Developing and harmonizing the modules of a multi-threaded text analyser program

Balázs Indig

Summary of the PhD dissertation

Pázmány Péter Catholic University
Faculty of Information Technology and Bionics
Roska Tamás Doctoral School of Sciences and Technology

Supervisor:
Gábor Prószéky PhD
Doctor of Science

Budapest, 2017

1 Introduction

Traditionally, natural language processing tools work like a *pipeline* with all its advantages and disadvantages. The pipeline starts with raw text and ends with the desired level of processing. The traditional modules are the following:

- Sentence splitter, Tokeniser
- Morphological analyser (MA), Part-of-Speech (POS) tagger
- Noun phrase (NP) chunker, Named-entity recogniser (NER) (shallow parsing)
- Syntactic parser
- Semantic parser
- Information-retrieval, Machine translation, etc.

Due to the simplicity of the pipeline, the individual modules do not need to know anything about the other modules. Thus, traditionally, their testing is performed on a *gold standard corpus* which is also called reference data, i.e. each module has to produce a perfect output from the perfect input. But in real life, the modules that do not have direct contact with the raw text must handle the accumulated errors of the modules that preceded them in the pipeline, thus their input is far from perfect. However, there are only a few references in the literature on how robust these systems are – as a part of a pipeline – in case of incorrect input.

This fact motivates the question: how could these modules work better together so that they could eliminate potential errors in time, in order to prevent their magnification later in the pipeline? Therefore, in my dissertation I review the freely available Hungarian state-of-the-art methods and I seek solutions to the problems related to their harmonisation.

In Hungarian, simple sentences can be divided into two distinct components. Firstly, there are *immediate constituent structures* – such as noun phrases – that have strictly bound word order and do not move freely in the sentence. Secondly, there are *verb phrases*. Among their components, we can find the aforementioned immediate constituent structures as arguments of the verbs. In my dissertation, I discuss these two classes in detail.

2 Methods

The different types of machine learning programs and pattern searching methods used in the dissertation can not yield results without comparison and analysis under standard conditions. Therefore, to evaluate the used programs, I applied the ordinary methodology: *precision*, *recall* and their harmonic mean, the so called *F-measure*. As input and expected output, I used the well-known available corpora.

Texts intended for scientific investigation are available in the form of corpora produced by concerning special criteria. These corpora also contain – mostly automatic – annotation of the texts. Methods based on supervised learning also need reference data which is manually prepared using a predetermined format and procedural order. The cost of such corpora is high because of the need for human resources. In Hungarian, the *Szeged Corpus* (Csendes et al. 2003) is the only manually annotated corpus available. It contains 70 000 sentences and 1 194 348 tokens. The *Szeged Treebank* (Vincze et al. 2010) is a variant of the Szeged Corpus which additionally contains dependency relations. In my dissertation I used this corpus as the Hungarian training corpus for maximal NP chunking.

For language modelling, it is enough to collect the largest amount of text possible, since processing does not require human resources. The only criterion for these texts is that they should be in the appropriate language in a normalised form and should form a coherent whole. Thanks to the increasing rate of internet communication, it is now very easy to get texts of different quality systematically from the Internet, so the number of automatically analysed corpora available for Hungarian is increasing (Indig 2018).

In my dissertation, I used two corpora for language modelling. One of them is the first and the second version (2.0.3) of the *Hungarian Gigaword Corpus* (Oravecz, Váradi, and Sass 2014). The first version contains 187 million and the second version 785 million words (978 million tokens) from various sources (transcripts of spoken texts, cross-border newspapers, legal texts, parliamentary diaries, etc.). The second corpus is the *Pázmány Corpus* (Endrédy 2016). It is entirely made of texts collected from the Internet and contains 1.2 billion words.

The dissertation presents measurements performed on the *InfoRádió Corpus* which contains only edited short news in the form of one or more sentence long utterances. With its 2 million words, it is a small domain-specific corpus which is a prototype of texts of the ideal input for the text analyser model presented in the dissertation. For the task of English arbitrary phrase chunking, I used the *CoNLL-2000 corpus* (Tjong Kim Sang and Buchholz 2000) as the gold standard, consisting of 259 104 tokens.

3 New scientific results

In my dissertation, I presented the natural language processing pipeline that is currently used for Hungarian. The pipeline architecture has many advantages and disadvantages. Nowadays, along with the long-known benefits, the disadvantages are slowly coming to light. Each module has contact only with the neighbouring modules so their input and output can vary greatly. As there are now many tools available to solve a given task, it is necessary to standardize and investigate their cooperation.

I have outlined the conditions necessary for the functioning of the ecosystems of such modules. I presented the requirements and expectations of the psycholinguistically motivated text analysis model, called AnaGramma framework, which I developed aiming at creating a computer-based text analysis model similar to a human analyser. The further goal of AnaGramma is to eliminate errors from the pipeline architecture which are magnified at the end of the pipeline and make the result unusable. The analytical system I have produced is inherently parallel, i.e. all modules work together and help each other to fix errors during operation. In the dissertation, I reviewed the procedures for the aforementioned model architecture. In Hungarian, noun phrases have been a good starting point because of the tight order of the elements in them, as well as their important role as arguments of verbs in a sentence. Thus, the dissertation is based on the harmony of these two groups.

Theses presented in my dissertation can be divided into four groups. From the state-of-the-art method of identifying noun phrases, through the various common properties of sequential labelling, I arrive methodically to the analysis of n-gram models to introduce the novel application of corpus samples needed for the text analyser framework. Subsequently, after examining the identification of noun phrases as verb arguments, I study the method of linking existing resources and transferring language-independent information to make the identification of verb frames more precise. Finally, I describe the architecture of the analyser framework and I present two explored and handled linguistic phenomena.

I. Automatic identification of noun phrases

In the first group of my theses, I focused on the task of maximal noun phrase chunking. I have examined the state-of-the-art methods currently used for Hungarian and English in order to understand how they could be used in the architecture of the text analyser framework. I wanted to adapt the English-language state-of-the-art method (Shen and Sarkar 2005) – its main contribution is per-

formance gain achieved by the simple majority voting on different IOB representations of the corpus trained and tagged by relying on the strengths of the individual representations – to Hungarian. However, during the reproduction of the English method it was found that the gain obtained by voting between the different IOB representations resulted from measurement error and was an artefact. Thus, the best method in English was not applicable to Hungarian since the result achieved with it is not real.

Thesis 1. *I have shown by measurements that contrary to the statement in the literature, voting between different IOB-representations does not improve the quality of the English noun phrase identification task significantly.*

Publications supporting the thesis: [3]

After that, during the examination of the state-of-the-art method for maximum NP chunking for Hungarian (Recski and Varga 2012), I discovered that the state-of-the-art method uses only the bigram transition model, as it has its origins in the method of Hungarian named-entity recognition. I have verified by measurements that the results of the model can be improved using a trigram transition model.

Thesis 2. *Using the HunTag3 program that I developed, I have verified (together with a co-author) that using trigrams can improve the results compared to bigrams in the task of Hungarian maximal noun phrase identification.*

Publications supporting the thesis: [8]

II. Lexicalisation methods

In the second group of theses, based on my investigations on the task of NP chunking, I recognised that sequential labelling tasks differ in many features but they also have many in common. This observation helped me in the design of the architecture of the text analysis framework. Therefore, I began to work on methods that are commonly applicable to all sequential labelling tasks which have the common property of handling text strictly left to right, just like the human parser. I used these methods to improve my existing results. In connection with this, I presented the inner-workings and the effect of the lexicalisation methods I examined.

Thesis 3. *I created a new, general lexicalisation method for sequential labelling, and its first application helped to improve the task of identification of arbitrary phrases.*

Publications supporting the thesis: [2, 3]

By using the lexicalisation method which I invented and the determination and application of the optimal threshold, I have improved the performance of the state-of-the-art method in arbitrary phrase chunking in English.

Thesis 4. *The method that I developed for the English noun phrase identification task outperforms verifiably the currently known methods according to the F-measure metric.*

Publications supporting the thesis: [2, 3]

Consequently, it can be stated that greater precision can be achieved with a finer classification and where precision is the goal, my method can improve the results. Therefore, I thought it was important to examine whether the converter during the conversion to different `IOB` representations and the tagger during testing can maintain the well-formedness of the parentheses of the output label sequences or not. To measure this, I developed a metric that I applied in practice to the task of English arbitrary phrase chunking.

Thesis 5. *I developed a parenthesis method which acts as a metric in order to rank the different methods for the labelling task by quality.*

Publications supporting the thesis: [2, 3]

III. Linking resources

Since the maximal noun phrases function as verb arguments in the sentence, I examined the available verb frame resources in Hungarian (Indig, Vadász, and Kalivoda 2017; Kalivoda 2016; Kornai, Nemeskey, and Recski 2016; Sass 2015; Sass et al. 2010). During my investigations, I did not find any semantic information in the verb frame resources which would help one to further refine the classification of nouns. I adapted the notion of *Linked Data*[1] and described the resource-related version of the method. I also presented some examples of linked resources in English (Prószéky, Mihăltz, and Kuti 2013; Vossen et al. 1998). After that, my goal was to link the bilingual Hungarian-English MetaMorpho database (Prószéky, Tihanyi, and Ugray 2004) with the English VerbIndex database (Loper, Yi, and Palmer 2007) in order to automatically translate a language-independent semantic annotation from the much richer information of VerbIndex into the MetaMorpho database.

[1] `http://linkeddata.org/`

Thesis 6. *I created an automatic method for linking the parallel verb frames with 1-, 2- and 3 arguments, which resulted in the successful transfer of appropriate thematic roles from English to Hungarian.*

Publications supporting the thesis: [11, 12, 4, 22]

As a part of the interconnection, the two ontologies describing the constraints of the elements in each resource had to be harmonised. I created the interoperability between the two resources with a unified ontology containing bridging concepts.

Thesis 7. *I created an ontology that links the verb frame descriptions of the Hungarian MetaMorpho with the syntactic and semantic categories of the English VerbIndex.*

Publications supporting the thesis: [11, 12, 4]

Based on existing information, an interconnection could be established between Hungarian and English WordNets as well. I have also put these links in place to improve the quality, but they have not proved to be appropriate to the task.

Thesis 8. *I demonstrated by measurement that the inclusion of the Hungarian and English WordNets can not improve the quality of the aforementioned ontology.*

Publications supporting the thesis: [11, 12, 4]

After all, I have successfully transmitted semantic information in high quality to a fairly well-organized subclass of verb frames by automatic means, which opened up further classification options.

IV. The architecture of the psycholinguistically motivated text analysing system

Following the lessons learned from the various linguistic phenomena, I presented the practical operation of the AnaGramma text analyzer, also from the aspects of the implementation. I defined the window used for handling language phenomena, the idea of which comes from the two phase parsing model called *Sausage Machine* (Frazier and Fodor 1978). The window – which corresponds to the first PPP phase of the *Sausage Machine* – and the search methods defined therein provide an efficient solution to the task of left-to-right analysis, in a similar manner to that assumed in case of human readers.

Thesis 9. *I created the foundations of a new approach (the so-called window-based text analysis model) – for which the theory was developed together with a co-author – which enables the Hungarian input text to be processed efficiently and similarly to human processing, from left to right.*

Publications supporting the thesis: [14, 5, 21, 34, 25, 26]

I have described the handling of the unmarked (nominative) and the marked (*-nAk* suffixed) possessive structures (Bánréti et al. 1992) using the aforementioned window.

In the first phase of the two-phased sentence analysis, a case disambiguation of nouns with zero case marking is performed using the lookahead window which clarifies the role of the nouns in the sentence. We have to decide whether the noun is a possible candidate for a possible nominative case or the possessor of a possible unmarked possessive structure. In the supply-and-demand framework, the possessed noun will be the one indicating demand for a possessor, which becomes a possessive edge between the possessed and the possessor when a supplied possessor is found.

Thesis 10. *I created an algorithm for disambiguating unmarked grammatical cases (for example, an effective real-time disambiguation of the Hungarian possessive structure and the subject), which theory was developed with a co-author.*

Publications supporting the thesis: [5, 21, 25]

Based on corpus measurements, it is also possible to use the supposed two token long analysis window within the framework of the AnaGramma text analyser system to create the edges between the finite verb and its preverb and the infinite verb and the finite verb. Using the pool and the window, the VFrame search engine connects the verbal elements (finite and infinite verbs) and their preverbs correctly in the sentence.

Thesis 11. *I created the VFrame procedure – for which the theory was developed together with a co-author – which enabled the discovery of the correct preverb for the finite verb knowing the distribution patterns of the possible preverbs and helped to filter the possible verb frames.*

Publications supporting the thesis: [14, 26, 27]

The introduced methods performed equally well, in the context of precision and recall.

4 Applications

The presented results are not widely applied due to their freshness, but the results of the noun phrase chunking and sequential labelling tasks have attracted great interest at international conferences. I believe that the results currently examined only in English can be adapted to Hungarian and some other agglutinating languages with some minor changes.

One of them could be the incorporation of the annotation marking of the border of NPs and other phrases with the existing part-of-speech tagger methods to be able to run the two tasks in parallel. The presented mild lexicalisation can be applied to various tasks beyond the ones presented in the dissertation. The application of the metric developed for checking well-formedness can save researchers from some inconvenience in all sequential tagging tasks in the future.

My theoretical results on linked resources can also be useful for those who are planning similar resource linking. It can be seen that in the system a number of errors can be made by humans due to the predominance of rule-based components, therefore, one should consider my experience before starting another similar project. One can see from my measurements that the current rule-based and statistical resources are not yet usable in combination. We are better off using only the rule-based systems for the presented task.

The presented work can be applied for example to create quality resources that contain valuable semantic information and enable high-precision semantic analysis. These resources can be the basis of various theoretical linguistic research.

Indeed, I have already applied this result in the module of connecting the right preverb–verb pairs in the presented parsing model. This application can be an example for researchers on the field of computational linguistics. A long-term goal could be the transformation of the language-agnostic information in a reliable way to Hungarian by using the presented ontologies, but with the fast evolution of the neural networks WordNet and other similar hand-made resources is shadowed in favour of the statistically well-founded resources, therefore there are doubts in their long term usability.

When designing the AnaGramma text analyser architecture, I used my former results presented in my dissertation. The theoretical significance of these results contributed greatly to the newly emerging results. The results described in the last chapter are also important for theoretical linguistics. Their application in psycholinguistic research can be expected as well.

The author's publications

International journal papers and book chapters

[1] Garay, Barnabás Miklós and **Balázs Indig** (2015). „Chaos in Vallis' asymmetric Lorenz model for El Niño". In: *Chaos, Solitons & Fractals* 75.1, pp. 253–262. issn: 0960-0779.

[2] **Indig**, **Balázs** (2017a). „Less is More, More or Less... – Finding the Optimal Threshold for Lexicalization in Chunking". In: *Computación y Sistemas* 21.4.

[3] **Indig**, **Balázs** and István Endrédy (2018). „Gut, Besser, Chunker – Selecting the best models for text chunking with voting". In: *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part I (Lecture Notes in Artificial Intelligence)*. Ed. by Alexander Gelbukh. Cham: Springer International Publishing. Chap. 29, 409–423. isbn: 978-3-319-75476-5. doi: 10.1007/978-3-319-75477-2_29.

[4] **Indig**, **Balázs**, András Simonyi, and Márton Miháltz (2018). „Exploiting Linked Linguistic Resources for Semantic Role Labeling". In: *Human Language Technology. Challenges for Computer Science and Linguistics. 7th Language and Technology Conference, LTC 2015, Poznań, Poland, November 27-29, 2015. Revised Selected Papers (Lecture Notes in Artificial Intelligence 10930)*. Ed. by Zygmunt Vetulani, Joseph Mariani, and Marek Kubis. Cham: Springer International Publishing. isbn: 978-3-319-93781-6. doi: 10.1007/978-3-319-93782-3.

[5] **Indig**, **Balázs**, Noémi Vadász, and Ágnes Kalivoda (2016). „Decreasing Entropy: How Wide to Open the Window?" In: *Theory and Practice of Natural Computing (Lecture Notes in Computer Science volume 10071)*. Ed. by Carlos Martín-Vide, Takaaki Mizuki, and Miguel A. Vega-Rodríguez. Cham: Springer International Publishing, 137–148. isbn: 978-3-319-49001-4. doi: 0.1007/978-3-319-49001-4_11.

National journal papers and book chapters

[6] Prószéky, Gábor and **Balázs Indig** (2015a). „Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel". In: *Alkalmazott nyelvtudomány* 15.1-2, pp. 29–44.

[7] Prószéky, Gábor, **Balázs Indig**, and Noémi Vadász (2016). „Performanciaalapú elemző magyar szövegek számítógépes megértéséhez”. In: *“Szavad ne feledd!”: Tanulmányok Bánréti Zoltán tiszteletére*. Ed. by Bence Kas. Budapest: MTA Nyelvtudományi Intézet, pp. 223–232.

International conference papers

[8] Endrédy, István and **Balázs Indig** (2015). „HunTag3: a general-purpose, modular sequential tagger – chunking phrases in English and maximal NPs and NER for Hungarian”. In: *7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. (Poznań, Poland, Nov. 27–30, 2015). Poznań, Poland: Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu, pp. 213–218. isbn: 978-83-932640-8-7.

[9] **Indig**, **Balázs** (2017b). „Mosaic n-grams: Avoiding combinatorial explosion in corpus pattern mining for agglutinative languages”. In: *8th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. (Poznań, Poland, Nov. 17–19, 2017). Poznań, Poland: Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu, pp. 147–151. isbn: 978-83-64864-94-0.

[10] **Indig**, **Balázs** (2018b). „The stability of the parameter transformation with Zipfian distributions across corpora”. In: *Computational Linguistics and Intelligent Text Processing: 19th International Conference, CICLing 2018, Hanoi, Vietnam, April 18–24, 2018, Revised Selected Papers, Part I (Lecture Notes in Artificial Intelligence)*. Ed. by Alexander Gelbukh. (Accepted, in press). Cham: Springer International Publishing.

[11] **Indig**, **Balázs**, Márton Mihaltz, and András Simonyi (2015). „Exploiting Linked Linguistic Resources for Semantic Role Labeling”. In: *7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. (Poznań, Poland, Nov. 27–30, 2015). Poznań, Poland: Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu, pp. 140–144. isbn: 978-83-932640-8-7.

[12] **Indig**, **Balázs**, Márton Mihaltz, and András Simonyi (2016). „Mapping Ontologies Using Ontologies: Cross-lingual Semantic Role Information Transfer”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Por-

torož, Slovenia: European Language Resources Association (ELRA), pp. 2425–2430. isbn: 978-2-9517408-9-1.

[13] **Indig**, **Balázs**, András Simonyi, and Noémi Ligeti-Nagy (2018). „What's Wrong, Python? – A Visual Differ and Graph Library for NLP in Python". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA). isbn: 979-10-95546-00-9.

[14] **Indig**, **Balázs** and Noémi Vadász (2016b). „Windows in Human Parsing – How Far can a Preverb Go?" In: *Proceedings of the Tenth International Conference on Natural Language Processing (HrTAL2016) 2016, Dubrovnik, Croatia, September 29-October 1., 2016*. Ed. by Marko Tadić and Božo Bekavac. (Accepted, in press).

[15] Miháltz, Márton, Bálint Sass, and **Balázs Indig** (2013). „What Do We Drink? Automatically Extending Hungarian WordNet With Selectional Preference Relations". In: *Proceedings of the Joint Symposium on Semantic Processing: Textual Inference and Structures in Corpora*. (Nov. 20–22, 2013). Ed. by Octavian Popescu and Alberto Lavelli. Trento, Italy: Association for Computational Linguistics (ACL), 105–109. isbn: 978-1-6299353-9-3.

[16] Váradi, Tamás, Eszter Simon, Bálint Sass, Iván Mittelholcz, Attila Novák, **Balázs Indig**, Richárd Farkas, and Veronika Vincze (2018). „E-magyar – A Digital Language Processing System". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA). isbn: 979-10-95546-00-9.

National conference papers

[17] **Indig**, **Balázs** (2013b). „PureToken: egy új tokenizáló eszköz". In: *IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013)*. Ed. by Attila Tanács and Veronika Vincze. Szegedi Tudományegyetem Infor-

matikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 305–309.

[18] **Indig**, **Balázs** (2018a). „Közös crawlnak is egy korpusz a vége – Korpuszépítés a CommonCrawl .hu domainjából”. In: *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Ed. by Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 125–135.

[19] **Indig**, **Balázs**, László János Laki, and Gábor Prószéky (2016). „Mozaik nyelvmodell az AnaGramma elemzőhöz”. In: *XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016)*. Ed. by Attila Tanács, Viktor Varga, and Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 260–270.

[20] **Indig**, **Balázs** and Gábor Prószéky (2013). „Ismeretlen szavak helyes kezelése kötegelt helyesírás-ellenőrző programmal”. In: *IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013)*. Ed. by Attila Tanács and Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 310–317.

[21] Ligeti-Nagy, Noémi, Noémi Vadász, Andrea Dömötör, and **Balázs Indig** (2018). „Nulla vagy semmi? Esetegyértelműsítés az ablakban”. In: *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Ed. by Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 25–37.

[22] Miháltz, Márton, **Balázs Indig**, and Gábor Prószéky (2015). „Igei vonzatkeretek és tematikus szerepek felismerése nyelvi erőforrások összekapcsolásával egy kereslet-kínálat elvű mondatelemzőben”. In: *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*. Ed. by Attila Tanács, Viktor Varga, and Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 298–302.

[23] Novák, Attila, György Orosz, and **Balázs Indig** (2011). „Javában taggelünk”. In: *VIII.. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2011)*. Ed. by Attila Tanács and Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 310–317.

[24] Prószéky, Gábor, **Balázs Indig**, Márton Miháltz, and Bálint Sass (2014). „Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási

modell felé". In: *X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014)*. Ed. by Attila Tanács, Viktor Varga, and Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 79–87.

[25] Vadász, Noémi and **Balázs Indig** (2018). „A birtokos esete az ablakkal". In: *LingDok: nyelvész-doktoranduszok dolgozatai*. Ed. by György Scheibl. Szegedi Tudományegyetem. Nyelvtudományi Doktori Iskola, pp. 85–99.

[26] Vadász, Noémi, Ágnes Kalivoda, and **Balázs Indig** (2017). „Ablak által világosan – Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével". In: *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017)*. Ed. by Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 3–12.

[27] Vadász, Noémi, Ágnes Kalivoda, and **Balázs Indig** (2018). „Egy egységesített magyar igei vonzatkerettár építése és felhasználása". In: *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Ed. by Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 3–15.

[28] Váradi, Tamás, Eszter Simon, Bálint Sass, Mátyás Gerőcs, Iván Mittelholcz, Attila Novák, **Balázs Indig**, Gábor Prószéky, Richárd Farkas, and Veronika Vincze (2017). „Az `e-magyar` digitális nyelvfeldolgozó rendszer". In: *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017)*. Ed. by Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 49–60.

Other publications

[29] **Indig**, **Balázs** (2013a). „An extended spell checker for unknown words". In: *Pázmány Péter Catholic University PhD Proceedings* 8, pp. 29–32.

[30] **Indig**, **Balázs** (2014a). „Towards a Psycholinguistically Motivated Performance-Based Parsing Model". In: *PhD Proceedings Annual Issues of the Doctoral School Faculty of Information Technology and Bionics* 2014, pp. 133–136.

[31] **Indig**, **Balázs** (2014b). „Towards recognizing thematic roles for verbal frames by linking two independent language resources for a parser based on the supply and demand paradigm". In: *PhD Proceedings Annual Is-*

sues of the Doctoral School Faculty of Information Technology and Bionics 2015, pp. 159–161.

[32] **Indig**, **Balázs** and Noémi Vadász (2016a). *POS Comes with Parsing: a Refined Word Categorisation Method*. Konfereciaabsztrakt (konferenciakötetbe nem került), 4th International Conference on Statistical Language and Speech Processing (SLSP 2016), Csehország, Plzeň, 2016. október 11-12. Pilseň, Czech Republic. url: http://grammars.grlmc.com/SLSP2016/Download/slides/pos-comes-with-parsing-abstract.pdf.

[33] **Indig**, **Balázs**, Noémi Vadász, and Ágnes Kalivoda (2017). *Manócska – integrált igeivonzatkeret-adatbázis*. url: https://github.com/ppke-nlpg/manocska.

[34] Prószéky, Gábor and **Balázs Indig** (2015b). *Natural parsing: a psycholinguistically motivated computational language processing model*. Konfereciaabsztrakt (konferenciakötetbe nem került), 4th International Conference on the Theory and Practice of Natural Computing (TPNC 2015), Spanyolország, Astruias, Mieres, 2015. december 15-16. Mieres, Astruias, Spain. url: http://grammars.grlmc.com/TPNC2015/Slides/d1s503natural_parsing_abstract.pdf.

References

Bánréti, Zoltán, István Kenesei, András Komlósy, Tibor Laczkó, and Anna Szabolcsi (1992). *Strukturális magyar nyelvtan I: Mondattan*. Ed. by Ferenc Kiefer and Zsófia Róbert. Akadémiai Kiadó. isbn: 963-05-6468-8.

Csendes, Dóra, Csaba Hatvani, Zoltán Alexin, János Csirik, Tibor Gyimóthy, Gábor Prószéky, and Tamás Váradi (2003). „Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz". In: *I. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003)*. Ed. by Zoltán Alexin and Dóra Csendes. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 238–245.

Endrédy, István (2016). „Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz". PhD thesis. Budapest: PPKE-ITK.

Frazier, Lyn and Janet Dean Fodor (1978). „The Sausage Machine: A New Two-Stage Parsing Model". In: *Cognition* 6.4, pp. 291–325.

Indig, Balázs (2018). „Közös crawlnak is egy korpusz a vége – Korpuszépítés a CommonCrawl .hu domainjából". In: *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Ed. by Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 125–135.

Indig, Balázs, Noémi Vadász, and Ágnes Kalivoda (2017). *Manócska – integrált igeivonzatkeret-adatbázis*. url: `https://github.com/ppke-nlpg/manocska`.

Kalivoda, Ágnes (2016). „A magyar igei komplexumok vizsgálata". MA thesis. PPKE-BTK. url: `https://github.com/kagnes/hungarian_verbal_complex`.

Kornai, András, Dávid Márk Nemeskey, and Gábor Recski (2016). „Detecting Optional Arguments of Verbs". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA). isbn: 978-2-9517408-9-1.

Loper, Edward, Szu-Ting Yi, and Martha Palmer (2007). „Combining lexical resources: mapping between PropBank and VerbNet". In: *Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg*, pp. 118–128.

Oravecz, Csaba, Tamás Váradi, and Bálint Sass (2014). „The Hungarian Gigaword Corpus". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA). isbn: 978-2-9517408-8-4.

Prószéky, Gábor, Márton Miháltz, and Judit Kuti (2013). „Lexikális szemantika: a számítógépes nyelvészet és a pszicholingvisztika határán". In: *Általános Nyelvészeti Tanulmányok* XXV, pp. 143–172.

Prószéky, Gábor, László Tihanyi, and Gábor Ugray (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), pp. 138–142.

Recski, Gábor and Dániel Varga (2012). „Magyar főnévi csoportok azonosítása". In: *Általános Nyelvészeti Tanulmányok* XXIV. Ed. by Gábor Prószéky, Tamás Váradi, and István Kenesei.

Sass, Bálint (2015). „28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet". In: *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*. Ed. by Attila Tanács, Viktor Varga, and Veronika Vincze. Szegedi Tudományegyetem Informatikai Intézet. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 399–403.

Sass, Bálint, Tamás Váradi, Júlia Pajzs, and Margit Kiss (2010). *Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára.* Budapest: Tinta Könyvkiadó.

Shen, Hong and Anoop Sarkar (2005). „Voting Between Multiple Data Representations for Text Chunking". In: *Proceedings of the Advances in Artificial Intelligence, 18th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2005, Victoria, Canada, May 9-11, 2005*. Ed. by Balázs Kégl and Guy Lapalme. Vol. 3501. Lecture Notes in Computer Science. Springer, pp. 389–400.

Tjong Kim Sang, Erik F. and Sabine Buchholz (2000). „Introduction to the CoNLL-2000 Shared Task: Chunking". In: *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*. ConLL '00. Lisbon, Portugal: Association for Computational Linguistics, pp. 127–132.

Vincze, Veronika, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik (2010). „Hungarian Dependency Treebank". In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid

Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias. Valletta, Malta: European Language Resources Association (ELRA), pp. 1855–1862. isbn: 2-9517408-6-7.

Vossen, Piek, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters (1998). *The EuroWordNet base Concepts and Top Ontology*. Tech. rep.