SK-SRB-2016-0017

A1		Základné informácie o projekte Project basic information
01	Evidenčné číslo projektu Project ID	SK-SRB-2016-0017
02	Dátum podania Date of submission	28.06.2016 17:02
03	Názov projektu	Kvantitatívna analýza slabík v slovanských jazykoch (ruština, slovenčina, srbčina)
	Project title	Quantitative analysis of syllables in Slavic languages (Russian, Slovak, Serbian)
04	Akronym projektu	QASSL
	Acronym of the project	QASSL
05	Odbor výskumu a vývoja	10199 - Ostatné príbuzné odbory matematických vied
	R&D specialization	10199 - Other disciplines related to mathematical sciences
06	Začiatok riešenia projektu Project start	01.01.2017
07	Koniec riešenia projektu Project end	31.12.2018
08	Anotácia	

Projekt je zameraný na kvantitatívnu analýzu slabík v slovanských jazykoch, konkrétne v ruštine, slovenčine a srbčine. Tieto tri jazyky reprezentujú tri geografické skupiny slovanských jazykov (východné, západné, južné). Slabiky na rozdiel od iných jazykových jednotiek zatiaľ neboli systematicky matematicky modelované, hlavným dôvodom sú problémy spojené s ich definovaním (s rozdelením slova na slabiky). Cieľom projektu je zaplniť túto medzeru, pričom na určenie hraníc medzi slabikami sa dá použiť algoritmický postup, ktorý okrem iného využíva aj štatistické testy. Konkrétne sa sústredíme na modelovanie početností slabík a dĺžky slabík v troch vyššie spomenutých jazykoch, pričom budeme pracovať na úrovni grafém. Predpokladáme, že tieto modely budú súvisieť s už známymi modelmi pre početnosti grafém a pre dĺžku slov.

Annotation

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.

09	Žiadateľ	Univerzita Komenského v Bratislave
	Applicant	Comenius University in Bratislava
10	Požadované finančné prostriedky z APVV v EUR Required budget from the agency in EUR	4 482,00

SK-SRB-2016-0017

11	Celkové náklady na projekt v EUR Total project budget in EUR	4 482,00

SK-SRB-2016-0017

A2		**Základné informácie o riešiteľských organizáciách** **Applicant basic information**
Žiadateľ		
Applicant		
01	Názov organizácie	Univerzita Komenského v Bratislave - Fakulta matematiky, fyziky a informatiky
	Name of the organization	Comenius University in Bratislava - Faculty of Mathematics Physics and Informatics
02	Adresa organizácie / Organization address	Šafárikovo námestie 6, 81499 Bratislava,
03	IČO / ID	00397865
04	Príslušnosť k rezortu	MŠVVaŠ SR
	Governmental branch	Education
05	Typ organizácie	Žiadateľ
	Organization type	Applicant
06	Forma hospodárenia	vysoká škola
	Form of economy	higher-education institution
07	Telefón / Phone	02/6542 7086
	Fax	02/6542 5882
	E-mail	so@fmph.uniba.sk
08	Štatutárny zástupca I / Statutory representative I	prof., RNDr. Karol Mičieta, PhD.
09	Štatutárny zástupca II / Statutory representative II	

SK-SRB-2016-0017

A3		**Zoznam všetkých riešiteľov** **List of all participants**		
01		Zoznam slovenských a partnerských zamestnancov priamo sa podieľajúcich na riešení projektu List of Slovak and partner participants involved in project		
Meno a Priezvisko Name and Surname		Tituly Titles	Pracovné zaradenie Position	Rola v RK Participant role
Ján Mačutek		doc. Mgr. PhD.	docent	Zodpovedný riešiteľ
Michaela Koščová		Mgr.	doktorand	Doktorand
Ivan Obradović		prof.	profesor	Zodpovedný riešiteľ partnerskej organizácie
Marija Radojičić				Doktorand
Biljana Lazić				Doktorand
02	Ostatní zamestnanci Other staff	Celkový počet ostatných osôb Total number of other staff		
03	Spolu Total	Celkový počet zamestnancov Total number of involved staff		5

SK-SRB-2016-0017

A4		Informácie o slovenskom zodpovednom riešiteľovi a prehľad jeho výstupov odbornej činnosti Information about Slovak Principal Investigator and the overview his/her outputs
01	Meno a priezvisko Name and surname	doc., Mgr. Ján Mačutek, PhD.
02	Pohlavie	Muž
	Gender	Male
03	Telefón Phone	+421 2 60295717 +421 914 166578
04	Email	jmacutek@yahoo.com

SK-SRB-2016-0017

A4	Informácie o slovenskom zodpovednom riešiteľovi a prehľad jeho výstupov odbornej činnosti / Information about Slovak Principal Investigator and the overview his/her outputs
05	Publikácie v zahraničných a domácich periodikách pokrytých CC za posledných 5 rokov / CC publications in the foreign and domestic periodicals in the last 5 years

1. Ferrer-i-Cancho, R., Baixeries, J., Hernández-Fernández, A., Dębowski, Ł., Mačutek, J. (2014). When is Menzerath-Altmann law mathematically trivial? A new approach. Statistical Applications in Genetics and Molecular Biology 13, 633-644.
2. Čech, R., Mačutek, J., Žabokrtský, Z. (2011). The role of syntax in complex networks: local and global importance of verbs in a syntactic dependency network. Physica A: Statistical Mechanics and its Applications 390, 3614-3623.

	Počet / Number
2	

06	Publikácie v zahraničných a domácich periodikách nepokrytých CC za posledných 5 rokov / Non-CC publications in the foreign and domestic periodicals in the last 5 years

Edited volumes
1. Mikros, G.K., Mačutek, J. (eds.) (2015). Sequences in Language and Text. Berlin, Boston: de Gruyter. 260 pp.
2. Tuzzi, A., Benešová, M., Mačutek, J. (eds.) (2015). Recent Contributions to Quantitative Linguistics. Berlin, Boston: de Gruyter. 284 pp.
3. Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.) (2014). Empirical Approaches to Language and Text Analysis. Lüdenscheid: RAM-Verlag. 231 pp.
Refereed papers
4. Čech, R., Mačutek, J., Liu, H. (2016). Syntactic complex networks and their applications. In: Mehler, A., Lücking, A., Banisch, S., Blanchard, P., Frank-Job, B. (eds.), Towards a Theoretical Framework for Analyzing Complex Linguistic Networks: 167-186. Berlin, Heidelberg: Springer.
5. Mačutek, J., Koščová, M., Čech, R. (2016). Lexical compactness across genres in works by Karel Čapek. In: Mayaffre, D., Poudat, C., Vanni, L., Magri, V., Follette, P. (eds.), Statistical Analysis of Textual Data: 825-832. Nice: University Nice Sophia Antipolis.
6. Čech, R., Mačutek, J., Koščová, M. (2015). On the relation between verb full valency and synonymy. In: Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015): 68-73. Uppsala: Uppsala University.
7. Mačutek, J. (2015). Type-token relation for word length motifs in Ukrainian texts. In: Tuzzi, A., Benešová, M., Mačutek, J. (eds.), Recent Contributions to Quantitative Linguistics: 63-73. Berlin, Boston: de Gruyter.
8. Mačutek, J., Melicherová, B. (2015). Automatic classification of Ukrainian texts based on distances between words of equal length. Mathematical Linguistics 1(1), 57-69.
9. Mačutek, J., Mikros, G.K. (2015). Menzerath-Altmann law for word length motifs. In: Mikros, G.K., Mačutek, J. (eds.). Sequences in Language and Text: 125-131. Berlin, Boston: de Gruyter.
10. Mačutek, J. (2014). Complex networks are not (so much) privileged: Comment on “Approaching human language with complex networks” by Cong and Liu. Physics of Life Reviews 11, 635-636.
11. Mačutek, J., Wimmer, G. (2014). A measure of lexical text compactness. In: Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.), Empirical Approaches to Language and Text Analysis: 132-139. Lüdenscheid: RAM-Verlag.
12. Kelih, E., Mačutek, J. (2013). Number of canonical syllable types: a continuous bivariate model. Journal of Quantitative Linguistics 20, 241-251.
13. Mačutek, J., Čech, R. (2013). Frequency and declensional morphology of Czech nouns. In: Obradović, I., Kelih, E., Köhler, R. (eds.), Methods and Applications of Quantitative Linguistics: 59-68. Belgrade: Akademska Misao.
14. Mačutek, J., Wimmer, G. (2013a). Alternative methods of goodness-of-fit evaluation applied to word length data. In: Köhler, R., Altmann, G. (eds.), Issues in Quantitative Linguistics 3: 282-290. Lüdenscheid: RAM-Verlag.
15. Mačutek, J., Wimmer, G. (2013b). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. Journal of Quantitative Linguistics 20, 227-240.
16. Wimmer, G., Mačutek, J. (2012). New integrated view at partial-sums distributions. Tatra Mountains Mathematical Publications 51, 183-190.

SK-SRB-2016-0017

17. Mačutek, J. (2011). Regularity of rhythmic patterns in examples from Slovak poetry. In: Scherr, B.P., Bailey, J., Kazartsev, E.V. (eds.), Formal Methods in Poetics: 306 313. Lüdenscheid: RAM-Verlag.
18. Mačutek, J., Rovenchak, A. (2011). Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In: Kelih, E., Levickij, V., Matskulyak, Y. (eds.), Issues in Quantitative Linguistics 2: 136-147. Lüdenscheid: RAM-Verlag.
19. Mačutek, J., Švehlíková, Z., Cenkerová, Z. (2011). Towards a model for rank-frequency distributions of melodic intervals. Glottometrics 21, 60-64.

	Počet / Number
19	

07	Aplikačné výstupy za posledných 5 rokov / Application outputs in the last 5 years (uveďte aj v anglickom jazyku)

žiadne, riešiteľ sa venuje základnému teoretickému výskumu
none, the applicant is envolved in basic theoretial research

	Počet / Number

08	Riešené projekty a iné výstupy za posledných 5 rokov / Supported projects and other outputs in the last 5 years (uveďte aj v anglickom jazyku)

01/2015ff. Discrete and continuous probabilistic models and their applications (Slovakia, VEGA 2/0047/15)
01/2012-12/2014 New Methods of Mathematical Statistics (Slovakia, VEGA 2/0038/12)
10/2012-02/2014 Linguistic and lexicostatistic analysis in cooperation of linguistics, mathematics, biology and psychology (EU – Czech Republic, ESF)
01/2011-06/2011 Length Motifs in Slovak Texts (Austria - Slovakia, Aktion Österreich-Slowakei, Ernst Mach Grant)

	Počet / Number
4	

SK-SRB-2016-0017

A5	Informácie o partnerskej organizácii / Information about partner organization
01	Názov partnerskej organizácie / Name of partner organization (uveďte aj v anglickom jazyku)
Univerzita v Belehrade University of Belgrade	
02	Adresa partnerskej organizácie / Address of partner organization
University of Belgrade Studentski trg 1 11000 Beograd Serbia	
03	Zoznam členov riešiteľského kolektívu partnerskej organizácie / List of participants of partner organization
prof. Ivan Obradović Biljana Lazić (PhD student) Marija Radojičić (PhD student)	
04	Meno, priezvisko a podpis zodpovedného riešiteľa partnerskej organizácie / Name, surname and signature of Principal Investigator of partner organization
prof. Ivan Obradović	
05	Meno, priezvisko a podpis štatutárneho zástupcu partnerskej organizácie / Name, surname and signature of Statutory Representative of partner organization
prof. Dušan Polomčić, dean	
06	Čestne vyhlasujem, že všetky informácie obsiahnuté v žiadosti sú pravdivé. / I, hereby declare that, all information in the application concerning the partner organization is true.
Áno / Yes	

SK-SRB-2016-0017

A6	Informácie o spolupracujúcich organizáciách / Information about cooperating organizations
01	Zoznam slovenských spolupracujúcich organizácii a k ním prislúchajúci členovia riešiteľského kolektívu / List of the Slovak cooperating organizations within the Slovak project participants (uveďte aj anglický názov slovenských spol. organizácií)

SK-SRB-2016-0017

B	Charakteristika projektu
01	Konkrétne ciele, originálnosť a aktuálnosť výskumného zámeru (maximálne 15 000 znakov)

Tento projekt patrí do oblasti kvantitatívnej lingvistiky. Je interdisciplinárneho charakteru; spája v sebe matematiku, najmä štatistiku, s lingvistikou a automatickým spracovaním textu. Kvantitatívna lingvistika sa ako vedná disciplína zaoberá konštruovaním teórie jazyka a jazykových procesov, pričom používa aparát matematického modelovania a štatistických testov. Köhler et al. (2005) poskytuje súhrn výsledkov dosiahnutých týmito metódami v nie príliš vzdialenej minulosti.
Jeden z najnádejnejších prístupov ku konštrukcii teórie jazyka vo vyššie uvedenom zmysle je využitie princípu synergie (pozri Köhler 2005). Je založený na predpoklade, že jednotlivé úrovne jazyka (napr. fonémy/grafémy, slabiky/morfémy, slová, klauzy, vety, atď.) a ich vlastnosti nie sú navzájom nezávislé, ale naopak, navzájom sa ovplyvňujú. Väčšina z týchto úrovní bola buď pomerne dôkladne preskúmaná (pre grafémy pozri napr. Grzybek et al. 2009 a Grzybek a Rusko 2009 a zoznam citovanej literatúry v týchto článkoch, pre slová pozri Popescu et al. 2009), alebo sa dajú nájsť aspoň čiastkové, aj keď nie systematické výsledky (napr. články Best 2005a a Best 2005b sú prehľadom výsledkov dosiahnutých pre morfémy a pre vety).
Úroveň slabík je však, čo sa týka kvantitatívneho prístupu k lingvistike, takmer nedotknutá. Je zrejmé, že hlavným dôvodom je postavenie slabiky, ktoré je v lingvistike predmetom sporov. Na jednej strane je slabika hlavným konštituentom morfém, slovných foriem, lexém atď., ale na druhej strane je chápaná najmä ako fonetická artikulačná jednotka, ktorá sama osebe nie je nositeľom sémantickej informácie. Konceptuálna dôležitosť slabiky však bola nedávno zdôraznená v psycholingvistike, keďže lexikón slabík hrá veľkú úlohu pri produkcii a vnímaní ľudského jazyka (pozri Levelt a Wheeldon 1994) .
Vo všeobecnosti je dobre známym problémom definícia a presné určenie hraníc medzi slabikami v slove. Každá empirická porovnávacia štúdia viacerých jazykov si vyžaduje spoľahlivú definíciu slabiky, a to takú, ktorá sa dá použiť na automatickú analýzu textov. Doteraz boli v kvantitatívnej lingvistike urobené iba pokusy o systematický výskum vlastností slabík, a ani tých nebolo veľa (pozri Bektaev 1973, Zörnig a Altmann 1993, Schiller et al. 1996, Obradović et al. 2010, Kelih a Mačutek 2012).
Hlavným cieľom tohto projektu je začať systematický kvantitatívny porovnávací výskum slabiky a jej vlastností v slovanských jazykoch. V prvom kroku, ktorým je tento projekt, sa sústredíme na tri slovanské jazyky, konkrétne na ruštinu, slovenčinu a srbčinu. Tieto tri jazyky reprezentujú tri bežne zaužívané skupiny slovanských jazykov (východo-, západo- a juhoslovanské). Tento projekt má teoretický a empirický charakter (ide o základný výskum, budeme pracovať s dátami). Jeho predpokladané výsledky prispejú k plnšiemu zaradeniu slabiky do jazykového modelu, ktorý navrhli Köhler (2005) a Kelih (2012).
Vyššie spomenutý problém chýbajúcich (technických, t.j., algoritmických, vo všeobecnosti aplikovateľných) procedúr na určenie presných hraníc medzi slabikami (teda rozdelenie slova na slabiky) bude vyriešený na základe postupu, ktorý navrhol Pulgram (1970). Jeho návrh neskôr podstatne vylepšili Lehfeldt (1971) a Kempgen (2003). V rámci nášho projektu vytvoríme semi-automatický algoritmus na sylabifikáciu ruských, slovenských a slovinských textov pričom budeme pracovať s textami v ortografickej podobe. Algoritmus je založený na dôležitej štrukturálnej vlastnosti slabík, konkrétne na ich grafotaktickom/fonotaktickom "správaní sa". V skratke, algoritmus využíva
– rôzne funkcie a správanie sa samohlások a spoluhlások,
– podobnosť štruktúr na začiatkoch slov a slabík,
– podobnosť štruktúr na koncoch slov a slabík,
– uprednostňovanie otvorených slabík (CV),
– princíp minimálnej kódy a maximálneho onsetu, a
– princíp nepravidelnej kódy.
"Štartovacím bodom" pri sylabifikácii slovných foriem je otvorená slabika (CV) . Spoluhláskové zhluky sú rozdelené na základe empiricky získaných štruktúr na začiatkoch a koncoch slov. Tento prístup si vyžaduje systematickú analýzu spoluhláskových zhlukov nachádzajúcich sa na začiatkoch slov, na ich koncoch a v ich strede. Na túto analýzu bude použité veľké množstvo textov. Žiadatelia majú k dispozícii korpusy ruských, slovenských a srbských textov (na stránke http://www-gewi.uni-graz.at/quanta/site.php?show=12 sa dá nájsť obsiahly opis takzvaného Quanta-Project-u z Univerzity v Grazi, teda z pracoviska, kde pred časom pôsobil zodpovedný riešiteľ zo slovenskej strany). Slabika je určená na základe princípov maximálneho onsetu a minimálnej kódy, teda onset bude oveľa dlhší ako kóda. Tento algoritmus sa dá aplikovať tak na písané, ako aj na fonologicky kódované texty.
Proces sylabifikácie v sebe zahŕňa nasledujúce kroky (pozri Pulgram 1970: 48ff):
1. urč slovné formy na základe ortografických kritérií (vylúč skratky a slová z iných jazykov),
2. urč počty slovných foriem,

SK-SRB-2016-0017

3. dočasne je každá samohláska pokladaná za hranicu slabiky (tento krok predpokladá dôslednú analýzu každej samohlásky, najmä čo sa týka funkcie a formy samohláskových zhlukov, akými sú napr. dvojhlásky),
4. hranica slabiky je zafixovaná, ak sa samohláska vyskytuje na konci slova; potom sa krok po kroku analyzujú spoluhláskové zhluky predchádzajúce samohláskam, až kým sa nenájde taký onset slabiky, ktorý sa vyskytuje aj na začiatku niektorého slova,
5. kóda slabiky sa fixuje analogicky ku kroku 3,
6. vyskytujúce sa nepravidelné spoluhláskové zhluky (ktoré sa nenachádzajú na začiatkoch slov) sú reanalyzované, rozhoduje sa v prospech kódy.

Jedna z hlavných modifikácií tohto prístupu zahŕňa do rozhodovania aj početnosti spoluhláskových zhlukov na začiatkoch a na koncoch slov. Lehfeldt (1971) navrhol rozlišovať marginálne a nemarginálne spoluhláskové zhluky. Toto rozlišovanie využíva štatistické testy, ktoré porovnávajú teoreticky očakávanú početnosť s tou pozorovanou. Výhodou tohto prístupu je skutočnosť, že sporné prípady sylabifikácie sa dajú rozhodnúť na základe informácie o početnostiach. Takáto sylabifikácia založená na pravdepodobnostnom prístupe už bola použitá na textovom materiáli v ukrajinčine (Lehfeldt 1971), zatiaľ čo Kempgen (2003) poskytuje informácie o spoluhláskových zhlukoch na začiatkoch a na koncoch slov v ruštine. Unuk (2003) sa zaoberá slovinskými dátami. Ďalšie referencie o slabikách v slovanských jazykoch vrátane slovenčiny uvádza Bethin (1998).

Na základe tohto algoritmu bude v rámci projektu vyvinutý počítačový program, ktorý určí v texte hranice medzi slabikami v slovách, navyše vytvorí zoznam slabík, ktoré sa v texte vyskytujú, spolu s ich početnosťami, dĺžkami a ďalšími vlastnosťami.

Jazykový materiál, ktorý bude analyzovaný, pozostáva z ruských románov “Kak zakaljalas' stal'” od N. Ostrovského a “Master i Margerita” od M. Bulgakova a ich prekladov do slovenčiny a srbčiny (pozri Kelih 2009). Zodpovední žiadatelia tak zo slovenskej (J. Mačutek), ako aj zo srbskej strany (I. Obradović) úzko spolupracujú s Emmerichom Kelihom z Viedenskej univerzity, ktorý je autorom týchto dvoch paralelných korpusov, texty teda majú žiadatelia k dispozícii. Navyše majú aj skúsenosti s prácou s týmito korpusmi.

Analýza spoluhláskových zhlukov na začiatkoch a na koncoch slov sa bude vykonávať na vyvážených korpusoch ruských, slovenských a srbských textov z už spomenutého projektu Quanta. Analýza paralelných korpusov sa pokladá za spoľahlivý spôsob na dosiahnutie maximálnej možnej homogenity dát, keď sa pracuje s rôznymi jazykmi. V tomto štádiu výskumu sa budeme zaoberať najmä dvoma základnými vlastnosťami slabík, a to ich početnosťami a ich dĺžkou. Dá sa očakávať, že početnosti slabík v troch skúmaných jazykoch sa budú dať modelovať jedným spoločným modelom, pričom parametre modelu budú závisieť od jazyka. Budeme skúmať aj otázku, akú úlohu zohrávajú pozoruhodné rozdiely v inventároch foném v ruštine, slovenčine a srbčine.

To isté – jeden model s rôznymi hodnotami parametrov – sa dá očakávať aj pre dĺžku slabík (meranú počtom foném). Keď budú matematické modely – ktoré by mali byť špeciálnymi prípadmi všeobecného modelu používaného v lingvistike, pozri Wimmer a Altmann (2005) – empiricky potvrdené, zabudujeme ich do siete známych lingvistických modelov. Konkrétne budeme hľadať najmä vzťahy medzi a) modelom pre početnosti slabík a pre početnosti grafém/foném (negatívne hypergeometrické rozdelenie, pozri Grzybek et al. 2009), a b) modelom pre dĺžku slabík a pre dĺžku slov (Poissonovo rozdelenie a jeho modifikácie, pozri Grzybek 2007).

Tento projekt poskytne aj dáta pre ďalší výskum. Strauss et al. (2008) uvádzajú niekoľko hypotéz o slabikách a ich vlastnostiach; zatiaľ však neboli dosiahnuté žiadne štatisticky podložené výsledky (snáď s výnimkou kanonickej štruktúry slabík, pozri Zörnig a Altmann 1993, Obradović et al. 2010, Kelih a Mačutek 2012), pričom hlavným dôvodom sú práve chýbajúce dáta.

Zoznam literatúry:

Bektaev, K.B. (1973). Alfavitno-častotnyj slovar' slogov kazakhskogo jazyka. In: Statistika kazakhskogo teksta: 566-611. Alma-Ata: Nauka.

Best, K.-H. (2005a). Morphlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative Linguistics. An International Handbook: 255-260. Berlin, New York: de Gruyter.

Best, K.-H. (2005b). Satzlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative Linguistics. An International Handbook: 298-304. Berlin, New York: de Gruyter.

Bethin, C.Y. (1998): Slavic Prosody: Language change and phonological theory. Cambridge: Cambridge University Press.

Grzybek, P. (ed.) (2007). Contributions to the Science of Text and Language. Word Length Studies and Related Issues. Dordrecht: Kluwer.

Grzybek, P., Kelih, E., Stadlober, E. (2009). Slavic letter frequencies: A common discrete model and regular parameter behavior? In: Köhler, R. (ed.), Issues in Quantitative Linguistics: 17-33. Lüdenscheid: RAM-Verlag.

Grzybek, P., Rusko, M. (2009). Letter, grapheme and (allo-)phone frequencies: the case of Slovak. Glottotheory 2/1, 30-48.

Kelih, E. (2009). Slavisches Parallel-Textkorpus: Projektvorstellung von “Kak zakaljalas' stal' (KZS)”. In: Kelih, E., Levickej, V, Altmann, G. (eds.), Methods of Text Analysis: 106-124. Chernivci: ChNU.
Kelih, E. (2012). Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation. München: Otto Sagner.
Kelih, E., Mačutek, J. (2012). Number of canonical syllable types: a continuous bivariate model. Journal of Quantitative Linguistics 20, 241-251.
Kelih, E. (2013). Grapheme inventory size and repeat rate in Slavic languages. Glottotheory 4(1), 56–71.
Kempgen, S. (2003). Phonologische Silbentrennung im russuschen. In: Kempgen, S. et al. (eds.), Rusistika – Slavistika –
Linguistika: 195-211. München: Otto Sagner.
Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative Linguistics. An International Handbook: 760-774. Berlin, New York: de Gruyter.
Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) (2005). Quantitative Linguistics. An International Handbook. Berlin, New York: de Gruyter.
Lehfeldt, W. (1971). Ein Algorithmus zur automatischen Silbertrennung. Phonetica 24, 212-234.
Levelt, W.J.M., Wheeldon, L. (1994). Do speakers have a mental syllabary? Cognition 50, 239–269.
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.
Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Mehler, A., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). Word Frequency Studies. Berlin, New York: de Gruyter.
Pulgram, E. (1970). Syllable, word, nexus, cursus. Den Haag: Mouton. Schiller, N.O., Meyer, A.S, Baayen, R.H., Levelt, W.J.M. (1996). A comparison of lexeme and speech syllables in Dutch. Journal of Quantitative Linguistics 3,
8-28.
Strauss, U., Fan, F., Altmann, G. (2008). Problems in Quantitative Linguistics 1. Lüdenscheid: RAM-Verlag.
Unuk, D. (2003). Zlog v slovenskem jeziku. Ljubljana: Slavistično društvo Slovenije.
Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative Linguistics. An International Handbook: 791-807. Berlin, New York: de Gruyter.
Zörnig, P., Altmann, G. (1993). A model for distribution of syllable types. Glottometrika 14, 190-196.

02	Harmonogram riešenia výskumného zámeru s ohľadom na charakter výzvy (maximálne 6 000 znakov)

Projekt je plánovaný na dva roky od januára 2017 do decembra 2018, s nasledujúcim harmonogramom:
január 2017 - stretnutie v Belehrade (3 dni)
- príprava metodologicky orientovaného článku o projekte
- koordinácia práce na výbere srbských textov z dostupných databáz
február 2017 – marec 2017
- príprava vybratých srbských textov (odstránenie cudzích slov, skratiek, čísel, atď.)
- vytvorenie počítačového programu pre srbské texty
apríl 2017 - stretnutie v Bratislave (3 dni)
- kontrola prvých výsledkov analýzy srbských textov
- doladenie programu pre srbské texty
máj 2017 – jún 2017
- počítačové spracovanie srbských textov
- matematické modelovanie početností slabík v srbčine, testovanie modelu
júl 2017 - stretnutie v Belehrade (3 dni)
- koordinácia práce na výbere ruských a slovenských textov z dostupných databáz
- príprava abstraktu na konferenciu QUALICO 2018 (miesto konania zatiaľ nie je určené, konať sa bude pravdepodobne na jeseň 2018)
august 2017 – november 2017
- príprava vybratých ruských a slovenských textov (odstránenie cudzích slov, skratiek, čísel, atď.)
- vytvorenie počítačového programu pre ruské a slovenské texty
december 2017 - stretnutie v Bratislave (3 dni)
- kontrola prvých výsledkov analýzy ruských a slovenských textov
- doladenie programu pre ruské a slovenské texty
január 2018 – marec 2018
- počítačové spracovanie ruských a slovenských textov
- matematické modelovanie početností slabík v ruštine a slovenčine, testovanie modelu

apríl 2018 - stretnutie v Bratislave (3 dni)
- príprava článku, v ktorom bude prezentovaný model pre početnosti slabík v ruštine, slovenčine a srbčine
máj 2018 – jún 2018
- matematické modelovanie dĺžky slabík v ruštine, slovenčine a srbčine
júl 2018 - stretnutie v Belehrade (3 dni)
- príprava spoločnej prezentácie na konferenciu QUALICO 2018
august 2018 – september 2018
- interpretácia parametrov modelu pre početnosti slabík, najmä vo vzťahu k inventáru foném/grafém
- odvodenie vzťahu medzi modelmi pre početnosti grafém a slabík v troch skúmaných jazykoch
október 2018 - stretnutie v Belehrade (3 dni)
- príprava článku, v ktorom bude prezentovaný model pre dĺžku slabík v ruštine, slovenčine a srbčine
november 2018
- analýza kanonickej štruktúry slabík v ruštine, slovenčine a srbčine
- interpretácia parametrov modelu pre dĺžku slabík
december 2018 - stretnutie v Bratislave (3 dni)
- príprava článku o kanonickej štruktúre slabík
- úprava vytvoreného počítačového programu do podoby user-friendly softvéru a jeho sprístupnenie na internetovej stránke

03	Zdôvodnenie významu a nevyhnutnosti medzinárodnej spolupráce pri riešení výskumného zámeru (maximálne 6 000 znakov)

Tento projekt je v pravom zmysle slova interdisciplinárny a vyžaduje si tak prácu lingvistu, ako aj matematika. Veľmi zhruba a zjednodušene, výskum v kvantitatívnej lingvistike postupuje takto:
1. lingvista sformuluje lingvisticky relevantnú hypotézu;
2. hypotéza je “preložená do matematického/štatistického jazyka”, t.j., je preformulovaná tak, aby sa dala opísať aparátom matematiky/štatistiky;
3. je vyvinutý matematický model, ktorý zodpovedá hypotéze;
4. sú vytvorené/nájdené/pozorované lingvistické dáta;
5. matematický model je testovaný na dátach;
6. výsledok testu, t.j., (ne-)zamietnutie hypotézy, je lingvisticky interpretované.
Je zrejmé, že lingvista pracujúci sám nebude schopný vyvinúť matematické modely a testovať skúmané hypotézy (keďže modely a testy majú isté predpoklady, ktoré lingvistické dáta môžu, ale nemusia spĺňať). Na druhej strane matematik pracujúci sám nebude schopný posúdiť lingvistickú korektnosť modelovanej a testovanej hypotézy a jeho práca by niesla riziko, že pôjde o aplikáciu matematiky pre samotnú matematiku o hranie sa so vzorcami a lingvistickými dátami bez toho, aby boli dosiahnuté výsledky zmysluplné pre lingvistiku.

Kvantitatívna lingvistika je relatívne mladá, aj keď rýchlo sa rozvíjajúca vedná disciplína, ktorej sa ešte stále venuje len pomerne málo vedcov. Keďže lingvisti a matematici dostávajú veľmi odlišné vzdelanie, väčšinou nie je ľahké zostaviť taký výskumný tím, v ktorom by spolupráca išla od začiatku hladko. Tento výskumný tím však takéto problémy nemá.
J. Mačutek dostal prestížne Štipendium Lise Meitnerovej od rakúskej grantovej agentúry FWF, v rámci ktorého pracoval dva roky na Katedre slovanských štúdií na Univerzite v Grazi (január 2009 – december 2010, potom dostal v Grazi ďalší 6-mesačný projekt), má teda priame skúsenosti s prácou v tíme lingvistov. I. Obradović sa tiež dlhodobo venuje aplikáciám matematiky v lingvistike. Obaja zodpovední riešitelia majú navyše množstvo kontaktov medzi lingvistami. Navyše obaja patria k malej skupine autorov odborných publikácií zameraných na matematické modelovanie vlastností slabík. Tento projekt predstavuje možnosť spojiť ich doterajšie skúsenosti a vybudovať systematický, jednotný prístup k tejto (doteraz pomerne rozdrobenej) problematike.

04	Očakávané výstupy a prínosy riešenia výskumného zámeru pre Slovensko s dôrazom na aspekt medzinárodnej spolupráce (maximálne 6 000 znakov)

Výsledkom tohto projektu bude okrem “priamych” vedeckých výstupov vo forme vedeckých článkov a prezentácií aj niekoľko “vedľajších produktov”, ktoré môžu prospieť (nielen) slovenskej strane. Projekt zvýši medzinárodnú prestíž riešiteľov. Keďže pôjde o prvý projekt systematicky venovaný kvantitatívnemu prístupu k slabikám, dá sa očakávať, že si vedecká komunita jeho výsledky všimne ako pilotnú štúdiu a že bude často citovaný.
Kvantitatívna lingvistika – vrátane skupiny vedcov, ktorí sa venujú výskumu v tejto oblasti – je stále kdesi na okraji záujmu hlavného prúdu lingvistiky. Keďže slabika je lingvistická jednotka, ktorej definícia sa dá dobre operacionalizovať pomocou algoritmického prístupu opierajúceho sa o matematické modely a štatistické testy, výsledky projektu zvýšia šancu na presadenie kvantitatívnych metód ako integrálnej súčasti lingvistiky. V dlhodobej perspektíve môže úspešné riešenia takýchto projektov viesť až k inštitucionálnemu

zakotveniu (k vzniku Centra pre kvantitatívnu lingvistiku napr. v rámci Katedry aplikovanej matematiky a štatistiky).
J. Mačutek je vedúcim Seminára z kvantitatívnej lingvistiky, ktorý sa koná od septembra 2014. Zúčastňujú sa na ňom vedci z Katedry aplikovanej matematiky a štatistiky Univerzity Komenského a z dvoch ústavov SAV (Jazykovedný ústav Ľudovíta Štúra a Ústav informatiky). Srbskí riešitelia budú počas svojich pobytov v Bratislave okrem iného aj prezentovať výsledky svojej práce. Slovenskí vedci, vrátane doktorandov, tak budú mať možnosť rozšíriť si svoje obzory v lingvistike a spoznať modernú výskumnú metodológiu.
Vzhľadom na to, že účastníci seminára z Ústavu informatiky SAV pracujú najmä v aplikovanom výskume (analýza a syntéza reči), tento projekt, aj keď je primárne teoretického charakteru, môže prispieť aj k rozvoju alebo vylepšeniu aplikácií.

05	Charakteristika riešiteľského kolektívu z pohľadu zamerania navrhovaného výskumného zámeru na slovenskej strane, ako aj z pohľadu zapojenia doktorandov a/alebo mladých pracovníkov výskumu a vývoja (do 35 rokov) do projektu (maximálne 6 000 znakov)

Navrhnutý výskumný tím má piatich členov. Vzhľadom na to, že ide o projekt interdisciplinárneho charakteru, tím je ideálnou zmesou dvoch riešiteľov, ktorí dlhodobo pôsobia tak v matematickom, ako aj v lingvistickom výskume (J. Mačutek a I. Obradović), dvoch doktorandiek - matematičiek s rôznymi špecializáciami (M. Koščová sa venuje diskrétnym rozdeleniam pravdepodobnosti, ktoré patria k najčastejšie používaným modelom pre lingvistické dáta; M. Radojičić bude v rámci projektu zodpovedná za navrhovanie a realizáciu algoritmov) a jednej doktorandky na Fakulte filológie (B. Lazić).
Tím je vyvážený aj čo sa týka skúseností jeho členov s výskumom. Jeho členmi sú dvaja skúsení vedci a tri doktorandky. J. Mačutek a I. Obradović sa aktívne podieľajú na výskume v kvantitatívnej lingvistike už viac ako desať rokov. Ich aktivitu v tejto blasti odráža aj množstvo publikácií (pozri http://www.iam.fmph.uniba.sk/ospm/Macutek/jm_publications.pdf a www.rgf.bg.ac.rs/LicnePrezentacije/ivan_obradovic/radovi_kategorije.htm). Obaja sa angažujú v Medzinárodnej kvantitatívnolingvistickej asociácii (IQLA). J. Mačutek je pokladníkom IQLA od roku 2009, I. Obradović bol hlavným organizátorom konferencie QUALICO organizovanej IQLA v roku 2012. J. Mačutek je členom redakčných rád časopisov Journal of Quantitative Linguistics, Journal of Language Modelling, Glottotheory, Mathematical Linguistics a Glottometrics).
M. Koščová je doktorandkou v odbore aplikovaná matematika, pričom jej školiteľom je J. Mačutek. Už jej diplomová práca sa zaoberala štatistickými analýzami lingvistických dát (výsledky z tejto práce sú schválené na publikovanie v Journal of Quantitative Linguistics, pracovná verzia tohto článku, ktorého spoluautormi sú J. Mačutek a E. Kelih, sa dá nájsť na http://arxiv.org/pdf/1504.03608.pdf). Jej dizertačná práca sa sústredí na diskrétne rozdelenia pravdepodobnosti, ktoré sú jedným z najčastejšie používaných matematických modelov. Zapojením sa do tohto projektu získa dvojakú výhodu. Po prvé, práca s dvoma skúsenými vedcami bude pre ňu jedinečnou šancou na rozšírenie si obzorov v matematickom modelovaní v lingvistike; po druhé, matematické/štatistické analýzy (nielen lingvistických) dát často otvárajú nové pohľady na samotné matematické modely a metódy a dávajú tak impulzy na ďalší matematický výskum na teoretickej úrovni. Okrem toho skúsenosti získané pri medzinárodnej spolupráci, jej zaradenie sa do medzinárodnej vedeckej komunity a publikácie v rámci tohto projektu značne zvýšia jej šancu začať po skončení doktorandského štúdia akademickú kariéru.

SK-SRB-2016-0017

B	Project proposal
01	Research aim targets, originality and topicality of the research objective (max. 15 000 characters)

The present project locates itself in the field of quantitative linguistics: it is interdisciplinary by nature, joining the disciplines of mathematics, mainly statistics, on the one hand, and linguistics and text scholarship, on the other. As a scientific discipline, quantitative linguistics is concerned with the construction of a theory of language and language processes, using the apparatus of mathematical modelling and statistical testing. A comprehensive overview of results achieved in not so distant past can be found in the handbook edited by Köhler et al. (2005).

One of the most promising steps towards a linguistic theory in the abovementioned sense is a synergetic approach (see Köhler 2005). It is based on the assumption that particular language levels (e.g., phonemes/graphemes, syllables/morphemes, words, clauses, sentences, etc.) and their properties are not independent, but they rather influence each other. For most of these levels, either quite extensive investigations were performed (see, e.g., Grzybek et al. 2009 and Grzybek and Rusko 2009, and references therein for graphemes, and Popescu et al. 2009 for words), or at least some partial results can be found (papers by Best 2005a and Best 2005b provide overview of results achieved for morphemes and for sentences, respectively).

However, the level of syllables is barely touched, as far as the quantitative approach to linguistics is concerned. Obviously the main reason is that the status of the syllable is disputed in linguistics. On the one hand it is the main constituent of morphemes, word forms, lexemes etc., but on the other hand the syllable is mainly understood as a phonetic unit of articulation, not bearing autonomously semantic information. However, the conceptual importance of the syllable was emphasized in psycholinguistics recently, since a separate syllable lexicon plays a crucial role in human language production and perception (see Levelt and Wheeldon 1994).

One notorious problem in general is the definition and the exact determination of the syllable boundaries. For any empirical crosslinguistic study, a reliable definition of the syllable, including a potential applicability for an automatic analysis, is required. So far in quantitative linguistics only few attempts have been made to study quantitative properties of the syllable systematically (see Bektaev 1973, Zörnig and Altmann 1993, Schiller et al. 1996, Obradović et al. 2010, Kelih and Mačutek 2012).

The main aim of this project is to start a crosslinguistic study of the syllable and its quantitative properties in Slavic languages. In the first attempt we will focus on three Slavic languages, namely, on Russian, Slovak, and Serbian. The three chosen languages represent three established groups of Slavic languages (East, West, and South, respectively).

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”. In a nutshell, the algorithm utilizes

– the different functions and behaviour of vowels and consonants,
– the similarity of the word-initial and syllable-initial structure,
– the similarity of the word-final and syllable-final structure,
– a preference towards an open syllable (CV),
– the principle of the minimal coda and the maximal onset, and,
– the principle of the irregular coda.

Thus, an open syllable (CV) is the starting point of the syllabification of word forms. Consonantal clusters are split according to the empirically obtained word-initial and word-final structure. This implies the systematic analysis of word-initial, word-final and word-medial consonant clusters on a huge amount of text data. The applicants have text corpora for Russian, Slovak, and Serbian at their disposal (see http://www-gewi.uni-graz.at/quanta/site.php?show=12 for the comprehensive description of the so-called Quanta-Project from the University of Graz, i.e., from the institution where the principal investigator from the Slovak side worked some time ago). The syllable is determined according to the maximal onset and minimal coda principle, implying in fact a much longer onset than coda. The algorithm is applicable both for written and phonologically encoded texts. The syllabification process includes the following steps (see Pulgram 1970: 48ff):

SK-SRB-2016-0017

1. determine word forms based on orthographical criteria (exclude abbreviations, non-native vocabulary has to be removed manually),
2. determine the number of word form types and tokens (both levels can be analysed, depending on the linguistic hypothesis analysed) ,
3. temporarily, every vowel is determined as a syllable boundary (which implies the in-depth analysis of each vowel, especially the function and the form of occurring vowel clusters, e.g., diphthongs),
4. fix the syllable boundary in case that the vowel occurs in the word-final position, followed by a step-by-step analysis of occurring consonantal clusters in a pre-vocal position until a valid word- initial syllable onset is found,
5. in analogy to Step 3, the syllable coda is fixed,
6. a re-analysis of occurring irregular consonant clusters in favour of the coda.

One of the main modifications of this approach concerns the integration of the frequency of word-final and word-initial consonant clusters. Lehfeldt (1971) suggested to differentiate marginal and non-marginal consonantal clusters, which can be extracted by the means of statistical tests, calculating a theoretically expected value in addition to the empirically observed frequency. The advantage of this approach is that disputed cases of the intersyllabic division can be solved, based on the frequency information. This probability based syllabification algorithm has already been applied to Ukrainian material (Lehfeldt 1971), whereas Kempgen (2003) provides data on consonantal clusters in word-initial and word-final positions for Russian; see also Unuk (2003) for Slovene data. Some further references for the syllable in Slavic languages in general, including Slovak, can be found in Bethin (1998).

Based on this algorithm, 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 form 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.

At this stage of research, our main goal is to investigate two basic properties of syllables, namely their frequencies and their length. We expect that one common mathematical model for syllable rank-frequency distribution will be found, its parameters will probably be language specific. Finally, it has to be analyzed which role the remarkable differences of the phoneme inventory sizes of Russian, Slovak, and Serbian play. The same is expected for the model of syllable length (measured in the number of phonemes). When the mathematical models - which should be special cases of one general model used in linguistics, see Wimmer and Altmann (2005) - are corroborated on the language material, we will embed them into the network of other known models, i.e., we will look mainly for the relations between a) the model for syllable frequencies and grapheme/phoneme frequencies (the negative hypergeometric distribution, see Grzybek et al. 2009), and b) the model for syllable length and word length (the Poisson distribution and its modifications, see Grzybek 2007).

The project will also provide data for further research. Several hypotheses regarding syllables and their properties can be found in Strauss et al. (2008); however, almost no statistically corroborated results (perhaps with the exception of the canonical syllable structure, see Zörnig and Altmann 1993, Obradović et al. 2010, Kelih and Mačutek 2012) have been achieved so far, mainly because of the lack of data.

List of literature:

Bektaev, K.B. (1973). Alfavitno-častotnyj slovar' slogov kazakhskogo jazyka. In: Statistika kazakhskogo teksta: 566-611. Alma-Ata: Nauka.

Best, K.-H. (2005a). Morphlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative Linguistics. An International Handbook: 255-260. Berlin, New York: de Gruyter.

Best, K.-H. (2005b). Satzlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative Linguistics. An International Handbook: 298-304. Berlin, New York: de Gruyter.

Bethin, C.Y. (1998): Slavic Prosody: Language change and phonological theory. Cambridge: Cambridge University Press.

Grzybek, P. (ed.) (2007). Contributions to the Science of Text and Language. Word Length Studies and Related Issues. Dordrecht: Kluwer.

Grzybek, P., Kelih, E., Stadlober, E. (2009). Slavic letter frequencies: A common discrete model and regular parameter behavior? In: Köhler, R. (ed.), Issues in Quantitative Linguistics: 17-33. Lüdenscheid: RAM-Verlag.

Grzybek, P., Rusko, M. (2009). Letter, grapheme and (allo-)phone frequencies: the case of Slovak. Glottotheory 2/1, 30-48.
Kelih, E. (2009). Slavisches Parallel-Textkorpus: Projektvorstellung von “Kak zakaljalas' stal' (KZS)”. In: Kelih, E., Levickej, V, Altmann, G. (eds.), Methods of Text Analysis: 106-124. Chernivci: ChNU.
Kelih, E. (2012). Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation. München: Otto Sagner.
Kelih, E., Mačutek, J. (2012). Number of canonical syllable types: a continuous bivariate model. Journal of Quantitative Linguistics 20, 241-251.
Kelih, E. (2013). Grapheme inventory size and repeat rate in Slavic languages. Glottotheory 4(1), 56–71.
Kempgen, S. (2003). Phonologische Silbentrennung im russuschen. In: Kempgen, S. et al. (eds.), Rusistika – Slavistika –
Linguistika: 195-211. München: Otto Sagner.
Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative Linguistics. An International Handbook: 760-774. Berlin, New York: de Gruyter.
Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) (2005). Quantitative Linguistics. An International Handbook. Berlin, New York: de Gruyter.
Lehfeldt, W. (1971). Ein Algorithmus zur automatischen Silbertrennung. Phonetica 24, 212-234.
Levelt, W.J.M., Wheeldon, L. (1994). Do speakers have a mental syllabary? Cognition 50, 239–269.
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.
Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Mehler, A., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). Word Frequency Studies. Berlin, New York: de Gruyter.
Pulgram, E. (1970). Syllable, word, nexus, cursus. Den Haag: Mouton.
Schiller, N.O., Meyer, A.S, Baayen, R.H., Levelt, W.J.M. (1996). A comparison of lexeme and speech syllables in Dutch. Journal of Quantitative Linguistics 3, 8-28.
Strauss, U., Fan, F., Altmann, G. (2008). Problems in Quantitative Linguistics 1. Lüdenscheid: RAM-Verlag.
Unuk, D. (2003). Zlog v slovenskem jeziku. Ljubljana: Slavistično društvo Slovenije. Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), Quantitative
Linguistics. An International Handbook: 791-807. Berlin, New York: de Gruyter. Zörnig, P., Altmann, G. (1993). A model for distribution of syllable types. Glottometrika 14, 190-196.

02 Project schedule and activities with regard to the call (max. 6 000 characters)

The research project is planned for two years, from January 2017 until December 2018, with a schedule as follows:
January 2017 - meeting in Belgrade (3 days)
- preparation of methodologically oriented paper about the project
- coordination of the work on extracting Serbian texts from available databases
February 2017 – March 2017
- preparation of selected Serbian (removing foreign words, abbreviations, numbers, etc.)
- creation of computer program for Serbian texts
April 2017 - meeting in Bratislava (3 days)
- revision of the first outputs of the analysis of Serbian texts
- last modifications of the program for Serbian texts
May 2017 – June 2017
- computer processing of Serbian texts
- mathematical modelling of syllable frequencies in Serbian, model testing
July 2017 - meeting in Belgrade (3 days)
- coordination of the work on extracting Russian and Slovak texts from available databases
- preparation of an abstract for QUALICO 2018 conference (location not yet determined, it will take place probably in autumn 2018)
August 2017 – November 2017
- preparation of selected Russian and Slovak texts (removing foreign words, abbreviations, numbers, etc.)
- creation of computer program for Russian and Slovak texts
December 2017 - meeting in Bratislava (3 days)
- revision of the first outputs of the analysis of Russian and Slovak texts
- last modifications of the program for Russian and Slovak texts
January 2018 – March 2018

- computer processing of Russian and Slovak texts
- mathematical modelling of syllable frequencies in Russian and Slovak, model testing
April 2018 - meeting in Bratislava (3 days)
- preparation of paper which will present the model for syllable frequencies in Russian, Slovak, and Serbian
May 2018 – June 2018
- mathematical modelling of syllable length in Russian, Slovak, and Serbian
July 2018 - meeting in Belgrade (3 days)
- preparation of common presentation for QUALICO 2018 conference
August 2018 – September 2018
- interpretation of parameters of the model for syllable frequencies, especially with respect to phoneme/grapheme inventory sizes
- work on the relation between models for grapheme frequencies and syllable frequencies in the investigated languages
October 2018 - meeting in Belgrade (3 days)
- preparation of paper which will present the model for syllable length in Russian, Slovak, and Serbian
November 2018
- analysis of canonical syllable structure in Russian, Slovak and Serbian
- interpretation of parameters of the model for syllable length
December 2018 - meeting in Bratislava (3 days)
- preparation of paper which will present the canonical syllable structure
- modification of the created computer program to a user-friendly software and making it available on a webpage

03	Importance and meaningfulness of bilateral cooperation at international project participation (max. 6 000 characters)

The project, being of a genuinely interdisciplinary character, requires inputs both from the side of linguistics and mathematics/statistics. Very roughly, and admittedly with a bit of simplification, researchers in quantitative linguistics work as follows:
1. a linguist formulates a linguistically relevant hypothesis;
2. the hypothesis is “translated into the mathematical/statistical language”, i.e., it is reformulated so that it can be described by the apparatus of mathematics/statistics;
3. a mathematical model corresponding to the hypothesis is developed;
4. linguistic data are found/observed/created;
5. the mathematical model is tested on the data;
6. the result of the test, i.e., the (non-)rejection of the hypothesis, is linguistically interpreted.
It is obvious that a linguist alone will not be able to develop mathematical models and to test the hypothesis under investigation (as the models and tests have some assumptions, which linguistic data can, but do not have to satisfy). On the other hand, a mathematician alone will not be able to assess the linguistic soundness of the hypothesis modelled and tested, and such a work would bear a risk of l'art-pour-l'art applications of mathematics, of playing with mathematical formulas and substituting linguistic data into them, without actually achieving results meaningful for linguistic research.
Quantitative linguistics is a relatively young – albeit quickly developing – discipline, in which still not too many scientists work. Given (very) different educational backgrounds of linguists and mathematicians, it is usually not easy to build a team of researchers being able to cooperate smoothly. This is, however, not the case of this research team.
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. Both principal investigators have, in addition, many contacts among linguistists, and both of them belong to a small group of authors who published papers on mathematical modelling of syllable properties. This project presents a possibility to connect their work and to build a systematic, unified approach to this (so far quite fragmented) research area.

04	The anticipated contribution of international research cooperation for Slovakia (socio-economic benefits) (max. 6 000 characters)

The project will provide, in addition to its "direct" scientific outcomes in the form of scientific papers and presentations, also several "by-products", from which (not only) the Slovak side can drive benefits. The project will increase the international prestige of the participants. As it will be the first project systematically dedicated to the quantitative approach to the level of syllables, it can be expected that its outcomes will be studied in the related scientific community, they will be regarded as a pilot study in that area, and, consequently, they will often be cited.

SK-SRB-2016-0017

Quantitative linguistics – and the group of scientists conducting research in this area – is still somewhere on the edge of the interest of the mainstream linguistics. Given that the syllable is a linguistic unit definition of which can be operationalized by the means of an algorithmic approach, which relies on mathematical models and statistical tests, the outcomes of the project will increase a chance to establish quantitative approaches to linguistic research as an integral part of linguistics. In the long-term perspective, even an institutional development (something like a Centre for Quantitative Linguistics at the Department of Applied Mathematics and Statistics) could become realistic.
J. Mačutek is the leader of the Seminar on quantitative linguistics which has been organized since September 2014. Researchers from the Department of Applied Mathematics and Statistics of the Comenius University and from two institutes of the Slovak Academy of Sciences (Ľudovít Štúr Institute of Linguistics and Institute of Informatics) take part in it. During his stays in Bratislava, E. Kelih will present results of his research at the seminar. Slovak scientists, including PhD students, will thus have an opportunity to enlarge their overview in linguistics and in the modern linguistic methodology.
As seminar participants from the Institute of Informatics of the Slovak Academy of Sciences are engaged mainly in applied research (speech analysis and synthesis), this project, although theoretically oriented, may contribute also to the development or improvement of applications.

05	Characteristics of the research team in terms of professional orientation of the proposed research plan/objective on the Slovak side as well as in terms of involvement of PhD students and/or young R&D staff (up to 35 years) in the project (max. 6 000 ch.)

The proposed research team consists of five members. 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. J. Mačutek and I. Obradović have been active in quantitative linguistics research for more than a decade. They publish extensively in this area (see http://www.iam.fmph.uniba.sk/ospm/Macutek/Publications.pdf and www.rgf.bg.ac.rs/LicnePrezentacije/ivan_obradovic/radovi_kategorije.htm for the respective lists of publications). Both of them are active members of the International Quantitative Linguistics Association (IQLA). J. Mačutek has been the Treasurer of IQLA since 2009, I. Obradović was the main organizer of the IQLA conference (QUALICO) in 2012. J. Mačutek is a member of editorial boards of Journal of Quantitative Linguistics, Journal of Language Modelling, Glottotheory, Mathematical Linguistics, and Glottometrics).
M. Koščová is a PhD student in applied mathematics, with J. Mačutek serving as her supervisor. Already her diploma thesis dealt with statistical analyses of linguistic data (results from the diploma thesis will be published in the Journal of Quantitative Linguistics; a draft of the paper, of which J. Mačutek and E. Kelih are co-authors, can be found at http://arxiv.org/pdf/1504.03608.pdf). Her dissertation thesis is focused on discrete probability distributions, which are one of the most commonly used mathematical models in quantitative linguistics. Her participation in the project will give her a twofold advantage – first, working with two experienced researchers she will have a unique opportunity to gain insight into mathematical modelling in linguistics; second, a mathematical/statistical analysis of (not only linguistic) data often opens up new vistas on mathematical models and methods themselves, and thus provides an impetus towards further mathematical research on a theoretical level. Moreover, experience she will obtain from an international cooperation, her integration into the international scientific community, and publications from the project will significantly increase her chances to start an academic career after her PhD study.

C – Rozpočet		**Rozpočet projektu v EUR Project budget in EUR**		
Žiadateľ: Univerzita Komenského v Bratislave				
Rok / Year		**2017**	**2018**	**Celkovo / Total**
01	Cestovné a pobytové náklady / Travel and stay expenses	2 241,00	2 241,00	4 482,00
Položky celkom / Cost items summary		**2 241,00**	**2 241,00**	**4 482,00**

Zdôvodnenia	**2017**	**Zdôvodnenie pre organizáciu: Univerzita Komenského v Bratislave**
01	Cestovné a pobytové náklady	
4 spiatočné letenky Viedeň-Belehrad (2 krát 2 osoby) 600€ (4 x 150€) Ubytovanie pre srbských riešiteľov (3 osoby, 2 pobyty, 3 noci oba pobyty) 900€ (18 x 50€) Pobytové náklady (3 osoby, 2 pobyty, 3 dni oba pobyty) 741€ (18 x 41.16€)		
01	Travel and stay expenses	
4 return air-tickets Vienna-Belgrade (2 times 2 persons) 600€ (4 x 150€) Accomodation costs for Serbian participants (3 persons, 2 stays, 3 nights each stay) 900€ (18 x 50€) Per diem allowances (3 persons, 2 stays, 3 days each stay) 741€ (18 x 41.16€)		

Zdôvodnenia	**2018**	**Zdôvodnenie pre organizáciu: Univerzita Komenského v Bratislave**
01	Cestovné a pobytové náklady	
4 spiatočné letenky Viedeň-Belehrad (2 krát 2 osoby) 600€ (4 x 150€) Ubytovanie pre srbských riešiteľov (3 osoby, 2 pobyty, 3 noci oba pobyty) 900€ (18 x 50€) Pobytové náklady (3 osoby, 2 pobyty, 3 dni oba pobyty) 741€ (18 x 41.16€)		
01	Travel and stay expenses	
4 return air-ticket Vienna-Belgrade (2 times 2 persons) 600€ (4 x 150€) Accomodation costs for Serbian participants (3 persons, 2 stays, 3 nights each stay) 900€ (18 x 50€) Per diem allowances (3 persons, 2 stays, 3 days each stay) 741€ (18 x 41.16€)		

SK-SRB-2016-0017

C – Rozpočet		Rozpočet projektu v EUR Project budget in EUR		
Sumárny rozpočet projektu / Budget summary				
Rok / Year		**2017**	**2018**	**Celkovo / Total**
01	Cestovné a pobytové náklady / Travel and stay expenses	2 241,00	2 241,00	4 482,00
Položky celkom / Cost items summary		**2 241,00**	**2 241,00**	**4 482,00**

D	**Ciele a výstupy projektu**	
01	Očakávané výstupy riešenia	
Výstupy	Rok 2017	Rok 2018
Príprava spoločných publikácií a iných výstupov	1	3
Aktívna účasť na konferenciách ako výstup zo spoločných aktivít VaV	0	1
Zapojenie doktorandov a/alebo mladých vedeckých pracovníkov (do 35 rokov)	1	1

D	**Project objectives and outcomes**	
01	Anticipated results	
Outputs	Year 2017	Year 2018
Preparation of the joint publication and other outputs	1	3
Active participation at conferences, organization of joint scientific activities	0	1
Participation of PhD students and/or young researchers (up to 35 years)	1	1

E – Podpisy a inštitucionálne schválenie – slovenský žiadateľ		
01	Podpis zodpovedného riešiteľa	
02	Dátum	
03	Meno a priezvisko štatutárneho zástupcu	prof., RNDr. Karol Mičieta, PhD.
04	Podpis štatutárneho zástupcu	
05	Dátum	

Ja, vyššie podpísaný/á prof., RNDr. Karol Mičieta, PhD. , štatutárny zástupca čestne vyhlasujem, že:

- organizácia má platné osvedčenie o spôsobilosti vykonávať výskum a vývoj v zmysle § 18 ods. 2 písm. f) zákona č. 172/2005 Z. z. v znení neskorších predpisov a je evidovaná v zozname osôb spôsobilých vykonávať výskum a vývoj,
- som oprávnený žiadateľ v zmysle výzvy Slovensko – Srbsko 2016,
- všetky informácie obsiahnuté v žiadosti sú pravdivé,
  žiadosť zaslaná v elektronickej podobe je po obsahovej stránke zhodná so žiadosťou v listinnej podobe,
- organizácia nie je daňovým dlžníkom,
- organizácia nie je dlžníkom poistného na sociálnom poistení (dôchodkovom, nemocenskom, garančnom a úrazovom poistení, poistení v nezamestnanosti, poistení do rezervného fondu solidarity) a dlžníkom príspevkov na starobné dôchodkové sporenie,
- organizácia nie je dlžníkom poistného na zdravotnom poistení (za každú zdravotnú poisťovňu),
- organizácia nie je v likvidácii,
- voči organizácii nie je vedený výkon rozhodnutia,
- voči organizácii nebolo začaté konkurzné/reštrukturalizačné konanie,
- na majetok organizácie nie je vyhlásený konkurz,
- voči organizácii nebol namietnutý návrh na vyhlásenie konkurzu pre nedostatok majetku,
- organizácia neporušila zákaz nelegálnej práce a nelegálneho zamestnávania podľa osobitného predpisu za obdobie od jeho účinnosti (1. apríl 2005) a v prípade porušenia nelegálneho zamestnávania cudzinca podľa § 2 ods. 2 písm. c) zákona č. 82/2005 Z. z. o nelegálnej práci a nelegálnom zamestnávaní a o zmene a doplnení niektorých zákonov v znení neskorších predpisov uplynulo päť rokov od porušenia tohto zákazu,
- organizácia má vysporiadané finančné vzťahy so štátnym rozpočtom,
- som si vedomý povinnosti poskytnutia informácií o výskume a vývoji, na ktorý boli poskytnuté finančné prostriedky zo štátneho rozpočtu na účel zverejnenia v súlade s platnou legislatívou SR a s týmto súhlasím,
- žiadateľ o financovanie projektu nežiadal finančné prostriedky z iných zdrojov štátneho rozpočtu,
- som si vedomý povinnosti zverejnenia informácií o technickej infraštruktúre výskumu a vývoja, ktorú žiadateľ buduje z poskytnutých prostriedkov štátneho rozpočtu alebo z prostriedkov Európskej únie na centrálnom informačnom portáli a s týmto súhlasím.

V prípade vyzvania zo strany APVV sa zaväzujem nahradiť toto čestné vyhlásenie aktuálnymi potvrdeniami príslušných úradov.

V prípade zistenia, že údaje uvedené v predmetnom vyhlásení nie sú pravdivé, beriem na vedomie, že žiadosť bude vyradená alebo zmluva o poskytnutí prostriedkov nebude podpísaná, prípadne dôjde k odstúpeniu od zmluvy a zaväzujem sa nahradiť všetky škody spôsobené APVV z dôvodu poskytnutia nepravdivých a nesprávnych údajov a vyhlásení.