Markert, K. and Nissim, M. (2002). Metonymy Resolution as a Classification Task. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 204–213, Philadelphia. Association for Computational Linguistics.

[Markert and Nissim, 2007a] Markert, K. and Nissim, M. (2007a). Metonymic Proper Names: A Corpus-based Account. In Stefanowitsch, A. and Gries, S. T., editors, *Corpus-Based Approaches to Metaphor and Metonymy*, pages 152–174. Mouton de Gruyter.

[Markert and Nissim, 2007b] Markert, K. and Nissim, M. (2007b). SemEval-2007 Task 08: Metonymy Resolution at SemEval-2007. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 36–41, Prague. Association for Computational Linguistics.

[Russell, 2000] Russell, B. (2000). Descriptions. In Stainton, R. J., editor, *Perspectives in the Philosophy of Language – A Concise Anthology*. Broadview Press.

[Stallard, 1993] Stallard, D. (1993). Two kinds of metonymy. In *Proceedings of ACL*, pages 87–94.

[Szarvas et al., 2006] Szarvas, Gy., Farkas, R., and Kocsor, A. (2006). A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In *Proceedings of Discovery Science 2006*, pages 267–278. Springer Verlag.

[Tjong Kim Sang, 2002] Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In Roth, D. and van den Bosch, A., editors, *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

[Tjong Kim Sang and De Meulder, 2003] Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Daelemans, W. and Osborne, M., editors, *Proceedings of CoNLL-2003*. Edmonton, Canada.