Serbian-Slovakian bilateral project proposal for period 2017-2018.

INTRODUCTION - short description of the project:
The project is focused on quantitative analysis of syllables in Slavic languages, namely, in Russian, Slovak, and Serbian. These three languages represent three geographical groups of Slavic languages (East, West, and South). Syllables, as opposed to other language units, have not been mathematically modelled systematically, the main reason being problems with their definition (i.e., with word syllabification). The aim of the project is to fill this gap. The syllabification can be performed algorithmically, the approach makes use, among others, of statistical tests. In particular, we will focus on models for syllable frequency and syllable length in the three abovementioned languages. We will work with orthographic input of texts. It is expected that the new models will be related to the already known models for grapheme frequencies and word length.

OBJECTIVES - indicate the objectives and the scientific and/or educational basis of the project (highlight reasons for collaborating with the foreign partner):
The project is of a theoretical and empirical character. Supposed results will contribute to a fuller integration of the syllabic level of the language into the model suggested by Köhler (2005) and Kelih (2012). The above mentioned problem of missing (technical) procedures for the determination of exact syllable boundaries (i.e., splitting words into syllables) will be solved by following the suggestion presented by Pulgram (1970), which was later substantially improved by Lehfeldt (1971) and by Kempgen (2003). Within the project, a semi-automatic algorithm for the syllabification of Slavic texts (in particular for Russian, Slovak, and Serbian) will be developed, based on an orthographical input of texts. The algorithm is conceptually based on an essential structural property of syllables, namely their graphotactical/phonotactical “behaviour”.

METHODOLOGY - indicate in an analytical form the research and/or educational stages, emphasizing the roles played by the Serbian and Slovakian research units:
Based on the algorithm developed within this project, we will develop a computer program which will provide syllable boundaries within words in a text as its output, with a possibility of an automatic creating of the list of syllables occurring in the text, their frequencies, lengths and other properties. The language material which will be analyzed consists of the Russian novels “Kak zakaljalas' stal'” by N. Ostrovsky and “Master and Margerita” by M. Bulgakov, and their translations into Slovak and Serbian (see Kelih 2009). Principal investigators both from the Slovak and the Serbian side closely cooperate with Emmerich Kelih (University of Vienna), who is the author of the two parallel Slavic corpora, hence the texts are available and the applicants have experience in working with them. The analysis of word-final and word-initial clusters will be based on the balanced corpora of Russian, Slovak, and Serbian from the abovementioned Quanta-project. The analysis of parallel corpora seems to be a reliable way to achieve a maximal homogeneity of the data taken from the three Slavic languages.

RESOURCES - indicate the Serbian and Slovakian financial and human resources (title and reference number of national research project):
I. Obradović is participant of the national project 178006 "Serbian language and its resources: the theory, description and application", which deals with related issues. When it comes to human resources, given the interdisciplinary character of the project, the team presents an ideal mixture of two researchers who work both in the areas of mathematics and linguistics (J. Mačutek and I. Obradović), two PhD students in mathematics, with two different specializations (M. Koščová investigates discrete probability distributions, which belong to the most common models applied to linguistic data; M. Radojičić will be responsible for the development and realization of algorithms) and one PhD student at the Faculty of philology (B. Lazić). The team is balanced as far as research experience of its members is concerned. It includes two established scientists and three PhD students.

EXPECTED RESULTS - indicate the expected results with particular regard to technological transfer and/or development of human resources, impact on scientific and technological relations:
This project presents a possibility to further strengthen the collaboration of two principal investigators and to build a systematic, unified approach to this (so far quite fragmented) research area. Both of them have many contacts among linguistists, and belong to a small group of authors who published papers on mathematical

modelling of syllable properties. J. Mačutek was awarded the prestigious Lise Meitner Grant from the Austrian research-funding agency FWF, and he worked at the Department of Slavic Studies at the University in Graz (January 2009 – December 2010, followed then by another six-month project), hence, he has direct experience with working in a linguistic research team I. Obradović has been working on applications of mathematics in linguistics for a long time as well. Given the state of the art in quantitative analysis of syllables, expected results (four scientific papers, a presentation at an international conference in 2018, and a freely available software) are likely to attract a wide attention from researchers working in the field of quantitative linguistics. Consequently, the results will be cited. The members of the proposed research team and their universities/faculties will profit from such an international attention given to their results.

COLLABORATIONS - already developed with the partner Country - if any:
Both principal researchers are active members of the International Quantitative Linguistics Association (IQLA), and initial collaboration between them was established within the scope of this organization and the IQLA conferences (QUALICO). J. Mačutek has been the Treasurer of IQLA since 2009, and I. Obradović was an IQLA council member from 2012 to 2014 and the main organizer of the QUALICO conference in 2012 in Belgrade.

INTRODUCTION - BIBLIOGRAPHY - Indicate max 5 publications relevant to the project:
1. Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) (2005). Quantitative Linguistics. An International Handbook. Berlin, New York: de Gruyter. 2. Grzybek, P. (ed.) (2007). Contributions to the Science of Text and Language. Word Length Studies and Related Issues. Dordrecht: Kluwer. 3. Kelih, E., Levickij, V, Altmann, G. (eds.), Methods of Text Analysis: 106-124. Chernivci: ChNU. 4. Obradović, I., Obuljen, A., Vitas, D., Krstev, C., Radulović, V. (2010). Canonical syllable types in Serbian. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives: 145-157. Wien: Praesens. 5. Kelih, E., Mačutek, J. (2012). Number of canonical syllable types: a continuous bivariate model. Journal of Quantitative Linguistics 20, 241-251.

Date:
03.07.2016

Signature of principal investigator:

Professor Dušan Polomčić

Dean of University of Belgrade Faculty of Mining and Geology