LANGUAGE TECHNOLOGY ALGORITHMS FOR AUTOMATIC CORPUS BUILDING AND MORE PRECISE DATA PROCESSING

SUMMARY OF THE PHD THESES

István Endrédy

Academic advisor:
Gábor Prószéky, D.Sc.

PÁZMÁNY PÉTER CATHOLIC UNIVERSITY

FACULTY OF INFORMATION TECHNOLOGY AND BIONICS

ROSKA TAMÁS DOCTORAL SCHOOL OF SCIENCES AND TECHNOLOGY

Budapest, 2016

1 Introduction

This thesis is interdisciplinary: it is built on the common points of linguistics and information technology, as it focuses on corpus building and corpus management. Several linguistic studies are based on big text corpora. On the one hand, building and processing such a large amount of text require the methods and tools of information technology; on the other hand, these corpora serve the purposes of linguistics. Below, I present a method of building corpora automatically from the web, then research done on this corpus is shown.

More than one Hungarian corpus is available: BME MOKK (Halácsy et al. 2004) with 600 million words, or the Hungarian National Corpus (Váradi 2002) with a bit less than 190 million words. The upgraded version of the latter, the MNSZ2 (Oravecz, Váradi, and Sass 2014) has recently 784 million tokens. However, a bigger, comprehensive, up-to-date and annotated corpus with billion tokens is also needed. It is rather expensive to build a large corpus. On the contrary, if a corpus is built based on texts from the internet, it is free and its size can constantly increase. Furthermore, an up-to-date corpus makes it possible for linguists to investigate the frequent structures of the given language, and the changes in usage over time.

The study touches the themes of stemming, lemmatization, diacritic restoration, noun phrase detection and sentence analysis as well. The workflow was corpus driven: ideas were inspired and verified by the corpus.

The first part of this study describes the tasks, problems and results of **automatic corpus building** through the description of several corpora created during this research. I developed a web crawler with a new **boilerplate removal algorithm** which is able to learn the specific properties of websites automatically. The available solutions (Pomikálek 2011; Kohlschütter, Fankhauser, and Nejdl 2010) extract the valuable part of a webpage only by the features of the HTML content itself (based on the number of the links, stop words, words and tags). They perform well, but they sometimes extract texts with similar HTML features as the main article text, although they should be excluded from the final result (e.g. titles of related articles, comments, other leading news, etc.). These contents will be not just duplicates in the corpus, but they even impair the cohesion of the text. The **Gold Miner** algorithm was developed for this case. It works on a higher level than the features of the text, the algorithm searches and learns the typical patterns of a web domain where the main content used to be. If the pattern is defined, boilerplate removal algorithms get only the essential part of the page; as a result they perform much better.

The processing of the building corpus required tools which were driven by the corpus: Hungarian comments on the internet are usually written without diacritics. This can be fixed by a diacritic restoration application. The more accurate processing of texts needs a more accurate stemmer or lemmatizer. The **lemmatizer based on the analysis of Humor** and applications for its evaluation were also created.

Finally, this study traces the **detection of noun phrases** in English and Hungarian. Noun phrase detection is an important step of sentence analysis. The best ever result for Hungarian NP chunking belongs to the tool HunTag. Its newer, modular version (HunTag3) was used in my tests. More transition models (Viterbi, bigram, trigram), three tag sets (MSD, KR, Humor) were tested. The trigram model and new defined features resulted in the best improvement (3%). A rule based system and a statistical learning algorithm were also created and evaluated.

The error analysis showed that most errors occur in the detection of the adjacent NPs (incorrectly joining them together or incorrectly separating them), and possessors and their possesses are often missed. The corrections avoiding these error types have the most significant improvement.

2 New scientific results

My main results are following: the **GoldMiner** algorithm which is able extract text from web pages in a self learning way; the **lemmatizer** which is used in many applications and it is evaluated as the best in Hungarian by the **stemmer evaluation system**; the results in **NP chunking**; and the created **Pázmány corpus**.

The results presented in my dissertation can be divided into three thesis groups. The first thesis group includes the boilerplate removal algorithm which was used in corpus building. The second one describes my Humor based lemmatizer and diacritic restoration application and their evaluation method. Finally, the third thesis group presents the improvements in connection with the noun phrase detection.

THESIS GROUP I.

The thesis group I. focuses the corpus building process, especially the boilerplate removal problem, how the valuable content of a webpage can be extracted in an algorithmic way. The several repeated and irrelevant template contents of a web page make it harder to filter the main content. Menus, headers, footers, advertisements, the repeated structure on each page may vary not only in different domains, but in

time as well. As a result, an algorithm was needed which can adapt these changes automatically, without human interaction. The GoldMiner algorithm performed better than the available boilerplate removal solutions (Table 1).

The boilerplate removal GoldMiner algorithm learns the specific patterns of the domains based on a sample. It records the most frequent HTML tag sequence which includes the most valuable content.

Algorithm		Sentences	Unique sentences	Characters	Characters in unique sentences	Rate of unique sentences	Rate of characters in unique sentences
origo.hu	all texts	264 423	63 594	16 218 753	7 048 011	24%	43%
	BTE	60 682	33 269	12 016 560	7 499 307	54%	62%
	JusText	58 670	30 168	8 425 059	4 901 528	51%	58%
	GoldMiner	22 475	21 242	3 076 288	3 051 376	**94%**	**99%**
nol.hu	all texts	509 408	144 003	25 358 477	12 570 527	28%	49%
	BTE	154 547	107 573	24 292 755	13 544 130	69%	55%
	JusText	186 727	128 782	14 167 718	11 665 284	68%	82%
	GoldMiner	162 674	123 716	12 326 113	11 078 914	**76%**	**89%**
index.hu	all texts	232 132	55 466	9 115 415	4 542 925	23%	49%
	BTE	51 713	26 176	5 756 176	4 061 697	50%	70%
	JusText	40 970	29 223	4 371 693	3 441 337	71%	78%
	GoldMiner	13 062	11 887	1 533 957	1 489 131	**91%**	**97%**

Table 1. Results on 2 000 pages from various news portals

A crawler was also developed which collected Hungarian texts from the internet with the help of GoldMiner. The downloaded texts form the basis of the Pázmány corpus with 1.2 billion tokens from more than 30 000 domains (Table 2). I separated the comments in corpus for further processing, because comments have different quality and structure.

subcorpus	tokens	sentences	NPs
main corpus	954 298 454	48 536 849	223 347 534
other contents	228 806 919	15 802 499	52 865 889
comments	58 985 126	3 505 818	13 867 066
total	**1 242 090 499**	67 845 166	290 080 489

Table 2. The contents of the Pázmány corpus

Thesis 1: I created the GoldMiner algorithm which extracts the articles from web pages more effectively than previous algorithms.

Thesis 1.a: I developed a web crawler which is able to build a corpus.

Published in: [1], [9]

Thesis 7: I created the 1.2 billion token Pázmány corpus with the help of the crawler.

THESIS GROUP II.

In the second thesis group I created a lemmatizer and a diacritic restoring application. The evaluation of the former involved an evaluation method and algorithm as well the lemmatizer and the diacritic restorer count the results based on the morphemes of the Humor analyses (Prószéky and Kis 1999).

The basic idea of the lemmatizer algorithm is to count the lemma based on the role of the morph labels. The analysis goes from left to right, and the surface forms of the morphemes are concatenated to build the lemma, except the last morpheme: in this case the lexical form is applied to lemma. No doubt, it is a crucial question which morpheme belongs to the lemma and which does not. In case of derivational suffixes it is a more important question. For instance, word *adósság* (debt) with removed suffix *-ság* will result lemma *adós* (debtor). But if suffix *-s* and *-ó* are also removed, then lemma will be *adó* (tax) or *ad* (give). As can be seen, applying suffixes has great impact on the final result. In general, it should be taken into consideration which solution can be better for the given application. More removed suffixes may increase the recall of lemmatization (more possible relevant results), but it may decrease its precisions (more irrelevant results as well).

Several properties of the lemmatizer can be tuned from a configuration file: one can customize the morphemes for a lemma, the final part-of-speech of suffixes, the label conversions, or even the algorithm of the lemma determination.

I evaluated the lemmatizers in two ways: evaluation on the direct output and the performance in an information retrieval (IR) system was also measured. The direct metrics made it possible to measure the quality of each solution. On the contrary, the IR based evaluation did not require to give the exact lemma (stemmers do not provide it by design), but it is enough to count a stem (a common root form) of the infected word forms, independent from the lemma accuracy. The stem need not be a full word. The IR evaluation may show us, for example, if a language does not need a lemmatizer, a stemmer is enough for it. The gold standards in both of the evaluations were generated from lemma annotated corpora in an automated way. To provide an overview, the evaluation was done with 10 stemmers and lemmatizers for 3

languages with different levels of inflectional morphology (English, Polish and Hungarian).

I applied more metrics in the evaluation on the direct output of the stemmers and lemmatizers. I measured the accuracy of the first lemma alternative and the correctness and ranking of lemma alternatives. It may give more precise overview about a stemmer if we know its first stem accuracy, the rate of its incorrect stems or the quality of its stem ranking. The defined six metrics on the direct output of the stemmers make it possible not only to compare stemmers but to give feedback about typical stemmer errors, which can be useful for developing stemmers.

The earlier stemmer evaluations (Hull 1996; Tordai and De Rijke 2006; Halácsy and Trón 2007) were based on pre-defined, hand selected queries and their expected document sets. These sets were called experimental collections. The main idea of my evaluation is that every corpus with lemmas can be used as an IR evaluation data set. The basis of the method is the following: every sentence in the corpus will be the result item of an IR (hit), and its words (in their original forms) are the queries. The queries (the words) are connected to the result sets (the sentences) through their lemma and PoS tag. These word-sentence connections will be used as a gold standard, and each stemmer will be tested against this: calculation of precision and recall is based on the sets of sentences determined by stemmers and by the gold standard (illustrated on Figure 1).

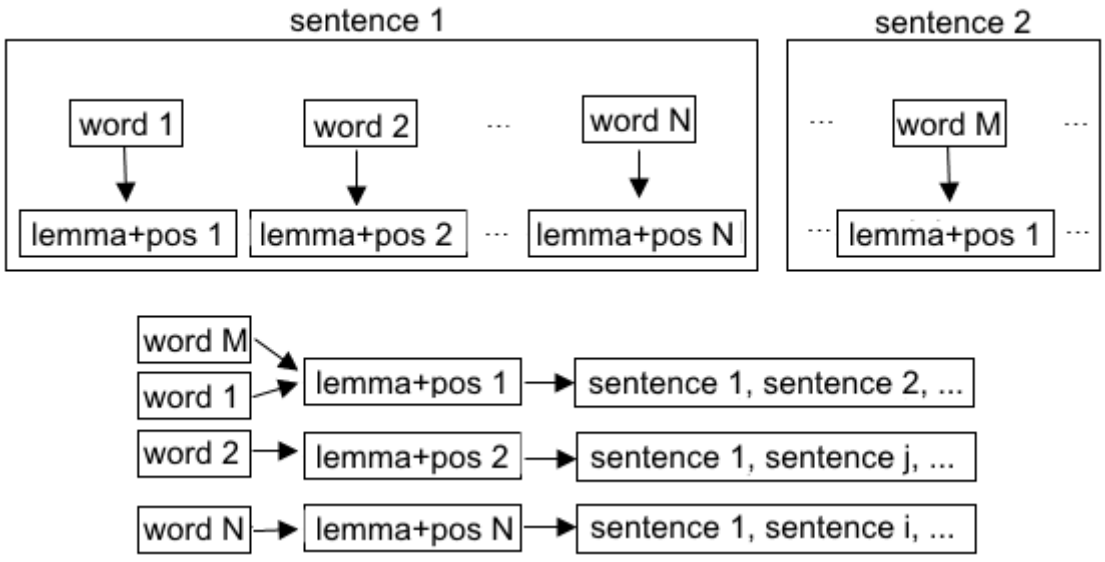

Figure 1. IR quality evaluation of a stemmer, based on a corpus with lemmas: sentences=documents and words=queries, sentence sets are compared to the gold standard

On the one hand, this evaluation method creates the gold standard automatically. As a result, the size of the test set can be big (for instance millions of doc in case of British National Corpus).

On the other hand, evaluation is made in two ways: evaluates all result items, and evaluates only the

first *n* results. This latter option reflects the mode when a human verifies the results: only the first *n* items matter. The possible incorrect results after *n* do not disturb the user: most probably nobody will visit them. The tests had an unwanted positive side effect: the evaluation of the first *n* results is able to evaluate even the ranking algorithm of the IR system itself. If this interval (the first *n* hits) does not contain the true positive results, then ranking is not good.

subcorpus	***No stemmer***	***Hunspell*** *(first stem)*	***Hunmorph-foma*** *(longest stem)*	***Hunmorph-compound*** *(first stem)*	***Hunmorph*** *(first stem)*	***Ocastem***	***Snowball***	***Humor*** *(longest stem)*
literature	52.4	86.6	76.2	86.6	86.4	**88.7**	58.3	88.4
students	52.9	**88.6**	78.1	88.2	88.1	88.0	57.0	88.3
newspaper	57.3	84.5	75.5	83.1	81.8	88.6	64.7	**92.8**
IT	57.9	81.9	75.8	81.7	79.3	87.9	68.6	**92.5**
juristic	62.0	81.8	77.4	82.4	80.8	86.7	72.4	**93.8**
business	55.5	78.1	68.9	80.2	78.9	87.6	65.2	**91.4**
Total	56.2	83.9	75.6	84.0	83.0	87.9	64.0	**91.0**

Table 3. Accuracy of the stemmer modules first/longest output on Szeged Corpus

domain	**no stemmer**	**Hu-light**	**Snowball**	**Hunspell**	**Humor**
literature	25.2	61.9	66.7	67.8	**78.3**
students	14.0	55.5	56.3	69.0	**75.4**
newspaper	16.5	79.6	81.1	77.1	**85.9**
IT	19.7	71.7	73.8	73.4	**81.8**
juristic	18.3	52.6	53.4	70.1	**75.3**
business	23.2	73.1	73.5	44.4	**87.9**
total	18.4	65.6	67.3	66.3	**80.1**

Table 4. IR quality evaluation of Hungarian stemmers on the sentences of Szeged TreeBank by Lucene

The best scores belong to my Humor based lemmatizer in both methods (Table 3 and 4). For this, the F-score IR was 80.1%, lemma accuracy was 91.0%, stem alternatives were 91-94%. The heavily agglutinative Hungarian language showed that only 18.4% can be achieved without any stemmer. Algorithmic stemming is not appropriate for this language.

domain	no stem		Stempfel		Morfologik		Hunspell		Humor	
	F	oov	F	oov	F	oov	F	oov	F	oov
literature	53.80	0	78.36	2.7	**89.34**	2.7	84.45	5.6	88.71	7.8
information	54.08	0	77.31	3.9	**89.09**	3.9	84.38	7.7	87.87	5.8
conversations	68.26	0	72.60	9.5	**83.62**	9.5	78.58	11.3	79.63	12.8
fiction	55.81	0	77.48	4.1	**88.98**	4.1	84.38	5.9	87.84	4.2
spoken radio	64.52	0	73.94	6.4	**86.11**	6.4	82.12	8.1	83.43	10.2
research & teaching	50.23	0	79.56	3.6	**89.64**	3.6	85.90	7.1	89.38	5.8
internet	55.27	0	77.15	3.2	**88.78**	3.2	83.48	8.0	86.02	7.8
journalistic	51.78	0	79.23	2.8	**90.03**	2.8	86.10	6.1	89.16	4.6
written-Parliament	51.24	0	78.77	4.3	89.85	4.3	85.48	7.4	**89.86**	2.7
utilities	52.22	0	80.82	1.2	90.65	1.2	87.48	7.5	**91.38**	2.3
not classified	54.38	0	78.72	2.0	**90.11**	2.0	85.55	3.9	88.52	3.9
total	54.21	0	78.2	3.6	**89.25**	3.6	85.02	6.7	88.08	5.4

Table 5. First lemma evaluation of Polish stemmers on gold standard lemmas of the PNC

domain	token	no stem	Stempel	Morfologik	Hunspell	Humor
literature	54 205	47.5	64.3	67.3	**71.6**	70.2
information	56 779	46.3	67.1	70.6	**73.9**	73.8
conversations	59 024	56.8	60.3	59.8	**64.4**	59.4
fiction	169 270	41.0	57.5	61.5	**65.4**	64.1
spoken radio	23 303	60.2	68.0	65.3	**71.6**	67.1
research & teaching	20 229	55.0	76.9	80.1	**82.5**	81.5
internet	72 273	50.0	63.4	65.3	**70.3**	67.3
journalistic	506 214	30.7	58.7	64.6	**67.5**	67.4
written-Parliament	66 315	42.9	66.7	79.9	82.6	**83.1**
utilities	30 998	47.8	77.2	81.2	83.3	**85.0**
not classified	10 140	64.7	73.3	72.3	**77.9**	76.5
total	1 028 671	36.9	60.4	65.3	**68.7**	68.0

Table 6. IR quality evaluation of Polish stemmers on the sentences of the PNC by Lucene

Polish results also show that stemming quality has great impact on the F-score of the IR. A stemmer can improve the F-score of an IR system at least 2 times (see Table 6), the best score belongs to Morfologik stemmer in lemma accuracy (Table 5), the best module is Hunspell in IR evaluation (Table 6).

In Hungarian, the characteristic of the two evaluation methods are similar. The better the quality of lemmatization is, the better the quality of the IR is. In the case of English, this connection is not verified. In the case of Polish and English, the characteristic of the two evaluation methods (direct output- and IR evaluation) were different.

Thesis 2. I developed a diacritic restoring module, designed with a co-author which has 94,3% precision with the help of a modified Humor lexicon.

Published in: [2], [11]

Thesis 3. I developed a lemmatizer engine, designed with a co-author which defines the lemma based on Humor analysis, and it has the best results for Hungarian according to the evaluations.

Published in: [2], [4]

Thesis 4. I created an evaluation method which can measure the (i) accuracy, (ii) IR quality, (iii) UI, OI, ERRT and (iv) further metrics of a stemmer. These evaluations were done on 10 stemmers for English, Polish and Hungarian.

Thesis 4a. I created a method for creating gold standard to evaluate stemmers in IR system.

Thesis 4b. I made the evaluation for English, Polish and Hungarian.

Thesis 4c. I presented the correlation between lemma accuracy and IR quality in the case of agglutinative languages (Polish, Hungarian).

Thesis 4d. I defined an IR stemmer evaluation which is able to give IR correlated evaluation results without native IR system.

Thesis 4e. I defined evaluations on the direct output of the stemmer which are able to compare stemmers, and they can give feedback about the typical errors of the module, for detection and fixing.

Thesis 4f. I presented and measured that the first *n* hits of an IR stemming evaluation set is able to evaluate the ranking algorithm of the IR as well.

Published in: [2], [4]

THESIS GROUP III.

In the third theses group, this study traces the detection of noun phrases in English and Hungarian. On the one hand, my research focused on feature defining and tuning for more precise noun phrase detection. I developed a tool which can measure the usefulness of features of a training set, and it gives an overview about the connection between features and output IOB labels. It cannot improve the features automatically, it is just an utility for the linguist. The tool combines the abstraction ability of a linguist and the power of a statistical engine.

On the other hand, the present state-of-the-art NP chunking system for Hungarian was produced by the error analysis of the HunTag, defining new features and applying HunTag3. The Szeged Treebank corpus was used (Csendes et al. 2005). The error analysis and tests showed that most of the errors were caused by neighboring NPs, the false unification of possessors and their possesses, and because the POS tag for participles and adjectives did not differ in the MSD formalism. Explicit features were defined for them, resulting in a significant improvement. The trigram transition model was the best in the case of Hungarian. All these improvements resulted in 93.59% F-score (illustrated in Table 7). The previous best result was 90.28% (Recski 2014). In English, CRF suite solution had the best scores according to my tests, but it does not outperform the state-of-the-art English result (95.23%).

	MSD	**KR**	**KR + better features**
T'n'T	68,52	70,95	-
baseline	81,71	88,72	-
HunTag	93,20	88,96	90,78
HunTag3 – bigram	93,43	89.10	90,72
HunTag3 – trigram	**93,59**	**89,83**	**91,50**
CRF	92,27	89,12	89,77

Table 7. Hungarian results on Szeged Treebank, F-scores with various POS tags and test sets. The best results were reached with the help of new POS category suggestions, and explicit features against typical HunTag mistakes

Thesis 5. I presented that the feature set can be further tuned for NP chunking, if (i) the given type of feature has only few value and (ii) if there is a more detailed classification where features correlate with IOB labels.

Thesis 5a. I divided the English part-of-speech tags into sub-categories which correlate better with IOB labels. (+2% improvement)

Thesis 5b. I created a method where the process described in Thesis 5.a can be run with help of synsets of WordNet.

Thesis 5c. I estimated the usefulness of a feature set counting the correlation between the label set and features, in a faster way than the original running time of training and test processes (NLTK unigram, bigram chunker; HunTag; SS05).

Thesis 5d. I demonstrated that the input annotation plays a very important role, it can be more important than the learning algorithm itself. Any external information helps which correlates in the output annotation.

Thesis 5e. I showed that there are only few features which correlate 100% with IOB labels, but all them are important.

Published in: [5], [6]

Thesis 6. I achieved better NP detection result by defining new features, applying a trigram transition model made with a co-author and applying HunTag3, which combined improvements resulted the best ever F-score in Hungarian NP detection (93.59%).

Thesis 6a. I presented at NP detection that the trigram transition model can perform better than bigram only in the case of more detailed IOB labels with more than three elements (IOB label with types or finer classes).

Published in: [6]

2.1 Application of the results

The lemmatizer presented here was built into applications of several companies: Microsoft Indexing Service, the document storing and searching system of the Országos Atomenergetikai Hivatal (Hungarian Atomic Energy Authority), the editor system of MTI, and PolyMeta search tool. The second edition of the Hungarian National Corpus (MNSZ2) was annotated with this tool as well.

The stemmer evaluation tool presented in the thesis group II can be used on other languages (in addition to English, Polish and Hungarian) and evaluation of further stemmer/lemmatizer modules as well. The python code developed for evaluation process has an option to generate automatically gold standard to measure lemma accuracy and IR quality. The evaluation script can be run for further stemmers.

The presented 1.2 billion token, lemma, part-of-speech and NP tagged corpus can be useful input for other language technology research.

Acknowledgments

I would like to say thank you to the persons without whom none of this would be possible.

I thank to my grammar teacher, Márta Bukovits, who could teach grammar in a very interesting way, and she gave the impression of grammar has mathematical precision. I thank to my professor Mátyás Naszódi at the university that he invited me to the world of language technology and to the MorphoLogic Ltd. The work and results presented here are strongly connected to my 13 years at MorphoLogic. That is why my dissertation can be considered a tribute to MorphoLogic. I am thankful for the years at this company, I learnt a lot from my colleagues: Balázs Kis the effective; Laci Tihanyi the fast; Miki Pál the accurate; Peti Kundráth the coder; Attila Novák the scientist; Zsolt Sebestyén I could always count on; Kati Hubay the problem solver; Andi Aggod the hard tester; Szabi Kincse the communicator, and everyone else who was great to work with.

I thank to my scientific supervisor, Gábor Prószéky, his encouragement and the conversations with good atmosphere. I was always welcomed at him, even he was very busy. His role was irreplaceable in my dissertation.

I thank to the reviewers of the PhD theses for their valuable suggestions.

I thank to my family, to my wife, Orsi, to my children Balázs, Kata, Dorka and Bence, that they supported me in this work, and they patiently accepted that I could spend less time with them. (I promised to my 9 year-old Dorka that we will build a workroom if I finish my dissertation. She asked me my twice a day: are you ready?)

I cannot thank enough to my parents for my life and the way they followed with attention this project too. To my sisters and brother, the way they love me.

I thank to the other PhD students for the common projects, the journeys. Balázs Indig takes the credit that he helped me in my deadlock – without him noticing it – when I wanted to give up. Special thanks to Nóra Wenszky for proofreading of my papers.

I would like to thank to the recent and previous leading professors of the university, Tamás Roska, Judit Nyékyné Gaizler and Péter Szolgay, who made it possible and supported my research in many ways. Thanks to Katinka Vida for the special attention, and to the administrative and financial staff for the work in the background.

Thanks for the holy masses on Wednesday at noon in the chapel of the university.

3 List of publications

Publications in journals

[1] **Endrédy, István**, Attila Novák. 2013. “More Effective Boilerplate Removal—The GoldMiner Algorithm” *Polibits Journal* 48: pp. 79–83.

[2] **Endrédy István**, Novák Attila. 2015. “Szótövesítők összehasonlítása és alkalmazásaik” In: Navracsics Judit (szerk.) *Alkalmazott Nyelvtudomány,* XV. évfolyam, 1-2. szám, pp. 7-27, Veszprém

Publication in book section

[3] Indig Balázs, **Endrédy István**. 2016. “Gut, Besser, Chunker - Selecting the best models for text chunking with voting” In: A. Gelbukh (Ed.) *Lecture Notes in Computer Science*: Computational Linguistics and Intelligent Text Processing Springer International Publishing, Berlin Heidelberg (*publishing in progress*)

Publications in proceedings of international conferences

[4] **Endrédy István**. 2015. “Corpus based evaluation of stemmers”, *7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 234-239, Poznań

[5] **Endrédy István**. 2015. “Improving chunker performance using a web-based semi-automatic training data analysis tool”, *7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 80-84, Poznań

[6] **Endrédy István**, Indig Balázs. 2015. “HunTag3, a general-purpose, modular sequential tagger – chunking phrases in English and maximal NPs and NER for Hungarian”, *7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 213-218, Poznań

[7] **Endrédy, István**. 2014. “Hungarian-Somali-English Online Dictionary and Taxonomy” In *Proceedings on “Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era,”* 38–43. Reykjavik, Iceland

[8] **Endrédy, István**, László Fejes, Attila Novák, Beatrix Oszkó, Gábor Prószéky, Sándor Szeverényi, Zsuzsa Várnai, and Beáta Wagner-Nagy. 2010. “Nganasan–Computational Resources of a Language on the Verge of Extinction” In *7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages (LREC 2010),* pp. 41-44 Valetta, Malta

Publications in proceedings of national conferences

[9] **István Endrédy,** Novák Attila. 2012. “Egy hatékonyabb webes sablonszűrő algoritmus – avagy miként lehet a cumisüveg potenciális veszélyforrás Obamára nézve.” In: *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pp 297–301. SZTE, Szeged

[10] Bakró-Nagy Marianne, **Endrédy István**, Fejes László, Novák Attila, Oszkó Beatrix, Prószéky Gábor, Szeverényi Sándor, Várnai Zsuzsa, Wagner-Nagy Beáta. 2010. “Online morfológiai elemzők és szóalakgenerátorok kisebb uráli nyelvekhez”. In: *VII. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 345–348, SZTE, Szeged

[11] Novák, Attila, **István Endrédy**. 2005. “Automatikus ë-jelölő program” In: *III. Magyar Számítógépes Nyelvészeti Konferencia*, pp 453–54. SZTE, Szeged

4 Bibliography

Csendes, Dóra, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. 'The Szeged Treebank.' In *Lecture Notes in Computer Science: Text, Speech and Dialogue*, 123–31. Springer.

Halácsy, Péter, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. 'Creating Open Language Resources for Hungarian.' *Proceedings of 4th Conference on Language Resources and Evaluation (LREC)*, 203–10.

Halácsy, Péter, and Viktor Trón. 2007. 'Benefits of Resource-Based Stemming in Hungarian Information Retrieval.' In *Evaluation of Multilingual and Multi-Modal Information Retrieval*, 99–106. Springer.

Hull, David A. 1996. 'Stemming Algorithms: A Case Study for Detailed Evaluation.' *JASIS* 47 (1): 70–84.

Kohlschütter, Christian, Peter Fankhauser, and Wolfgang Nejdl. 2010. 'Boilerplate Detection Using Shallow Text Features.' In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 441–50. WSDM '10. New York, NY, USA: ACM. doi:10.1145/1718487.1718542.

Oravecz, Csaba, Tamás Váradi, and Bálint Sass. 2014. 'The Hungarian Gigaword Corpus.' In *Proceedings of LREC*. Reykjavik.

Pomikálek, Jan. 2011. 'Removing Boilerplate and Duplicate Content from Web Corpora.' PhD dissertation, Masaryk University, Faculty of Informatics.

Prószéky, Gábor, and Balázs Kis. 1999. 'A Unification-Based Approach to Morpho-Syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages.' In *ACL*, edited by Robert Dale and Kenneth Ward Church. ACL. http://dblp.uni-trier.de/db/conf/acl/acl1999.html#ProszekyK99.

Recski, Gábor. 2014. 'Hungarian Noun Phrase Extraction Using Rule-Based and Hybrid Methods.' *Acta Cybernetica* 21 (3): 461–79.

Tordai, Anna, and Maarten De Rijke. 2006. *Four Stemmers and a Funeral: Stemming in Hungarian at Clef 2005*. Springer.

Váradi, Tamás. 2002. 'The Hungarian National Corpus.' In *LREC*.