Natural Language Processing of Large Parallel Corpora

Dániel Varga

Theses of Ph.D. Dissertation

Supervisors:
András Kornai D.Sc., András Lukács Ph.D.

Eötvös Loránd University
Faculty of Informatics
Department of Information Systems

Ph.D. School of Computer Science
András Benczúr D.Sc.

Foundations and Methods of Informatics Ph.D Program
János Demetrovics D.Sc. MHAS

Budapest, 2012

Introduction

The thesis presents novel research in morphological disambiguation, named entity recognition, noun phrase chunking, sentence alignment, and bitext querying. With the exception of the last task, which is a recent "Web 2.0" development, all other tasks are well-known problems in Natural Language Processing (NLP). Morphology is discussed in Chapter 2 of (Mitkov 2005), named entity recognition in Chapter 22.1 of (Jurafsky & Martin 2000) noun phrase chunking in Chapter 10.6.2 and 12.1.2 of (Manning & Schütze 1999), and sentence alignment in Chapter 25.6 of (Jurafsky & Martin 2000), all standard textbooks in NLP. Language Resources (LRs) are the fundamental language-specific data-sets of NLP, ranging from the largely unprocessed such as raw (unfiltered) corpora to the highly processed, such as machine readable dictionaries. The results are presented in three major categories, reflected in the structure of each section:

(1) NLP software components

(2) LRs created by the software – these are typically (but not always) Hungarian LRs

(3) Any further research uses that the software and the LRs have been put to

Chapter 1 of the thesis introduces the subject, Chapters 2 and 3 describe the mathematical and linguistic (morphological) background. These contain no new results.

For ease of cross-reference, the numbering of claims in this summary follows the numbering of chapters in the thesis, starting with 4, and using the three major categories listed above. Since the work presented here is generally highly collaborative, we add a discussion of provenance under (4) for each chapter.

All software and LRs discussed in the thesis are free and open source, available at the following links.

`webcorpus pipeline`	`https://github.com/zseder/webcorpus`
`hunner`, `hunchunk`	`https://github.com/recski/HunTag.git`
`hunalign`	`http://mokk.bme.hu/resources/hunalign`
`Hunglish bitext query`	`http://code.google.com/p/hunglish-webapp`
Szószablya Webcorpus	`http://mokk.bme.hu/resources/webcorpus`
Frequency Dictionaries	`http://hlt.sztaki.hu/resources/webcorpora.html`
Hunglish Corpus	`http://mokk.bme.hu/resources/hunglishcorpus`
JRC-Acquis Corpus	`http://langtech.jrc.it/JRC-Acquis.html`
Szószablya dictionary service	`http://szotar.mokk.bme.hu/szoszablya`
Hunglish bitext query service	`http://hunglish.hu`

4 Morphological Disambiguation

In any morphologically complex language such as Hungarian, morphological analysis will often return more than one possible analysis for a given word, compare *ment* 'is saving' and *ment*

'went'. The morphological disambiguation task is to decide, based on context, which of these alternatives is the correct one.

Claim (4.1) We designed and evaluated a system solving the morphological disambiguation task for Hungarian.

Our results show that purely statistical systems can be effectively combined with a rule-based morphological analyzer. For Hungarian, this was first demonstrated by (Oravecz & Dienes 2002). Our system has the advantage that the treatment of out of vocabulary (OOV) words is more robust.

Claim (4.2) We present our morphologically disambiguated Hungarian webcorpus and frequency dictionary. The corpus gathered is based on 18 million pages crawled. Its best quality stratum consists of 589m words harvested from 1.22m pages, making it the largest extant corpus of Hungarian to this day. As a comparison, the manually collected Hungarian National Corpus (Váradi 2002) has 188m words, the manually annotated Szeged Corpus (Csendes et al. 2004) has 1.2m words.

(4.3) The morphologically disambiguated frequency dictionary is frequently used by Hungarian linguists (Magyar & Szentgyörgyi 2011), (Rácz & Szeredi 2009), (Szeredi 2009) and psycholinguists (Racsmány et al. 2012), (Lukács et al. 2007). In Subsection 4.9.2 of the thesis we describe an application of the Szószablya frequency dictionary in the field of psycholinguistics (Pléh et al. 2011).

Currently our research team is in the process of creating an updated version of the Szószablya Hungarian Webcorpus and frequency dictionary. As we document it in (Halácsy et al. 2008) and (Zséder et al. 2012), the data processing pipeline has been completely rewritten for this task, mainly to increase processing speed, but also with improved data quality in mind. We have already used our pipeline to create and publish webcorpora and (morphologically non-disambiguated) frequency dictionaries for fifteen European languages.

(4.4) The main results of the Chapter originally appeared in (Halácsy et al. 2005) (in Hungarian) and (Kornai et al. 2006). The author of the thesis was an equal collaborator with Péter Halácsy and András Kornai in the design and creation of the morphological disambiguator systems presented in (Halácsy et al. 2005). The creation of the morphologically disambiguated Szószablya webcorpus and frequency dictionary was the result of a larger collaboration (Kornai et al. 2006). The author contributed software tools to an ongoing project to build webcorpora and frequency dictionaries for medium and small density languages. This work is documented in (Halácsy et al. 2008) and (Zséder et al. 2012). The subsection on a psycholinguistics application is the result of a collaboration with Csaba Pléh and co-workers (Pléh et al. 2011), where the author's contribution was the construction of the entropy models, and the statistical analysis of the empirical data.

5 Named Entity Recognition

Since much of our knowledge about the world is anchored to places, persons, and organizations, finding and correctly identifying parts of text that refer to *named entities* is key to understanding natural language input. This task is called Named Entity Recognition (NER).

(NER has its own disambiguation problems, like detecting and classifying *metonymic* readings such as place-for-event *'Vietnam was a great national trauma'*, see Section 5.5 of the thesis).

Claim (5.1) We developed the `hunner` named entity recognition system for Hungarian. The system employs a simplified Maximum Entropy Markov Model, and relies on character n-gram based features extensively. The system achieves state-of-the-art performance, with an F-score above 95% on a standard task (Varga & Simon 2007).

(5.2) Employing the `hunner` system, (Solymosi 2007) created the automatically named entity annotated 73.8 million token Origo NER Corpus based on a collection of news items from the `origo.hu` news portal.

(5.3) In Section 5.6 we discuss an application of the Origo NER Corpus in a network analysis study.

(5.4) The `hunner` system was joint work with Eszter Simon and appeared in (Varga & Simon 2006) and (Varga & Simon 2007). The author of the thesis is responsible for the design of the machine learning architecture, the software implementation, and the evaluation framework. Feature engineering and data collection was shared between the two original authors. The software was later reimplemented in a joint work with Gábor Recski (Recski & Varga 2009). The result of this reimplementation is the so-called `huntag` tool that is capable of named entity recognition and various chunking tasks, depending on the resources it is provided with. We present (5.2) and (5.3) only for completeness, the author had no significant role in these.

6 Noun Phrase Chunking

High accuracy full syntactic analysis, assigning a parse tree or other structural description to a sentence, is beyond the current state of the art. Fortunately, for many applications it is often sufficient to detect NPs (noun phrases) in a sentence. NP chunking is also a stepping stone toward full syntactic analysis, in part because analysis below the NP level can be performed by standard tools such as context free grammars and in part because locating the NPs can help higher level analysis, especially filling the valency structure of the main verb. The first Hungarian NP chunker, a cascaded regular grammar, was created by (Váradi 2003). The paper reports an F-score of 58.78%.

Claim (6.1) We developed the `hunchunk` noun phrase chunker.

(Miháltz 2011) evaluated `hunchunk` together with two rule-based NP-chunkers: a rule-based NP-chunker system for Hungarian (Váradi & Gábor 2004), and the Hungarian syntactic parser used by the MetaMorpho machine translation system (Prószéky et al. 2004). As Table 1 shows, `hunchunk` achieves higher scores than the rule-based systems. We note that the methodology is heavily biased towards `hunchunk`, as the test corpus is based on a subset of the Szeged Corpus, albeit a subset disjoint from our training corpus.
(Hócza 2004) documents a statistical rule-learner NP chunker algorithm for Hungarian, and reports an F-score of 83% on a corpus similar to ours, using tenfold cross-validation. Although our corpora are not completely comparable, we also performed tenfold cross-validation on our corpus, achieving an F-score of 89.30% (precision 89.75%, recall 88.86%).

	Precision	Recall	F-score
`hunchunk`	78.67	84.99	81.71
`MetaMorpho`	54.39	61.52	57.73
`NooJ`	37.57	59.28	45.99

Table 1: Comparison with two Hungarian rule-based maximal NP-chunkers by (Mihaltz 2011)

(6.2) (Recski et al. 2010) presents a system aligning NPs for Hungarian-English. The system relies on `hunchunk` for the chunking task.

(6.4) The results of this Chapter are joint work with Gábor Recski, originally presented in (Recski & Varga 2009) and (Recski & Varga 2012). The author had no role in the results presented in (Recski et al. 2010).

7 Sentence Alignment

Modern methods of machine translation and automated dictionary- and thesaurus building take parallel texts (also called bitexts) as their input, which contain the same material in two languages. *Sentence alignment*, finding which sentence in the source language corresponds to which sentence in the target language is a task that is key to preparing the data for machine translation and related tasks. Standard systems performing sentence alignment include `BSA` (Moore 2002) and `GMA` (Melamed 1998).

Claim (7.1) We developed the `hunalign` system and its `partialAlign` preprocessor. These are distinguished from `BSA`, `GMA`, and similar alternatives by performing, in the critical range of interest, an order of magnitude faster, without sacrificing accuracy.

Instead of a 'full IBM translation model' `hunalign` relies on a simpler but effective dictionary-building method. This has two advantages. First, our translation similarity score is very fast to calculate. While speed is often considered a mundane issue, `hunalign` is significantly faster than other modern aligners, and the increase in speed can be leveraged in many ways during the building of a parallel corpus with tens of thousands of documents. Second, we can exploit a bilingual lexicon, if one is available, and tune it according to frequencies in the target corpus. The `partialAlign` preprocessor uses a fast and accurate splitting algorithm, making it unnecessary to run `hunalign` directly on datasets over 20,000 sentences where its memory consumption would become a limiting factor.

After the publication of our results, (Krynicki 2006) compared `BSA`, `GMA`, and `hunalign` on Polish-English text for several corpora and lemmatization settings, finding that `BSA` often outperformed `hunalign` and `GMA` in precision, but `hunalign` always outperformed the other two aligners in F-scores thanks to better recall. (F-score was measured on the set of one-to-one correspondences.)

(Abdul Rauf et al. 2012) compares the performance of sentence alignment systems by training statistical machine translation (SMT) systems on their output, and evaluating the performance of the resulting SMT systems according to three popular evaluation metrics.

For each of the three evaluations of the French-English SMT system, `hunalign`'s score was 'statistically tied for first place', that is, within one standard deviation of the best score.

Like all of our NLP tools, `hunalign` is free software. In particular, parallel corpora aligned with it can be made freely available for any purpose, including commercial use. This distinguishes it from some of its alternatives which prohibit such use.

Claim (7.2) The JRC-Acquis Corpus (Steinberger et al. 2006) is a large multilingual parallel text corpus built by an international team of language technologists. With 230 citations according to Google Scholar, JRC-Acquis is one of the most important resources of multilingual natural language processing. Its main use is probably as training data for statistical machine translation systems such as (Turchi et al. 2009), but it was used in experiments in many fields such as cross-language information retrieval (Talvensaari 2008), multilingual sentiment analysis (Bautin et al. 2008), and cross-language plagiarism detection (Potthast et al. 2011).

Because of its accuracy, speed, and permissive licensing, the `hunalign` tool is widely used by the multilingual natural language processing community, collecting 98 citations at the time of writing according to Google Scholar. `hunalign` was used to create four of the five largest public multilingual parallel corpora we are aware of: (i) The author was directly involved in the creation of the influential JRC-Acquis corpus (Steinberger et al. 2006) discussed above; (ii) OPUS (Tiedemann 2009); (iii) Parasol (Waldenfels 2011); and (iv) InterCorp (Rosen & Vavřín 2012). The exception is (v) EUROPARL (Koehn 2005) which was aligned with the earlier Vanilla aligner, but in subsequent work (Koehn 2010) himself recommends the use of `hunalign`.

The largest English-Hungarian parallel corpus, the Hunglish Corpus (115m words in 4.15m sentence pairs, see (Varga et al. 2005)), was also created using `hunalign`. In Hungary, this corpus has been used as training data for machine translation (Hócza & Kocsor 2006), and word-sense disambiguation (Mihăltz & Pohl 2006). Internationally, it was employed as part of the training data in the '2008 ACL Workshop on Statistical Machine Translation' (Callison-Burch et al. 2008) and '2009 EACL Workshop on Statistical Machine Translation' `http://www.statmt.org/wmt09/translation-task.html` shared tasks.

(7.3) `hunalign` was integrated into several software tools as a component. The most notable are probably `UPlug` and `LF Aligner`. `UPlug` (Tiedemann 2002) is a framework for parallel text processing. It was used to create the OPUS parallel corpus (Tiedemann 2009) relying on its `hunalign` plug-in for the task of sentence alignment. As main components of the `LF Aligner` tool, `hunalign` and `partialAlign` are used by thousands of professional translators worldwide to create translation memories.

(7.4) The design, implementation, and evaluations of `hunalign` and `partialAlign` are the work of the present author. The creation of the corpora was a collaborative effort in each case. In the case of the Hunglish Corpus (Varga et al. 2005), the majority of work was done by the author, from data cleaning to packaging. In the case of the JRC-Acquis Corpus (Steinberger et al. 2006), the author was responsible for adapting and applying the `hunalign` technology to the automatic sentence alignment of sentence-segmented data for all possible language pairs.

8 Bitext Query

Searchable bitext (parallelized bilingual text) can be an important tool for human translators for finding translations of phrases in context. By offering the possibility of user feedback and document uploading, the system becomes a 'crowdsourced' tool for the semi-manual collection and fully manual validation of parallel corpus data.

Claim (8.1) We developed the Hunglish bitext query system `HBQ`. The system consists of an offline (asynchronous) and an online (synchronous) component. The offline component takes a pair of documents as input, extracts a set of parallel sentence pairs (bisentences) from them, and adds this set to an index. The online component answers user search queries on this index of bisentences. The two systems are integrated into a web-based application that is capable of answering search requests even while simultaneously adding new user-supplied bitexts to its database. It is also augmented with the feature that users can vote on the correctness of bisentences, thus improving the quality of the parallel corpus. Downvoted bisentences immediately drop to a lower position on the result list. The service is currently provided for Hungarian-English, but the software is entirely language(pair)-independent.

Claim (8.2) The service, still in early beta, and without a polished user interface, already achieved significant web presence. It serves 150,000 search queries monthly to 10,000 visitors. Visitors reported 2,000 erroneous bisentences in the Hunglish Corpus, and uploaded 37 new document pairs yielding 17,000 new bisentences.

(8.4) The offline component of `HBQ` is the result of joint work with Attila Zséder (Recski et al. 2009). The architecture of the software is the design of the present author. The implementation of the framework is the work of Attila Zséder. Plugging the `hun⋆` tools and other language processing utilities into this framework was the work of the present author. The online component is joint work with Péter Gergő Barna, previously unpublished. The architecture of the software is the result of joint work, implementation is largely the work of Péter Gergő Barna. The system is a full reimplementation of our bitext query system originally created by Péter Halácsy (Halácsy et al. 2004). The main new features of the reimplementation are the capability to add bidocuments to a running service, duplicate filtering, and user voting on bisentence quality.

Publications in English

Farkas, R., Simon, E., Szarvas, G. & Varga, D. (2007), GYDER: Maxent metonymy resolution, *in* 'Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)', Association for Computational Linguistics, Prague, Czech Republic, pp. 161–164.

Halácsy, P., Kornai, A., Németh, P. & Varga, D. (2008), Parallel creation of gigaword corpora for medium density languages - an interim report, *in* 'Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)', European Language Resources Association (ELRA), Marrakech, Morocco.

Kornai, A., Halácsy, P., Nagy, V., Oravecz, C., Trón, V. & Varga, D. (2006), Web-based frequency dictionaries for medium density languages, *in* 'Proceedings of the EACL 2006 Workshop on Web as a Corpus'.

Pléh, C., Németh, K., Fazekas, J. & Varga, D. (2011), Entropy measures and predictive recognition as mirrored in gating and lexical decision over multimorphemic Hungarian noun forms, *in* 'QMMMD Workshop, University of California, San Diego (Jan. 15-16)'.

Recski, G. & Varga, D. (2009), 'A Hungarian NP-chunker', *The Odd Yearbook* .

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. & Varga, D. (2006), The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, *in* 'Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)', Genoa, Italy.

Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L. & Varga, D. (2005), Hunmorph: open source word analysis, *in* 'Proceedings of the ACL 2005 Workshop on Software'.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. & Nagy, V. (2005), Parallel corpora for medium density languages, *in* 'Proceedings of the Recent Advances in Natural Language Processing 2005 Conference', Borovets. Bulgaria, pp. 590–596.

Varga, D. & Simon, E. (2007), 'Hungarian named entity recognition with a maximum entropy approach', *Acta Cybern.* **18**(2), 293–301.

Zséder, A., Recski, G., Varga, D. & Kornai, A. (2012), Rapid creation of large-scale corpora and frequency dictionaries, *in* 'Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)', European Language Resources Association (ELRA), Istanbul, Turkey.

Publications in Hungarian

Farkas, R., Szeredi, D., Varga, D. & Vincze, V. (2010), MSD-KR harmonizáció a Szeged Treebank 2.5-ben, *in* 'VII. Magyar Számítógépes Nyelvészeti Konferencia', pp. 349–353.

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V. & Varga, D. (2004), Hunglish: nyílt statisztikai magyar–angol gépi nyersfordító, *in* 'II. Magyar Számítógépes Nyelvészeti Konferencia', Szegedi Tudományegyetem, pp. 81–84.

Halácsy, P., Kornai, A. & Varga, D. (2005), Morfológiai egyértelműsítés maximum entrópia módszerrel (morphological disambiguation with the maxent method), *in* 'Proc. 3rd Hungarian Computational Linguistics Conf.', Szegedi Tudományegyetem.

Rebrus, P., Kornai, A. & Varga, D. (2012), 'Egy általános célú morfológiai annotáció', *Általános Nyelvészeti Tanulmányok* . to appear.

Recski, G. & Varga, D. (2012), 'Magyar főnévi csoportok azonosítása', *Általános Nyelvészeti Tanulmányok* . to appear.

Recski, G., Varga, D., Zséder, A. & Kornai, A. (2009), Főnévi csoportok azonosítása magyar-angol párhuzamos korpuszban, *in* 'VI. Magyar Számítógépes Nyelvészeti Konferencia', Szegedi Tudományegyetem.

Varga, D. & Simon, E. (2006), Magyar nyelvű tulajdonnév-felismerés maximum entrópia módszerrel, *in* Z. Alexin & D. Csendes, eds, 'IV. Magyar Számítógépes Nyelvészeti Konferencia', Szegedi Tudományegyetem, Szeged, pp. 32–38.

References

Abdul Rauf, S., Fishel, M., Lambert, P., Noubours, S. & Sennrich, R. (2012), Extrinsic evaluation of sentence alignment systems, *in* 'LREC Workshop on Creating Cross-language Resources for Disconnected Languages and Styles (CREDISLAS)', Istanbul (Turkey).

Bautin, M., Vijayarenu, L. & Skiena, S. (2008), International sentiment analysis for news and blogs, *in* 'Proceedings of the International Conference on Weblogs and Social Media (ICWSM)'.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. (2008), Further meta-evaluation of machine translation, *in* 'Proceedings of the Third Workshop on Statistical Machine Translation', StatMT '08, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 70–106.

Csendes, D., Csirik, J. & Gyimóthy, T. (2004), The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus, *in* 'Text, Speech and Dialogue: 7th International Conference, TSD', pp. 41–47.

Hócza, A. (2004), 'Noun phrase recognition with tree patterns', *Acta Cybern.* **16**(4), 611–623.

Hócza, A. & Kocsor, A. (2006), Hungarian-English machine translation using genpar, *in* 'Proceedings of the 9th international conference on Text, Speech and Dialogue', TSD'06, Springer-Verlag, Berlin, Heidelberg, pp. 87–94.

Jurafsky, Daniel & Martin, James, (2000), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall; 1st edition.

Koehn, P. (2005), Europarl: A Parallel Corpus for Statistical Machine Translation, *in* 'Conference Proceedings: the tenth Machine Translation Summit', AAMT, AAMT, Phuket, Thailand, pp. 79–86.

Koehn, P. (2010), *Statistical Machine Translation*, Cambridge University Press.

Krynicki, G. (2006), Compilation, Annotation and Alignment of a Polish-English Parallel Corpus, PhD thesis, Poznan University.

Lukács, A., Pléh, C. & Racsmány, M. (2007), 'Spatial language in Williams syndrome: evidence for a special interaction?', *Journal of Child Language* **34(2):311-43.**

Magyar, L. & Szentgyörgyi, S. (2011), Vowel zero alternations in Hungarian nominal inflectional and derivational paradigms: An analogy-based statistical approach, *in* '4th Syntax, Phonology and Language Analysis Conference, Budapest'.

Manning, C. & Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, MA.

Melamed, I. D. (1998), Empirical methods for exploiting parallel texts, PhD thesis, University of Pennsylvania, Philadelphia, PA, USA. AAI9829948.

Miháltz, M. (2011), Magyar NP-felismerők összehasonlítása, *in* 'VIII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged', pp. 333–335.

Miháltz, M. & Pohl, G. (2006), Exploiting Parallel Corpora for Supervised Word-Sense Disambiguation in English-Hungarian Machine Translation, *in* 'Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)', Genoa, Italy.

Mitkov, R. (2005), *The Oxford Handbook Of Computational Linguistics*, Oxford handbook, Oxford University Press.

Moore, R. C. (2002), Fast and accurate sentence alignment of bilingual corpora, *in* 'Proc 5th AMTA Conf: Machine Translation: From Research to Real Users', Springer, Langhorne, PA, pp. 135–244.

Oravecz, Cs. & Dienes, P. (2002), Efficient stochastic part-of-speech tagging for Hungarian, *in* 'Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002)', pp. 710–717.

Potthast, M., Barrón-Cedeño, A., Stein, B. & Rosso, P. (2011), 'Cross-language plagiarism detection', *Lang. Resour. Eval.* **45**(1), 45–62.

Prószéky, G., Tihanyi, L. & Ugray, G. (2004), Moose: a robust high-performance parser and generator, *in* 'Proceedings of the 9th Workshop of the European Association for Machine Translation', La Valletta, Malta, p. 138–142.

Racsmány, M., Conway, M., Keresztes, A. & Krajcsi, A. (2012), 'Inhibition and interference in the think/no-think task', *Memory and Cognition* **40(2):168-76.**

Rácz, P. & Szeredi, D. (2009), Testing usage-based predictions on Hungarian vowel reduction, *in* '17th Manchester Phonology Meeting, Manchester, UK'.

Recski, G., Rung, A., Zséder, A. & Kornai, A. (2010), NP Alignment in Bilingual Corpora, *in* N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias, eds, 'Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)', European Language Resources Association (ELRA), Valletta, Malta.

Rosen, A. & Vavřín, M. (2012), Building a multilingual parallel corpus for human users, *in* N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis, eds, 'Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)', European Language Resources Association (ELRA), Istanbul, Turkey.

Solymosi, A. (2007), Tulajdonnév-felismerés, személynevek azonosítása magyar nyelvű szövegben, Master's thesis, Budapest University of Technology and Economics.

Szeredi, D. (2009), Functional phonological analysis of the Hungarian vowel system, Master's thesis, Eötvös Loránd University, Theoretical Linguistics.

Talvensaari, T. (2008), *Comparable Corpora in Cross-language Information Retrieval*, Julkaisusarja A, University of Tampere, Department of Computer Sciences.

Tiedemann, J. (2002), Uplug - a modular corpus tool for parallel corpora, *in* L. Borin, ed., 'Parallel Corpora, Parallel Worlds', Rodopi, Amsterdam, New York, pp. 181–197. Proceedings of the Symposium on Parallel Corpora, Department of Linguistics, Uppsala University, Sweden, 1999.

Tiedemann, J. (2009), News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces, *in* N. Nicolov, G. Angelova & R. Mitkov, eds, 'Recent Advances in Natural Language Processing V', Vol. 309 of *Current Issues in Linguistic Theory*, John Benjamins, Amsterdam & Philadelphia, pp. 227–248.

Turchi, M., Flaounas, I., Ali, O., Bie, T., Snowsill, T. & Cristianini, N. (2009), Found in Translation, *in* 'Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II', ECML PKDD '09, Springer-Verlag, Berlin, Heidelberg, pp. 746–749.

Váradi, T. (2002), The Hungarian National Corpus, *in* 'Proceedings of the Third International Conference on Language Resources and Evaluation', Las Palmas, pp. 385–389.

Váradi, T. (2003), Shallow parsing of hungarian business news, *in* 'Proceedings of Workshop on Shallow Processing of Large Corpora, March 27 (SProLaC03)', Lancaster, UK.

Váradi, T. & Gábor, K. (2004), A magyar intex fejlesztéséről, *in* Z. Alexin & D. Csendes, eds, 'II. Magyar Számítógépes Nyelvészeti Konferencia', Szegedi Tudományegyetem, Szeged, pp. 3–10.

Waldenfels, R. v. (2011), Recent Developments in Parasol: Breadth for Depth and Xslt Based Web Concordancing with Cwb, *in* 'Proceedings of Slovko 2011, Modra, Slovakia, 20–21 October 2011', pp. 156–162.