# Natural Language Processing by Utilizing Crowds

Jozef HARINEK*

*Slovak University of Technology in Bratislava*
*Faculty of Informatics and Information Technologies*
*Ilkovičova 2, 842 16 Bratislava, Slovakia*
j.harinek@gmail.com

The amount of information stored in natural language on the web is huge and still growing. In order to make a better use of this information, we need to process the natural language and transform it to a form that machines are capable of understanding. This is performed by Natural Language Processing (NLP).

However, NLP is a difficult task. It has several levels in which it can be performed [1]. The first and most basic one is to recognize sounds. Then it goes to recognition of speech, in which the machine can divide sounds into words. Next step in NLP is morphological analysis. Here are the words analyzed and their grammatical categories are extracted. Next layer in analysis is syntactic layer. In this layer, the syntax of the text is recognized. It is represented by sentence components and relations between them. Another layer is morphemic layer which adds information about morphemic structure of the words. Last two layers are semantic and context layer, in which semantic structure is identified and put into context of the given text.

In our work we are focusing on syntactic analysis of Slovak language. This is a field that is still growing and much work is to be done, due to its specificity and grammatical structure [2]. To support NLP in Slovak language, we also plan to employ principles of crowdsourcing.

Crowdsourcing is a process, in which one uses the power of crowd to perform a task that is typically large and needs lots of experts to be involved [3]. To be successful, the crowd needs to be motivated and given tasks need to be small enough to be performed by non-professionals. There are several ways of motivating people. They can be motivated by a financial reward, by enjoyment they experience during fulfillment of the task (Games with a purpose), by added value of performing the task (they learn something), etc. [3]

In our work, we want to use the methods of crowdsourcing to help us annotate large scale texts, a task that normally needs lots of experts and time. We plan to verify our method in a software tool, which is specially designed to support education in

* Supervisor: Doctor Marián Šimko, Institute of Informatics and Software Engineering

*Spring 2014 PeWe Workshop, March 21, 2014, pp. 95–96.*

elementary schools. Children will perform sentence analysis assignments given by teacher. We also plan to verify data, that was manually annotated, obtained from Ľudovít Štúr Institute of Linguistics at the Slovak Academy of Sciences.

*Acknowledgement.* This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

## References

[1] Cimiano, Philipp. Ontology learning from text. Springer US, 2006.
[2] Čižmár, A., Juhár, J., Ondáš, S. (2010): Extracting sentence elements for the natural language understanding based on slovak national corpus, in Proceedings of the International Conference on Analysis of Verbal and Nonverbal Communication and Enactment, COST'10, LNCS Vol. 6800, Springer, 2010 pp. 171-177.
[3] Quinn, Alexander J., and Benjamin B. Bederson. "Human computation: a survey and taxonomy of a growing field." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2011.