Hybrid algorithms for preprocessing agglutinative languages and less-resourced domains effectively

Summary of the PhD dissertation

György Orosz

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

Supervisor:
Gábor Prószéky, DSc

Budapest, 2015

1 Introduction

Natural language technology eases everyday life by helping the information flow between humans and computers. Diverse applications of the field exist which can aid e.g. writing texts, understanding foreign languages or finding relevant information pieces. Text processing is a branch of language technology, which includes the automated analysis of textual data. Several processing layers can be distinguished such as text segmentation, morphological parsing, syntactic parsing or semantic analysis. Practical applications often build tools pipelining such components one after another. While many layers are not available for numerous languages, two preprocessing steps are indispensable for most of the cases. Words and sentences are the basic units of text mining applications, therefore segmentation must be performed first. Beside this, the lemmata and the part-of-speech (PoS) of words are also necessary components of such systems, thus morphological parsing is carried out next.

In most of the cases, text segmentation systems are accurate and remain robust amongst domains, hence the task is considered to be solved. However, there are numerous scenarios (such as the case of noisy texts) on which existing approaches fail, thus posing new challenges to researchers.

Further on, PoS tagging is another well-researched field of natural language processing (NLP), as diverse methods exist solving the problem in many languages. In practice, these algorithms mostly build on data-driven techniques thus restricting their applicability by the corpus they model. Furthermore, most of the tagging methods target English first, hence ignoring serious problems caused by rich morphological systems. In that way, disambiguating between part-of-speech labels becomes insufficient, thus full morphological tagging algorithms are required assigning full morpho-syntactic tags and computing lemmata as well. Therefore, language technology needs preprocessing methods which can handle morphologically rich languages efficiently and perform well on less-resourced scenarios at the same time.

The aim of this study is twofold. Firstly, morphological tagging algorithms are investigated which can handle agglutinative languages and domain adaptation scenarios effectively. Secondly, methods suitable for less-resourced noisy domains are examined.

First, we were interested in **how existing methods can be applied for full morphological tagging of agglutinative languages yet remaining suitable for domain adaptation tasks**. As a result we present an accurate lemmatization method and an efficient morphological tagging chain. Following this, we examined **how one can improve combination schemes of full morphological taggers in order to raise the overall annotation quality**. In that way, an architecture is introduced which fits well for agglutinative languages and improves the baselines used.

Beside tagging methods, their applications also played a central role in this study. We were interested in **creating a proper tagger tool for speech transcripts which can help linguists in their research**. In doing so, a methodology is described that can estimate morphosyntactic complexity of speech transcripts automatically.

The third part of this study deals with the preprocessing of electronic health records. In particular, first we were interested in **how one can develop a proper text segmentation algorithm using existing methods**. A hybrid framework was presented using diverse symbolic and machine learning components, thus resulting in a precise tokenization and SBD method. Following this, we looked into the questions **what the main pitfalls of morphological taggers are which target noisy clinical texts** and **how PurePos can be adapted for tagging medical texts properly**. This part presents several adaptation techniques relying on domain-specific knowledge, thus improving the annotation quality significantly.

2 Methods

In the course of our work, diverse corpora were used. First, the Szeged Corpus [22] was employed for developing and evaluating general tagging methods. Further on, these algorithms were tested on Old and Middle Hungarian [12] texts as well. Next, methods for speech transcripts were analyzed on the HUKILC corpus [2].

Beside existing ones, two new corpora were created manually from electronic health records. These texts enabled us to design algorithms for the clinical domain. Concerning their usage, texts were usually split into training, development and test sets.

As regards methods used, most of our work resulted in hybrid solutions. On the one hand, we built on symbolic morphological analyzers and rule-based (pattern matching) components. On the other hand, stochastic and machine learning algorithms were heavily utilized as well.

Morphological analyzers played a central role in our study, since their usage is inevitable for morphologically complex languages. In most of the cases we employed (adapted versions [12, 23, 16]) of Humor [24, 25, 26] but the MA of `magyarlanc` [27] was used as well.

As regards machine learning algorithms, tagging experiments were based on hidden Markov models [28, 29]. Our approach built on two well-known tools which are Brant's TnT [30] and HunPos [31] from Halácsy et al. Besides, other common methods such as n-gram modeling, suffix-tries and general interpolation techniques were utilized as well. Further on, the proposed combination scheme applied instance-based learning [32] implemented in the Weka toolkit [33].

Beside supervised learning, unsupervised techniques were employed as well. Identification of sentences was performed using the collocation extractions measure of Dunning [34]. In fact, we based on the study of Kiss and Strunk [35], which employs scaling factors for the $\log \lambda$ ratio.

The effectiveness of algorithms was measured calculating standard metrics. The performance of taggers were computed with accuracy as counting correct annotations of tokens and sentences. However, if the corpus investigated contained a considerable amount of punctuation marks, they were not involved in the computation. For significance tests, we used the paired Wilcoxon signed rank test as implemented in the SciPy toolkit [36]. Next, the improvement of taggers was examined calculating relative error rate reduction.

Simple classification scenarios were evaluated computing precision, recall and F-score for each class. Furthermore, overall accuracy values were provided as well. Finally, numeric scores were compared with mean relative error [37] and Pearson's correlation coefficient [37].

3 New scientific results

I Effective morphological tagging methods for morphologically rich languages

Full morphological tagging is a complex task composed of two parts. Beside identifying morpho-syntactic tags, lemmata of words must be computed as well. While the first task is a well-known problem of natural language processing, the latter one is often neglected. Results are summarized by describing the new lemmatization method first, followed by the full tagging systems.

> THESIS I.1. **I developed a new lemmatization method for agglutinative languages. The presented algorithm is based on the output of a morphological analyzer. It can handle both known and unknown words effectively by incorporating diverse stochastic models. Results presented show that the new system has high accuracy on Hungarian texts.**
> Publications: [18, 17, 10, 8]

The proposed algorithm performs lemmatization in two steps. First, it uses a morphological analyzer and a guesser component to generate lemma candidates, then disambiguation is performed using stochastic models. The latter part is carried out calculating the score (S) of each lemma (l) for a given word (w) and tag (t) using the interpolation of two different models:

$$S(l|w,t) = P(l)^{\lambda_1} P(l,t|w)^{\lambda_2} \tag{1}$$

The system combines a simple unigram model with the output of a suffix-based guesser. To calculate the lambda parameters, guesses of models are evaluated on the training data, then the better model's score gets increased while that of the worse one is decreased.

Several experiments have been presented on the Szeged Corpus showing that the proposed method has superior accuracy for Hungarian compared to other available tools.

— • —

THESIS I.2. I designed a hybrid morphological tagging system (PurePos[1]) for less-resourced and agglutinative languages. The method relies on stochastic methods incorporating the output of a morphological analyzer. Its lemmatization component utilizes algorithms presented in Thesis I.1. Furthermore, the tool is built up in a way to be able to incorporate domain-specific rules effectively. Experiments confirm its state-of-the-art accuracy for Hungarian and resource-scare scenarios.

Publications: [18, 17, 10, 8]

The architecture of PurePos (cf. Figure 1) is built up to allow multiple models cooperating effectively. The disambiguation is carried out in multiple steps. The data flow starts from a MA providing word analyses as *(lemma,*

[1]The presented system is open source and is freely available at https://github.com/ppke-nlpg/purepos

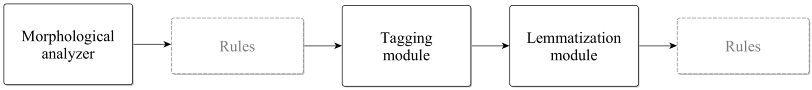

Figure 1 The architecture of the full morphological tagging tool

tag) pairs. Next, trigram-tagging methods (see [30, 31]) are employed for selecting morpho-syntactic labels of words. Finally, lemmatization is carried out employing the methods presented in Thesis I.1.

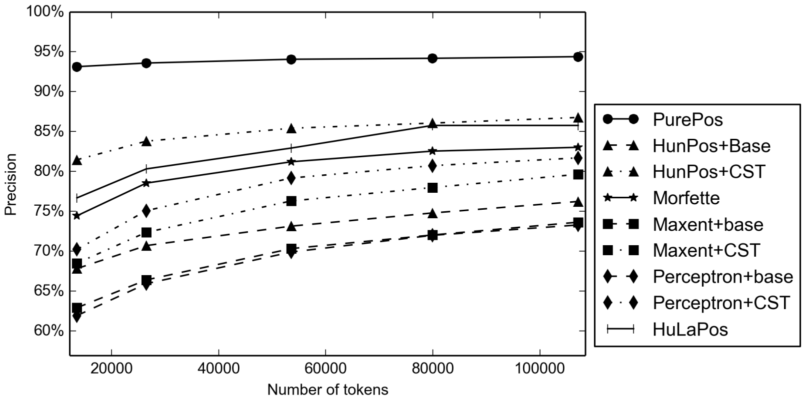

Figure 2 Learning curves of full morphological taggers on the Szeged Corpus (using Humor labels)

Several experiments were carried out measuring the performance of PurePos on the Szeged Corpus [22]. Results show that the new method yields very high (96.26%) full tagging accuracy on Hungarian. Moving on, I also compared existing tagging systems with the presented one on a less-resourced scenario. These experiments showed (cf. Figure 2) that PurePos can be successfully used even when the training dataset is limited. Finally, all the hybrid enhancements of PurePos ware evaluated one-by-one, showing that they can be used to fix several sorts of errors.

— • —

Although, methods of Thesis group I.2 have high accuracy, it was shown that they can be improved further. Therefore, a combination technique is presented increasing the ceiling of morphological tagging tools' performance for agglutinative languages.

THESIS I.3. I developed a methodology for combining morphological tagging systems effectively. The system presented selects the best lemma and tag candidates separately using two different combination methods. These components are trained with cross-validation using instance based learning. I showed that my method can significantly reduce the number of errors of existing annotation tools.
Publications: [20, 9, 5]

First of all, discrepancy of tagging systems was analyzed. For this, I designed a new metric (Own Error Rate) which measures the differences of output of taggers. It turned out that the most typical mistakes of HuLaPos [7] and PurePos are different enough to be aggregated.

Following this, the most common combination techniques were investigated considering their applicability to full morphological tagging. Next, a new combination method was presented involving adapted feature sets for a morphologically rich language. It utilizes instance based learning [32] and trains classifiers with cross-validation, which can employ the whole training dataset for both the baseline tools and the level-one learners. The novelty of the presented method is its architecture (cf. Figure 3) which allows us to utilize different combiners for the lemmatization and PoS tagging subtasks.

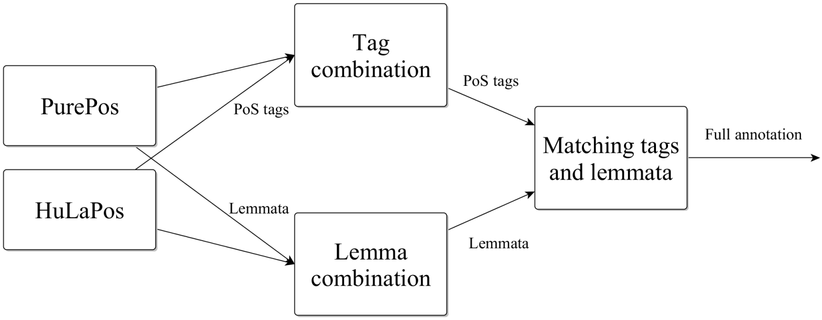

Figure 3 Combining the output of two PoS taggers and lemmatizers

Finally, evaluation experiments were presented indicating that the number errors of the best tagger can be decreased further. The new algorithm could reduce the number of errors of PurePos by 28.90%.

II Measuring morpho-syntactic complexity using morphological annotation algorithms

Measuring morpho-syntactic complexity is usually carried out calculating mean length of utterances. This metric is often computed in words for analytical languages, while morphemes (MLUm) are used for morphologically complex ones. Although automatic methods and tools exist for e.g English, other less-resourced languages lack such systems. Therefore, MLUm could be only computed manually, which is a rather time-consuming task.

This thesis group presents[2] methods for processing speech transcripts effectively and estimating mean length of utterance in morphemes automatically.

[2]This research has been conducted together with Kinga Jelencsik-Mátyus. My contributions are the construction of the tagging chain, its adaptation and the automatization of the MLUm calculation.

THESIS II.1. I developed a hybrid morphological tagging chain for Hungarian child-language transcripts. My method builds on top of the results presented in Thesis I.2 by adapting them to the domain. Evaluation shows that performance of the method is comparable with that of tagging methods for written corpora. Moreover, experiments indicate that the algorithm presented is accurate enough to be used in further applications.
Publications: [2, 4]

The proposed method adapts the algorithms introduced in Thesis I.2 for spoken Hungarian. For this, the Humor morphological analyzer was augmented first with analyses of words typical to the domain. Next, the output of PurePos was adjusted utilizing domain-specific knowledge.

For this, a gold corpus of about 1,000 utterances from the HUKILC was created by the manual annotation of texts. Additionally, a new tagging scheme was designed representing the characteristics of spoken language properly.

The evaluation of the chain resulted in 96% token-level precision, which is comparable with that of taggers for corpora of written language. Therefore, my investigation showed that PurePos is an appropriate base for tagging corpora of transcribed spoken texts.

— • —

THESIS II.2. I proposed a new algorithm for estimating morpho-syntactic complexity (calculating mean length of utterance in morphemes) in Hungarian child language transcripts. The method uses the morphological tagging chain of Thesis II.1 as a base. Evaluation of the system indicates that the methodology presented can properly replace the time-consuming manual computation of human annotators.
Publications: [2, 4]

The estimation method analyzes morphological annotations of tokens. Words known by the analyzer are decomposed by Humor, while lengths of unknown words are guessed based on their PoS labels. This is followed by morpheme counting rules implementing linguistic guidelines, thus providing relevant estimates.

As regards resources, a manually checked corpus was created for the experiments. Evaluation of the methods on this dataset shows that my results highly correlate (0.9901) with counts of human annotators. Further on, I showed that the mean relative error of the method is only 4.49%. Thus, the proposed algorithm can properly replace the labor-intensive human computation.

III Effective preprocessing methods for a less-resourced noisy domain

More and more electronic health records are produced in hospitals containing valuable but hidden knowledge. Since doctors cannot spend enough time on writing their reports properly, notes often contain numerous errors. Because of such mistakes, processing of these texts cannot be carried out using general-purpose tools. Moreover, while several algorithms are becoming available for English, Hungarian and other morphologically rich languages are still neglected.

> **THESIS III.1. I developed a new framework which segments noisy clinical records into words and sentences accurately. The method is built on top of well-known tokenization rules (e.g. [38]), however, it augments them with unsupervised heuristics. Evaluations showed that the algorithm can properly identify word and sentence boundaries in noisy clinical notes. Results also indicate that other systems available cannot handle such erroneous texts.**
> Publications: [5, 14, 3]

Figure 4 The architecture of the proposed method

The proposed method builds on pattern-matching algorithms taken from general-purpose tokenization tools. Even though these methods perform with high accuracy, their recall still stays low. Therefore, this study proposes a method (see Figure 4) which improves their performance using unsupervised heuristics and a domain-specific morphologic analyzer. First, the scaled $\log\lambda$ method [35] was adapted by introducing new scaling factors. Next, the Humor morphological analyzer was utilized to reveal further sentence boundaries.

The evaluation of the framework was carried out on a manually segmented corpus. Numerous metrics (such as precision, recall, F-score) were employed measuring the performance of the proposed tool. Moreover, existing Hungarian approaches were also compared with the proposed one.

Results show that other systems available can only produce low quality segmentation. Most of them yields F-scores less than 50% in sentence boundary identification. On the contrary, the method proposed can detect both token and sentence boundaries accurately, producing F-values over 90%.

— • —

THESIS III.2. I showed that tagging methods of Thesis I.2 can be applied for annotating electronic health records satisfactorily. In doing so, PurePos was adjusted with stochastic and symbolic domain adaptation techniques. The quality of the annotation produced is comparable with that of general written tagger tools.

Publications: [16, 1, 3]

First of all, an extended version of the Humor analyzer was used as a base of the tagging chain, since it was prepared[3] for electronic health records. Further on, the tagging chain was improved using a detailed error analysis of the baseline tagger.

For this, a manually annotated corpus was created containing texts of clinical notes. Results on this dataset show that the improved system performs significantly better (93.73%) than the baseline system (88.09%). However, future work might target the segmentation and tagging tasks with a unified framework, since both systems have the most problems with abbreviated terms.

4 Applications

The methods presented here solve basic preprocessing tasks such as text segmentation and morphological tagging. Since these are essential components of any language processing chain, our results can be applied in numerous fields of natural language technology. In general, text mining solutions and information extraction tools utilize such algorithms. Since our methods aim morphologically rich and less-resourced languages (and especially Hungarian), they can be used to boost tasks involving such languages.

Concerning general tagging methods of Theses I.1 and I.2, they have been successfully applied in several Hungarian projects. Their applications involve the following studies:

1. Laki et al. [7] have developed an English to Hungarian morpheme-based statistical machine translation method using PurePos,

2. Novák et al. [12] have annotated Old and Middle Hungarian texts employing our methods,

[3]The lexicon extension was carried out by Attila Novák [15] .

3. Endrédy et al. [39] have proposed a noun phrase detection toolkit utilizing the morphological tagging tool presented,
4. Indig and Prószéky have applied [40] the proposed tagger tool for a batch spelling-correction tool and
5. Prószéky et al. [41] have built their psycho-linguistically motivated parser on top of PurePos.

Next, Thesis group II presents methods and resources for analyzing transcripts of spoken language which can serve NLP applications of the domain. Besides, methods of Thesis II.2 estimate morpho-syntactic complexity of children language, thus can replace the labor-intense manual work. Furthermore, Jelencsik-Mátyus utilizes [42] these algorithms in her research investigating the language development of Hungarian kindergarten children.

Finally, the last (III) Thesis group details methods for processing noisy texts effectively. Algorithms of Thesis III.1 segment clinical texts accurately, providing proper output for information extraction applications. Furthermore, lessons learned from our tagging methods could help the development of accurate text mining tools in the target domain. Besides, an ongoing project [43, 44, 3] on processing Hungarian electronic health records benefits from the proposed methods.

Acknowledgement

"My help will come from the Lord, who made heaven and earth."

– Psalms 121:2

First of all, I would like to say thank you to my scientific advisor, Gábor Prószéky for guiding and supporting me over the years.

I am also thankful to Attila Novák for the fruitful conversations and his useful advice. I would like to give thanks to Nóra Wenszky as well for polishing my English and helping me to refine this study. However, I am solely responsible for all the mistakes, ambiguities and omissions that might have remained in the text.

Conversations, lunches and the always cheerful coffee breaks with my colleagues are greatly acknowledged. Thanks to László Laki, Borbála Siklósi, Balázs Indig, Kinga Mátyus for collaborating in numerous valuable studies. I am also thankful to members of room 314 István Endrédy, Győző Yang Zijian, Bálint Sass, Márton Miháltz, András Simonyi and Károly Varasdi for the friendly and intellectual atmosphere.

I am also grateful to the Pázmány Péter Catholic University and the MTA-PPKE Hungarian Language Technology Research Group, where I spent my PhD years. Thanks are due to current and former leaders Tamás Roska, Judit Nyékyné Gaizler, Péter Szolgay giving me the opportunity to conduct my research. I would like to give special thanks to Katalin Hubay and Lívia Adorján for organizing our conference trips.

Work covered in this dissertation was supported partly by the TÁMOP 4.2.1.B – 11/2/KMR-2011–0002 and 4.2.2/B – 10/1–2010–0014 projects.

Most importantly, I am thankful to my family. I cannot express enough thanks to my loving wife Jucus for tolerating my absence and encouraging me over the years. I am grateful to my parents and brother Tomi for their continuous support during my studies.

Bibliography

The author’s journal publications

[1] **György Orosz**, Attila Novák, and Gábor Prószéky. Lessons learned from tagging clinical Hungarian. *International Journal of Computational Linguistics and Applications*, 5(1):159–176, 2014. ISSN: 0976-0962.

[2] Kinga Mátyus and **György Orosz**. MONYEK: morfológiailag egyértelműsített óvodai nyelvi korpusz. *Beszédkutatás – 2014*:237–245, 2014. ISSN: 1218-8727.

[3] Borbála Siklósi, Attila Novák, **György Orosz**, and Gábor Prószéky. Processing noisy texts in Hungarian: a showcase from the clinical domain. *Jedlik Laboratories Reports*, II(3):5–62, 2014. Péter Szolgay, editor. ISSN: 2064-3942.

The author’s book section publications

[4] **György Orosz** and Kinga Mátyus. An MLU Estimation Method for Hungarian Transcripts. English. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*. Volume 8655, in Lecture Notes in Computer Science, pages 173–180. Springer International Publishing. ISBN: 978-3-319-10815-5.

[5] **György Orosz**, Attila Novák, and Gábor Prószéky. Hybrid text segmentation for Hungarian clinical records. In Félix Castro, Alexander Gelbukh, and Miguel González, editors, *Advances in Artificial Intelligence and Its Applications*. Volume 8265, in Lecture Notes in Computer Science, pages 306–317. Springer, Berlin Heidelberg, 2013. ISBN: 978-3-642-45114-0.

[6] **György Orosz**, László János Laki, Attila Novák, and Borbála Siklósi. Improved Hungarian Morphological Disambiguation with Tagger Combination. In Ivan Habernal and Václav Matousek, editors, *Text, Speech, and Dialogue*. Volume 8082, in Lecture Notes in Computer Science, pages 280–287. Springer, Berlin, Heidelberg, 2013. ISBN: 978-3-642-40584-6.

[7] László János Laki, **György Orosz**, and Attila Novák. HuLaPos 2.0 – Decoding Morphology. In Félix Castro, Alexander Gelbukh, and Miguel González, editors, *Advances in Artificial Intelligence and Its Applications*. Volume 8265, in Lecture Notes in Computer Science, pages 294–305. Springer Berlin Heidelberg, 2013. ISBN: 978-3-642-45113-3.

The author's international conference publications

[8] **György Orosz** and Attila Novák. PurePos 2.0: a hybrid tool for morphological disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*. INCOMA Ltd. Shoumen, Hissar, Bulgaria, 2013, pages 539–545.

[9] **György Orosz**, László János Laki, Attila Novák, and Borbála Siklósi. Combining Language Independent Part-of-Speech Tagging Tools. In *2nd Symposium on Languages, Applications and Technologies*. José Paulo Leal, Ricardo Rocha, and Alberto Simões, editors. In OpenAccess Series in Informatics (OASIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Porto, 2013, pages 249–257.

[10] **György Orosz** and Attila Novák. PurePos – an open source morphological disambiguator. In *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*. Bernadette Sharp and Michael Zock, editors. Wroclaw, 2012, pages 53–63.

[11] László Laki and **György Orosz**. An Efficient Language Independent Toolkit for Complete Morphological Disambiguation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pages 26–31.

[12] Attila Novák, **György Orosz**, and Nóra Wenszky. Morphological annotation of Old and Middle Hungarian corpora. In *Proceedings of the ACL 2013 workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Piroska Lendvai and Kalliopi Zervanou, editors. Association for Computational Linguistics, Sofia, Bulgaria, 2013, pages 43–48.

[13] Borbála Siklósi, **György Orosz**, Attila Novák, and Gábor Prószéky. Automatic structuring and correction suggestion system for Hungarian clinical records. In *8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages*. Guy De Pauw, Gilles-Maurice de Schryver, Mike L. Forcada, Francis M. Tyers, and Peter Waiganjo Wagacha, editors. Istanbul, 2012, pages 29–34.

The author's other conference publications

[14] **György Orosz** and Gábor Prószéky. Hol a határ? Mondatok, szavak, klinikák. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2014, pages 177–187.

[15] **György Orosz** and Attila Novák. PurePos 2.0: egy hibrid morfológiai egyértelműsítő rendszer. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2014, pages 373–377.

[16] **György Orosz**, Attila Novák, and Gábor Prószéky. Magyar nyelvű klinikai rekordok morfológiai egyértelműsítése. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2013, pages 159–169.

[17] **György Orosz**. PurePos: hatékony morfológiai egyértelműsítő. In *VI. Alkalmazott Nyelvészeti Doktoranduszkonferencia*. Tamás Váradi, editor. Budapest, 2012, pages 134–139.

[18] **György Orosz**, Attila Novák, and Balázs Indig. Javában taggelünk. In *VIII. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács and Veronika Vincze, editors. Szeged, 2011, pages 336–340.

[19] László János Laki and **György Orosz**. HuLaPos2 – Fordítsunk morfológiát. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2014, pages 41–49.

[20] László János Laki and **György Orosz**. Morfológiai egyértelműsítés nyelvfüggetlen annotáló módszerek kombinálásával. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2013, pages 331–337.

[21] Borbála Siklósi, **György Orosz**, and Attila Novák. Magyar nyelvű klinikai dokumentumok előfeldolgozása. In *VIII. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács and Veronika Vincze, editors. Szeged, 2011, pages 143–154.

Other references

[22] Dóra Csendes, János Csirik, and Tibor Gyimóthy. The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora LINC 2004 at The 20th International Conference on Computational Linguistics COLING 2004*, 2004, pages 19–23.

[23] Attila Novák and Nóra Wenszky. Ó- és középmagyar szóalaktani elemző. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács and Veronika Vincze, editors. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged, 2013, pages 170–181.

[24] Gábor Prószéky and Balázs Kis. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. Robert Dale and Kenneth Ward Church, editors. ACL, College Park, Maryland, 1999, pages 261–268.

[25] Attila Novák. Milyen a jó humor? In *Magyar Számítógépes Nyelvészeti Konferencia 2003*. Szeged, 2003, pages 138–145.

[26] Gábor Prószéky and Attila Novák. Computational Morphologies for Small Uralic Languages. In *Inquiries into Words, Constraints and Contexts.* Stanford, California, 2005, pages 150–157.

[27] János Zsibrita, Veronika Vincze, and Richárd Farkas. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *Proceedings of Recent Advances in Natural Language Provessing 2013*. Association for Computational Linguistics. Hissar, Bulgaria, 2013, pages 763–771.

[28] Lawrence Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[29] Christer Samuelsson. Morphological tagging based entirely on Bayesian inference. In *9th Nordic Conference on Computational Linguistics NODALIDA-93*. Stockholm University, Stockholm, Sweden, 1993.

[30] Thorsten Brants. TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*. Universität des Saarlandes, Computational Linguistics. Association for Computational Linguistics, 2000, pages 224–231.

[31] Péter Halácsy, András Kornai, and Csaba Oravecz. HunPos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Prague, Czech Republic, 2007, pages 209–212.

[32] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[33] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009. ISSN: 19310145.

[34] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[35] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 2006.

[36] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: open source scientific tools for Python. [Online; accessed 2014-11-26]. 2001–. URL: `http://www.scipy.org/`.

[37] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edition, 2011, page 629. ISBN: 978-0-12-374856-0.

[38] Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004, pages 1201–1204.

[39] István Endrédy. Corpus driven research: ideas and attempts. In *PhD Proceedings Annual Issues of the Doctoral School - 2014*. Faculty of Information Technology and Bionics, Pázmány Péter Catholic University., Budapest, Hungary, 2014, 137–140.

[40] Balázs Indig and Gábor Prószéky. Ismeretlen szavak helyes kezelése kötegelt helyesírás-ellenőrző programmal. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2013, pages 310–317.

[41] Gábor Prószéky, Balázs Indig, Márton Miháltz, and Bálint Sass. Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2014, pages 79–87.

[42] Kinga Jelencsik-Mátyus. A szociolingvisztikai stílus: Stratégiák a gyermek-felnőtt diskurzusban. PhD thesis. Szeged, Hungary: University of Szeged, 2015, page 192.

[43] Borbála Siklósi and Attila Novák. Identifying and clustering relevant terms in clinical records using unsupervised methods. In Laurent Besacier, Adrian-Horia Dediu, and Carlos Martín-Vide, editors, *Statistical language and speech processing*. Volume 8791, in Lecture Notes in Computer Science, pages 233–243. Springer International Publishing, 2014. ISBN: 978-3-319-11396-8.

[44] Borbála Siklósi and Attila Novák. A magyar beteg. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2014, pages 188–198.