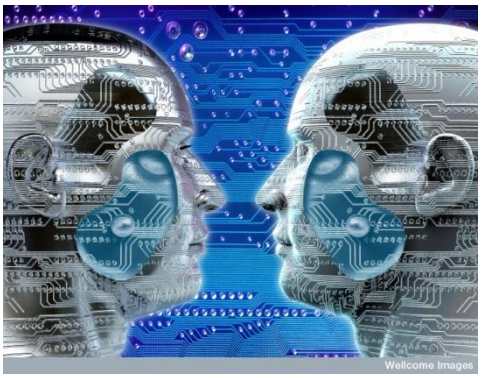
4. NATURAL LANGUAGE PROCESSING



GÁBOR PRÓSZÉKY, Professor; BORBÁLA SIKLÓSI, Assistant Professor; LÁSZLÓ LAKI, PhD; ATTILA NOVÁK, PhD
PhD Students: ISTVÁN ENDRÉDY, BALÁZS INDIG, GYŐZŐ ZIJIAN YANG

SHORT DESCRIPTION OF THE ACTIVITIES

The Natural Language Processing Group consists of faculty researchers, post-graduate researchers, PhD students, undergraduate students and programmers who work together developing algorithms that enable computers to process and understand human languages. Our research interest covers:

- corpus linguistic applications
- statistical machine translation
- syntactic parsing
- medical text mining
- morphologies
- spelling correction
- part-of-speech tagging

One of the most ambitious aims of the research group is to develop new methods and algorithms for syntactic parsing of the Hungarian language. Such a method must handle grammatically possible, but not correct analyses. That is why to deal with problems efficiently, *parallelism* is necessary. In practice for a human understanding is cooperatively done by several parts of the brain. To incorporate this knowledge the consideration of current state of the neurolinguistics and psycholinguistics is indispensable. The model that is to be researched is characterized by *performance*, while the state-of-the-art research results are considered from various field of applied linguistics. Since the current state of research does not provide any deeper understanding of how the ambiguous phrases are understood, in our project we incorporate parallel corpora to handle these—not necessarily multilingual—problems. With this, *new aspects of corpus linguistic* research are being revealed. The developed new methods are adapted to many aspects of Hungarian language. For using it for medical text processing tasks, we expect growing performance and precision. The developed algorithms are also planned to be adapted to other agglutinating languages and are expected to behave similarly well.

SELECTED PUBLICATIONS

[1] Borbála Siklósi, Attila Novák, Gábor Prószéky (2016): Context-aware correction of spelling errors in Hungarian medical documents. *Computer Speech & Language*, Vol.35, pp. 219─233.

[2] Endrédy István, Indig Balázs (2015): HunTag3, a general-purpose, modular sequential tagger – chunking phrases in English and maximal NPs and NER for Hungarian In: Zygmunt Vetulani, Joseph Mariani (szerk.) *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. pp. 213─218

[3] Attila Novák (2015): Making morphologies the `easy' way, In: A. Gelbukh (ed.) *Lecture Notes in Computer Science Volume 9041: Computational Linguistics and Intelligent Text Processing* Springer International Publishing, Berlin–Heidelberg. Part I pp. 127─138.

[4] Borbála Siklósi, Attila Novák (2014): Identifying and Clustering Relevant Terms in Clinical Records Using Unsupervised Methods. In: Besacier, L.; Dediu, A.-H. and Martín-Vide, C. (Eds.), *Lecture Notes in Computer Science Volume 8791: Statistical Language and Speech Processing* Springer International Publishing, Berlin Heidelberg. pp. 233─243

[5] György Orosz, Attila Novák, Gábor Prószéky (2014): Lessons learned from tagging clinical Hungarian. *International Journal of Computational Linguistics and Applications*, Vol. 5 no. 2.

[6] László János Laki, Attila Novák, Borbála Siklósi, György Orosz (2013): Syntax-based reordering in phrase-based English-Hungarian statistical machine translation. *International Journal of Computational Linguistics and Applications*, Vol. 4 no. 2. pp. 63─78.

[7] István Endrédy, Attila Novák (2013): More effective boilerplate removal – the GoldMiner algorithm. *Polibits 48*. pp. 79─83.

[8] György Orosz, László János Laki, Attila Novák Borbála Siklósi (2013): Improved Hungarian Morphological Disambiguation with Tagger Combination. In: Habernal, Ivan; Matousek, Vaclav (eds.) *Lecture Notes in Computer Science, Vol. 8082: Text, Speech, and Dialogue, 16th International Conference, TSD 2013.* Pilsen, Czech Republic. Springer, Berlin–Heidelberg. pp. 280─287.

[9] László János Laki, György Orosz, Attila Novák (2013): HuLaPos 2.0 – Decoding morphology. In: F. Castro, A. Gelbukh, M.G. Mendoza (eds.) *Lecture Notes in Computer Science, Vol. 8265: Advances in Artificial Intelligence and Its Applications*. Springer, Berlin–Heidelberg. pp. 294─305.

[10] György Orosz, László János Laki, Attila Novák, Borbála Siklósi (2013): Combining Language-Independent Part-of-Speech Tagging Tools. In: J. P. Leal, R. Rocha, and A. Simoes (eds.) *2nd Symposium on Languages, Applications and Technologies*. Porto: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. pp. 249─257

[11] Gábor Prószéky, Csaba Merényi (2012): Language Technology Methods Inspired by an Agglutinative, Free Phrase-Order Language. In: Walther von Hahn, Cristina Vertan (eds.) *Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation*. Cambridge: Cambridge University Press