

# **Предлог билатералног пројекта Србија-Словачка за циклус 2017-2018.**

## **УВОД - кратак опис пројекта:**

Фокус пројекта је на квантитативној анализи и математичком моделирању слогова у словенским језицима, наиме, руском, словачком и српском језику. Ова три језика представљају три географске групе словенских језика - источну, западну и јужну. Слогови, за разлику од других јединица језика, нису математички моделирани на систематски начин, а главни разлог је проблем са њиховом дефиницијом. Циљ пројекта је да попуни ту празнину. Силабификација се може обављати алгоритамски, при чему овај приступ користи, између осталог, статистичке тестове. У пројекту ћемо се посебно бавити моделима фреквенције слогова и дужине слогова у три горе поменута језика. При томе ће бити коришћени ортографски уноси текста. Очекује се да ће модели слогова развијени у овом пројекту кореспондирати са већ познатим моделима за фреквенције графема и дужине речи.

## **ЦИЉЕВИ - навести разлоге за билатералну сарадњу - комплементарност пројектних тимова:**

Пројекат је теоријског и емпиријског карактера. Очекивани резултати ће допринети потпунијој интеграцији слоговог нивоа у модел који су предложили Келер (2005) и Келих (2012). Проблем недостајућих (техничких) процедура за утврђивање тачних граница слогова (тј. поделе речи на слокове) ће бити решен на основу предлога који је изнео Пулграм (1970), а који су касније знатно побољшали Лехфелдт (1971) и Кемпген (2003). У оквиру пројекта биће развијен полуаутоматски алгоритам за силабификацију словенских текстова (посебно руски, словачки, и српски), на основу ортографског уноса текстова. Алгоритам ће се концепцијски заснивати на суштинским структурним особинама слогова, односно њиховом фонетском и графичком "понашању".

## **МЕТОДОЛОГИЈА - навести у аналитичкој форми актуелност пројекта и претходна истраживања у научној области са нагласком на улогу српског и словачког истраживачког тима:**

На основу алгоритма развијеног у оквиру овог пројекта, биће развијен рачунарски програм који ће одређивати границе слогова унутар речи у тексту, са могућношћу аутоматског креирања листе слогова који се јављају у тексту, њихове фреквенције, дужине и других особина. Језички материјал који ће бити анализиран чине руски романи "Како се калио челик" Н. Островског и "Мајстор и Маргарита" од М. Булгакова, и њихови преводи на словачки и српски језик (Келих 2009). Главни истраживачи, како са словачке тако и са српске стране, тесно сарађују са Емерихом Келихом (Универзитет у Бечу), који је аутор два паралелна словенска корпуса, те су им поменути текстови доступни, а главни истраживачи имају искуства у раду са корпусима. Анализа почетних и финалних кластера речи ће се заснивати на уравнотеженом корпусу руског, словачког, и српског из Quanta-пројекта. Полази се од претпоставке да анализа паралелног корпуса представља поуздан начин да се постигне максимална хомогеност података преузетих на три словенских језика.

## **ИЗВОРИ ФИНАНСИРАЊА - навести број националног пројекта који представља основу истраживања и/или друге евентуалне изворе финансирања:**

И. Обрадовић је учесник националног пројекта 178006 "Српски језик и његови ресурси: теорија, опис и примене", који обрађује сродну проблематику, те ће овај пројекат бити комплементаран са пројектом 178006. Када су у питању људски ресурси, с обзиром на интердисциплинарни карактер пројекта, пројектни тим представља идеалну комбинацију: два истраживача који раде у области математике и лингвистике (Ј. Мачутек и И. Обрадовић), два студента докторских студија из области математике, са две различите специјализације (М. Кошчова истражује дискретне расподеле вероватноће, које представљају моделе који се најчешће примењују на језичке податаке, а М. Радојичић ће бити одговорна за развој и реализацију алгоритама), као и једног студента докторских студија на Филолошком факултету (Б. Лазих). Тим је уравнотежен када је у питању истраживачко искуство његових чланова, два призната научника и три студента докторских студија.

## **ОЧЕКИВАНИ РЕЗУЛТАТИ - навести очекиване резултате са посебним нагласком на трансфер технологије и/или развој људских капацитета, евентуално пријављивање заједничког**

### **Хоризонт2020 пројекта, очекивана индустријска или практична корист пројектних резултата:**

Овај пројекат отвара могућност за даље јачање сарадње две главна истраживача и за изградњу систематског, обједињеног приступа овој (до сада прилично фрагментираној) истраживачкој области. Оба главна истраживача имају бројне контакте међу лингвистима, и припадају малој групи аутора који су објавили радове о математичком моделирања својстава слогова. Ј. Мачутек је добио престижну Lise Meitner стипендију од Аустријске агенције за финансирање истраживања FWF, и радио је у Одељењу за славистику на Универзитету у Грацу (јануар 2009 - децембар 2010, а затим и у оквиру једног другог шестомесечног пројекта), те има непосредно искуство у раду тима за лингвистичка истраживања а И. Обрадовић је такође радио на примени математичких метода у лингвистици већ дуже време. С обзиром на тренутно стање у квантитативној анализи слогова, очекује се да ће планирани резултати (четири научна рада, презентација на међународној конференцији у 2018, и слободно доступан софтвер) привући пажњу истраживача који раде у области квантитативне лингвистике, и да ће резултати пројекта бити цитирани. Чланови предложеног истраживачког тима и њихови универзитети/факултети ће имати користи од таквог међународног одјека њихових резултата.

### **ВРСТА САРАДЊЕ - А) успостављање сарадње први пут, или Б) развој раније успостављене сарадње**

Оба главна истраживача су активни чланови Међународног удружења за квантитативну лингвистику (IQLA), а сарадња између њих је и започела у оквиру ове организације и IQLA конференција (QUALICO). Ј. Мачутек је благајник IQLA од 2009. године, а И. Обрадовић је био члан савета IQLA од 2012. до 2014. године и главни организатор конференције QUALICO 2012. године у Београду.

### **БИБЛИОГРАФИЈА - наведите 5 најзначајних референци релевантних за пројекат: научни радови у референтним међународним часописима; књиге или поглавља у књигама међународног значаја, патенти:**

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.

**Датум:**

03.07.2016

**Потпис руководиоца пројекта:**



Проф. др Душан Полоччић

Декан Рударско-геолошког факултета Универзитета у Београду