Maja P. Miličević[*]
University of Belgrade
Faculty of Philology

SEMI-AUTOMATIC CONSTRUCTION OF COMPARABLE GENRE-ORIENTED CORPORA OF SERBIAN IN CYRILLIC AND LATIN SCRIPTS[†]

This article deals with methods for the semi-automatic construction of genre-oriented corpora from the web, drawing on the BootCaT toolkit. In particular, it reports the results of two parallel studies on Serbian. The first factor that makes Serbian interesting in this respect is its rich inflectional morphology; the second concerns the use of two alphabets, Cyrillic and Latin. Four different methods for the creation of genre-oriented corpora are compared for each script, based on keywords and sequences of words (n-grams) of different lengths (unigrams, bigrams and trigrams). The genre under scrutiny is that of cooking recipes, a genre that is very formulaic and also highly represented on the web. The analysis of the corpora created using the different methods shows that for the Latin script no single method substantially outperforms the others, with excellent results obtained across the board, while for Cyrillic there is a clear advantage for bigrams and trigrams over keywords and unigrams. As well as further confirming the potential of genre-oriented methods of corpus construction for languages with a rich system of inflectional morphology, the results also point to a functional split between the two scripts of Serbian.

Keywords: genre, semi-automatic corpora, BootCaT, keywords, n-grams, inflectional morphology, Serbian, Cyrillic script, Latin script

[*] Contact address: Maja Miličević, Faculty of Philology, Department of General Linguistics, Studentski trg 3, 11000 Belgrade, Serbia; m.milicevic@fil.bg.ac.rs.

[†] This paper was written as part of the project No. 178004, *The Standard Serbian Language: Syntactic, Semantic and Pragmatic Explorations* (*Standardni srpski jezik: sintaksička, semantička i pragmatička istraživanja*), funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia. It is based on presentations given at the workshop *BootCaTters of the world unite!* (University of Bologna at Forlì, 24 June 2013), and the conference *Languages and Cultures in Time and Space 3* (University of Novi Sad, 16 November 2013). Both audiences provided useful comments that helped improve the paper.

1. Introduction

In the past decades linguistic research has been heavily influenced by the development of corpus linguistics. An important milestone within this development was the advent of the World Wide Web, which made large amounts of text accessible to a wider group of potential corpora creators. Even though the web is often associated with lack of detailed information about the texts and their origin, it is at the same time a uniquely rich source of authentic and easily retrievable linguistic data. Moreover, corpora based on web data can nowadays be created not only by the technically savvy, but also by users who possess less technical knowledge, which is often the case with (non-computational) linguists and translators.

A widely used program for the transformation of web data into proper textual corpora is BootCaT (*Bootstrapping Corpora and Terms*; Baroni and Bernardini, 2004). BootCaT was originally devised for semi-automatic creation of small to medium-sized specialised corpora dedicated to a single topic (rock music, cat food, psychiatric diseases, etc.). The program queries the web based on criteria defined by the user, typically in the form of a list of keywords – terms characteristic of the chosen topic. After collecting the relevant data, in a final step it merges the textual material collected from the web into a single plain text file that can subsequently be used in a number of different ways. Attempts to outperform the BootCaT approach have so far not produced satisfactory results in the form of efficient and freely available instruments for creating user-defined corpora (see in particular Barbaresi, 2014).

On the other hand, several recent studies looked at the possibility of extending the “traditional” BootCaT procedure outlined above to include the creation of genre-oriented corpora (in addition to the topic-oriented ones). Since genre is mostly defined by text-external factors, automatic creation of genre-based corpora is normally seen as a highly complex task. However, attempts at identifying text-internal correlates of genre have found some fairly reliable indicators, one of them being frequent sequences of words, i.e. word-level n-grams. Studies employing BootCaT suggest that an approach based on n-grams could be on the right track for this software too, but with a caveat that it leads to different success rates for different languages; satisfactory results have been obtained for English (Dalan, 2012; Bernardini and Ferraresi, 2013) and Serbian (Miličević, Bernardini, and Ferraresi, 2014), but less so for Italian (Bernardini and Ferraresi, 2013; Miličević et al., 2014). The diverse success rates for different languages have mostly been attributed to differences in inflectional morphology.

In the present paper, we take this line of research as our starting point; we carry out a more in-depth investigation of Serbian by focusing on a very specific issue that affects the creation of genre-oriented corpora in this language, namely the use of two scripts – Cyrillic and Latin. We implement an approach that enables obtaining comparable corpora in the two scripts, using BootCaT; at the same time, for each script we evaluate four different corpus construction methods, based respectively on keywords and sequences of one, two and three words (unigrams, bigrams and trigrams). We partly rely on data from Miličević et al. (2014) and focus on the genre of culinary recipes, very widespread on the web and characterised by the use of highly formulaic language. The results show that fairly high success rates are obtained with all four methods for the Latin script, but only with bigrams and trigrams for the Cyrillic script; another interesting finding is that the Cyrillic corpora tend to be much larger. The detected differences are attributed to a partial functional split between the two scripts, with Latin being more present within the actual content of web pages dedicated to culinary recipes, and Cyrillic in books posted on websites such as scribd.org.

2. Textual genre and its automatic identification

As a linguistic category, genre is primarily defined based on functions performed by texts in the community to which they belong (cf. Sinclair, 2005). In the words of Swales, "[a] genre comprises a class of communicative events, the members of which share some set of communicative purposes" (1990, p. 58). Since extra-linguistic criteria such as communicative purpose of a text are difficult to operationalise in computational terms, automatic genre classification must rely on different kinds of text properties. As is often noted in the literature (see e.g. Swales, 1990; Crowston and Kwasnik, 2004), texts belonging to the same genre tend to have certain formal characteristics in common (in terms of style, structure and content), which allow for genre classification based on linguistic criteria. Among text-internal features considered useful for the identification of genre one can distinguish those strictly computational and not very intuitive, such as sequences of characters (character n-grams; cf. Kanaris and Stamatatos, 2009), those that are intuitive but difficult to obtain, such as sequences of parts of speech (POS n-grams; cf. Sharoff, 2007), and finally those that are both intuitive and easy to obtain, such as sequences of words.

Word sequences have proven particularly useful as genre indicators in English: uninterrupted sequences of four words were used by Biber and Conrad (1999) in a study of differences between conversation and academic prose; Gries and Mukherjee (2010) studied regional variations in Asian English by comparing sequences of words of varying length in the International Corpus of English (ICE); Gries, Newman, and Shaoul (2011) found that registers and sub-registers present in the ICE-GB and BNC-Baby corpora can be distinguished based on n-grams in general, and trigrams in particular; bigrams were identified as the best solution in studies conducted by Crossley and Louwerse (2007) and Louwerse and Crossley (2006). In other languages, (non-lemmatised) unigrams also appear to be reliable indicators of genre; Baroni et al. (2004), for instance, found this was the case for texts in the Italian newspaper corpus la Repubblica. When it comes to Serbian, Vitas, Pavlović-Lažetić, and Krstev (2006) observed very little overlap, i.e. very few frequent bigrams and trigrams shared between two corpora of narrative and two corpora of newspaper texts, which could be attributed to the relevance of n-grams for the classification of genre, or to data sparseness.

As for the studies that explored the possibility of creating genre-oriented corpora with BootCat, the use of trigrams, or alternatively a mixture of trigrams and keywords, was shown to be more efficient than keywords alone for English, but a similar result was not replicated for Italian. Dalan (2012) studied the genre of academic course descriptions in English and showed that combining trigrams and keywords led to better results than any of the individual methods; the method based on trigrams, however, was only slightly less efficient. Bernardini and Ferraresi (2013) compared the traditional keywords approach with a method based on trigrams frequent in the genre they studied (patient information leaflets), in English and Italian. The evaluation of the obtained corpora revealed that trigrams gave results superior to keywords only for English, suggesting for Italian a possible influence of its rich inflectional morphology (primarily conjugated verbs). Finally, Miličević et al. (2014) compared the usefulness of keywords, unigrams, bigrams and trigrams for the creation of genre-oriented corpora in Italian and Serbian, another highly inflected language. They found that the best result for Italian was achieved by the keyword method, while for Serbian no single method substantially outperformed the others; in addition, the results obtained for Serbian were consistently better than those for Italian, which was explained by the richer nominal morphology of Serbian compared to Italian, as well as the absence vs. presence of articles.[1]

While it can fairly confidently be concluded that n-grams can be used as a BootCaT method to retrieve texts belonging to a determined genre, it is clear that the success rate depends on the properties of specific languages. Serbian has been identified as a language in which several methods give good results, but previous work focused on one of its scripts (Latin), so new studies are needed to complete the picture with data on Cyrillic. The only previous study that did look at both scripts, Miličević (2013) did so only for keywords, finding that they were more reliable indicators of genre for the Latin than for the Cyrillic script for culinary recipes, and even more so for real estate advertisements.

3. The two scripts of Serbian

Serbian is one of the not so numerous languages written in two scripts. Until 2006, both the Cyrillic and the Latin script were listed as official in the highest-level legal documents regulating language use in Serbia. Since 2006, only Cyrillic is designated as the script of official correspondence by the Constitution of the Republic of Serbia. Recent editions of the Orthography of the Serbian Language also give more emphasis to the Cyrillic script, albeit without leaving out the Latin one (see Pešikan, Jerković, and Pižurica, 2010). However, outside the official correspondence and the education system, there is a general impression that more written communication in Serbian relies on the Latin script.

The medium that appears to particularly favour the Latin script is electronic communication, especially the Internet, in all likelihood due to coding issues and facilitated keyboard switching between Serbian and

[1] The prevalence of articles in Italian led to the most frequent bigrams and trigrams extracted from a manual corpus of culinary recipes not being sufficiently specialised.

other languages, or with the goal of enhancing mutual legibility with other speakers of former Serbo-Croatian. The Latin script is also dominant in the corpus linguistics / natural language processing arena, being widely represented in corpora and other resources; Cyrillic originals, if included, tend to undergo transliteration. This is the case with the Corpus of Contemporary Serbian Language (see Vitas et al., 2003; Utvić, 2013) and various related lexical resources (see e.g. Pavlović-Lažetić, Vitas, and Krstev, 2004), where the diacriticised letters such as *č* or *ć* are transformed into ASCII sequences not instantiated in Serbian (e.g. *č* is represented as *cy*), as well as with the Unicode-based web corpus srWaC (Ljubešić and Klubička, 2014). While being very practical and suitable for most linguistic purposes, such a situation impedes the study of functional differences between the two scripts, which is still often dealt with impressionistically.

One way to fill this lacuna is to use non-transliterated web corpora created with BootCaT. The software's semi-automatic procedure of corpus creation is particularly suitable for obtaining comparable data on the use of Latin and Cyrillic scripts in texts belonging to specific genres; findings from this first step can be used to help users make more informed decisions on which script to use for which type of corpora.

4. Creating semi-automatic Cyrillic and Latin genre-oriented corpora of Serbian

While previous research has already shown that BootCat can be used to create high-quality genre-oriented corpora for Serbian, it was limited to the Latin script, and data on Cyrillic is still missing. In this section we present a study designed to address this issue.[2]

4.1 Method

The standard BootCat approach is based on a first step in which the user selects a certain number of *seed* terms (individual keywords or phrases characteristic of the topic of interest). The seeds are automatically combined into groups (*tuples*) and sent to the search engine (at the time Bing); a tuple constitutes a query. As a first result the user obtains a list of URLs

[2] The part of the study concerning the Latin script has already been published in Milićević et al. (2014). For ease of exposition, we repeat all the relevant information here.

of pages that contain the requested tuples; (s)he then has the possibility to use additional options such as setting tuple length or including web domain limitations (e.g. looking only for pages within the .rs domain), and lastly the possibility to exclude URLs judged irrelevant. In the final step, the content of the selected pages is downloaded, cleaned from the HTML code, transformed into plain text and saved as a single .txt file.

The method based on n-grams, compared in this study to the traditional approach, involves an insertion, in the place of keywords, of sequences of words (n-grams) frequent in the chosen genre. Given the results of previous studies, which have identified as relevant not only sequences of several words, but also non-lemmatised unigrams (the most frequent words in the genre in question, in inflected form), in this study we created four different corpora for each script, based respectively on (1) keywords, (2) unigrams, (3) bigrams, and (4) trigrams. As for the textual genre, choosing culinary recipes is due to the fact that they constitute a familiar genre that is well represented on the web, while at the same being sufficiently specialised and conventional (see e.g. Arendholz et al., 2013; cf. also Bernardini and Ferraresi, 2013). Moreover, there is a growing attention towards this genre in the field of computational linguistics; culinary recipes in Serbian, for example, have recently been studied from the point of view of enlargement of the WordNet and morphological dictionaries for this language (Vujičić Stanković, Krstve, and Vitas, 2014).

Since both keywords and n-grams need to be frequent in the genre of interest, the starting point for the construction of semi-automatic corpora was a (semi-)manually constructed corpus of the studied genre. Specifically, eleven websites dedicated to culinary recipes were identified via simple Google queries; all websites were written in the Latin script. Pages from these websites were downloaded with BootCat; the procedure consisted in replacing tuples with website addresses preceded by the site: operator (cf. Bernardini et al. 2010). The subsequent manual clean-up of files, by means of which all text fragments that did not belong to the desired genres were removed (including the lines containing the URL, automatically added by BootCat), resulted in a corpus of 200,608 words.[3],[4]

[3] All corpora in this study were constructed using the command line procedure of *BootCaT*; the version used was the one described in Ljubešić (2013).

[4] A similar procedure for the creation of a web corpus of culinary recipes was used by Vujičić Stanković et al. (2014); these authors, however, used different, purpose-built programs.

For the semi-automatic corpora based on keywords, the keyword seeds were obtained using the "Keyword List" option in the *AntConc*[5] *program for textual analysis; this option compares the frequency of words in the corpus of interest to a reference corpus.*[6] *The reference corpus we used was created ad hoc* from a set of narrative and newspaper texts (comprising a total of 1,584,920 words).[7] N-grams were also extracted with AntConc, using the "Clusters > N-Grams" option and defining the length of the cluster as one, two and three;[8] the extraction of n-grams did not require a reference corpus.

In the creation of the semi-automatic corpora we used the first 50 words/sequences of the respective lists. In order to produce a similar number of words entered in the queries for different methods, tuple length was set to five for corpora based on keywords and unigrams, four for those based on bigrams, and three for those based on trigrams. Apart from conversion into lowercase letters (in the early stages of the selection of seed words), the creation of tuples and the selection of URLs were carried out without manual intervention. As for the construction of the Cyrillic corpora, to ensure full comparability, the tuples obtained for the Latin script were transliterated and directly inserted as queries. Examples of tuples used in the study are provided in Table 1.

	Tuples[1]
Keywords	1 pecite kašika fil stavite minuta mleka ulje šećera kuvajte
Unigrams	stavite preko od vode ne priprema staviti umutiti pa za
Bigrams	"i na" "ostaviti da" "i pecite" "se ohladi" "minuta na" "kada se" "100 g" "30 minuta"
Trigrams	"posolite i pobiberite" "u podmazan pleh" "čokolade za kuvanje" "papirom za pečenje" "u zagrejanoj rerni" "i beli luk"

Table 1: Examples of tuples used in the creation of semi-automatic corpora

[5] www.antlab.sci.waseda.ac.jp/antconc_index.html.

[6] Keywords were extracted using Log-Likelihood values.

[7] The precious help of several colleagues is gratefully acknowledged: most of the reference corpus texts were collected by Duško Vitas, Miloš Utvić and Cvetana Krstev, with the help of students from the Department of Informatics and Librarianship in the Faculty of Philology (University of Belgrade); another portion was kindly provided by Tanja Samardžić.

[8] Clearly, unigrams do not constitute actual clusters, but single words ordered by frequency; it would have also been possible to obtain them using the "Word List" function.

Twenty tuples were created for each seed list. The number of pages to download per tuple was also set to 20, without any limitations as to Internet domain.

4.2 Evaluation of corpora

The two tables provided below recap the main information about the corpora created; Table 2 shows the data for the Latin script, and Table 3 for the Cyrillic script corpora. The information provided for each corpus is the following: the number of URLs from which texts were downloaded, the total number of words in the corpus (excluding the lines containing the URLs), and the percentage of pages belonging to the studied genre, estimated via a sample-based analysis.

	Keywords	**Unigrams**	**Bigrams**	**Trigrams**
N URLs	314	339	324	321
N words	167,605	204,266	191,521	265,382
Relevant URLs	90%	72%	90%	88%

Table 2: Information about the Latin alphabet corpora

	Keywords	**Unigrams**	**Bigrams**	**Trigrams**
N URLs	234	308	88	51
N words	161,290	216,006	524,389	395,366
Relevant URLs	68%	50%	80%	92%

Table 3: Information about the Cyrillic alphabet corpora

As can be seen from the tables, the number of URLs is consistently higher for the Latin than for the Cyrillic script, with the difference being particularly prominent for bigrams and trigrams. However, somewhat unexpectedly, the number of words does not match this difference and is approximately equal for the two scripts for keywords (despite the higher number of URLs for the Latin script), and higher for the Cyrillic script in the remaining cases. In other words, while texts were retrieved from fewer web pages for the Cyrillic script, those pages had more content and led to larger corpora. A first manual inspection indicates that the differences are primarily due to the fact that several entire books in Cyrillic were downloaded, some of which were relevant (e.g. a well-known traditional cookbook called *Patin kuvar* ‘Pata’s Cookbook’), while others were not (these were literary works); the main source of the books was the scribd.

org website. Overall, while roughly even-sized texts were obtained for the Latin script, substantial differences in individual text lengths were noted for Cyrillic; the former was characterised by texts from websites dedicated to recipes, while the latter tended to be more "narrative", often including other text in addition to recipes. Another interesting finding in terms of content is that a lot of the pages in the Cyrillic corpora originated from websites whose content is in some way related to religion, especially the Serbian Orthodox Church (svetosavlje.org, pravoslavna-srbija.com, etc.); particularly prominent are recipes for fasting days, associated with various church holidays.

Of central interest for our analysis are the percentages of relevant URLs. In order to evaluate the different procedures for the creation of gender-oriented corpora in Serbian, we use the precision criterion, based on dividing the number of pages belonging to the target genre by the total number of pages evaluated. For each corpus we randomly selected and evaluated 50 pages (corresponding to between 14.75 and 98% of the total number of pages retrieved). The percentages of relevant URLs obtained on the basis of these samples were used to estimate the composition of entire corpora. It should be highlighted that we also treated as relevant pages in which only one portion of the content belonged to the target genre (e.g. if the text was composed of a recipe followed by comments by the author and/or the readers).

As evident from the bottom rows of Tables 2 and 3, in three out of four cases the results for the Latin script are better than those for the Cyrillic script. For the former, keywords and bigrams constitute equally successful methods (90%), with trigrams falling only slightly behind (88%); unigrams were the sole method that produced a corpus of a clearly lower quality (72%). The situation is somewhat different with the Cyrillic script, where keywords and unigrams pattern together (68 and 50% relevance respectively), opposed to bigrams and trigrams (80 and 92%); i.e., similarly to the Latin script, unigrams were the least successful method, while the best result, which was at the same time the best overall, was obtained using trigrams.

Another (partial) manual inspection of the corpora revealed that many pages not belonging to the target genre had a similar topic (containing mostly articles about food and health), or were isolated irrelevant sections of pages that do belong to the target genre (e.g. reader comments).[9] Among the less relevant content we found URLs with sections, or even entire content, in the wrong language (typically English or Russian), pages similar in genre but not in topic (e.g. household advice), as well as those not related to either the genre or the topic of culinary recipes (e.g. lists of proverbs, descriptions of holiday traditions, news pieces, an article about the history of the library of the Faculty of Electrical Engineering in Belgrade, and fiction); the last category was particularly highly represented in the Cyrillic corpora. It is also interesting to note that more texts in a different language were found for Cyrillic, which also contained instances of Serbian texts written in the Latin script.

The last criterion we evaluated was the rate of overlap between the URLs contained in the various corpora. For the Latin script, we found no URLs shared between two (or more) corpora. Cases of partial overlap (different pages from the same website) between the manual corpus and the semi-automatic corpora were also very few. Among the semi-automatic corpora, instances of partial overlap were not too numerous either, but when they did occur, they mostly concerned relevant websites (different pages of the same target genre website). For the Cyrillic script, overlaps between the manual corpus and the semi-automatic corpora are excluded due to the fact that the manual corpus was constructed from websites using the Latin script. As for the semi-automatic corpora, the situation is partly similar to that described for the Latin script, with some partial overlaps between relevant websites, but also between the irrelevant ones (especially Wikipedia). Overall, it can be concluded that despite being based on seeds derived from the same source, the corpora were very different from each other.

4.3 Discussion

To recap, the evaluation procedure has shown that (1) keywords are a reliable method for creating genre-oriented corpora of Serbian in the Latin script, but less so in the Cyrillic script; (2) unigrams give the lowest precision rate for both scripts; (3) bigrams and trigrams work well for both scripts, (4) more even-sized texts are obtained for the Roman script, and

[9] The presence of irrelevant sections is likely to be related to the page clean-up procedure used by BootCaT, where actual page text sometimes gets removed together with navigation menus and other irrelevant content.

these texts tend to originate from websites dedicated to the target genre, while substantial variation in text size is found for the Cyrillic script, where entire books are downloaded, leading to larger corpora overall. These results are in line with previous work in confirming that automatic selection of web pages based on textual genre is more complex than the identification of pages dealing with the same topic.

Looking at a wider cross-linguistic picture, this study has confirmed that bigrams and trigrams give good results for Serbian regardless of the script used, despite its rich inflectional morphology; the percentages of relevant URLs are in fact higher than those typically obtained for the morphologically poorer English, which tend to be close to 80% (see Dalan, 2012; Bernardini e Ferraresi, 2013). On the other hand, while keywords were an equally successful method for the Latin script, they led to a lower success rate than bigrams and trigrams for Cyrillic. As both sets of corpora were in the same language, and constructed based on the same words and word sequences, this difference can clearly not be due to factors having to do with inflectional morphology. Rather, it can be related to the availability of relevant texts on the Internet, as it appears that Serbian websites dedicated to culinary recipes tend to use the Latin script (cf. also Miličević, 2013, for real estate advertisements); this conclusion is also supported by the manual content analysis, where the Cyrillic corpora are found to deviate more from the target genre, as well as from the topics typically covered in this genre. While further studies looking at other genres (particularly those less tightly associated with a narrow range of topics) are needed in order to reach a more general conclusion, based on the data on culinary recipes, the two scripts used to write the Serbian language can be concluded to have a different presence on the web, likely resulting from some form of a functional split.

5. Conclusion

The results of two parallel studies based on Latin and Cyrillic input words and word sequences point to the script choice being highly relevant for creating semi-automatic genre-oriented Serbian corpora with BootCat. For the Latin script, keywords, bigrams and trigrams all turned out to be reliable as discriminators in the automatic selection of texts belonging to the textual genre of culinary recipes; for the Cyrillic script, however, keywords were a much less successful method than n-grams. The decisive factor behind this difference appears to be the availability of relevant texts in the two scripts, which in turn points to a partial functional split, to be explored further in studies looking at additional genres.

References

Arendholz, J., Bublitz, W., Kirner, M., and Zimmermann, I. (2013). Food for thought - or, what's (in) a recipe? In C. Gerhardt, M. Frobenius, and S. Ley (Eds), *Culinary Linguistics: The Chef's Special* (pp. 119-138). Amsterdam: John Benjamins.

Barbaresi, A. (2014). Finding viable seed URLs for web corpora: A scouting approach and comparative study of available sources. In *Proceedings of the 9th Web as Corpus Workshop* (pp. 1-8).

Baroni, M., and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004* (pp. 1313-1316). Lisbon: ELDA.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. and Mazzoleni, M. (2004). Introducing the “la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In *Proceedings of LREC* 2004 (pp. 1771-1774).

Bernardini, S., and Ferraresi, A. (2013). Old needs, new solutions: Comparable corpora for language professionals. In S. Sharoff, R. Rapp, P. Zweigenbaum and P. Fung (Eds), *Building and Using Comparable Corpora* (pp. 303-319). Dordrecht: Springer.

Bernardini, S., Ferraresi, A. and Gaspari, F. (2010). Institutional academic English in the European context: A web-as-corpus approach to comparing native and non-native language. In A. L. López e C. J. Rosalía (Eds), *Professional English in the European Context: The EHEA Challenge* (pp. 27-53). Bern: Peter Lang.

Biber, D., and Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard and S. Oksefjell (Eds), *Out of Corpora: Studies in Honour of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Crossley, S. A., and Louwerse, M. M. (2007). Multi-dimensional register classification using bi-grams. *International Journal of Corpus Linguistics, 12*, 453-478.

Crowston, K., and Kwasnik, B. H. (2004). A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences - Track 4* (pp. 40100a).

Dalan, E. (2012). *Costruzione automatica di corpora orientati al genere e fraseologia: Il caso delle guide web in inglese degli Atenei europei*. Unpublished MA thesis, University of Bologna.

Gries, S. Th., and Mukherjee, J. (2010). Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes. *International Journal of Corpus Linguistics, 15*, 520-548.

Gries, S. Th., Newman, J., and Shaoul, C. (2011). N-grams and the clustering of registers. *Empirical Language Research Journal, 5*.

Kanaris, I., and Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing and Management, 45*, 499-512.

Louwerse, M. M., and Crossley, S. A. (2006). Dialog act classification using n-gram algorithms. In G. Sutcliffe and R. Goebel (Eds), *Proceedings of the 19th International Florida Artifficial Intelligence Research Society* (pp. 758-763). Menlo Park, CA: AAAI Press.

Ljubešić, N. (2013). Helping *BootCaT* to catch the Babel fish: Getting encoding, content and language right. Presentation at the workshop "BootCaTters of the world unite!", Forlì, 24 June 2013.

Ljubešić, N., and Klubička, F. (2014). {bs,hr,sr}WaC — Web corpora of Bosnian, Croatian and Serbian. In F. Bildhauer and R. Schäfer (Eds), *Proceedings of the 9th Web as Corpus Workshop (WaC-9) - @ EACL 2014* (29-35). Gothenburg, Sweden.

Miličević, M. (2013). Izrada latiničnih i ćiriličnih korpusa srpskog jezika uz pomoć softvera BootCaT. Paper presented at the conference "Languages and Cultures in Time and Space 3". Novi Sad, 16 November 2013.

Miličević, M., Bernardini, S., and Ferraresi, A. (2014). Costruzione semiautomatica di corpora orientati al genere in lingue morfologicamente ricche: Un paragone fra l'italiano e il serbo. *Italica Belgradensia, 1/2014*, 99-114.

Pavlović-Lažetić, G., Vitas, D., and Krstev, C. (2004). Towards Full Lexical Recognition. In P. Sojka, I. Kopecek, and K. Pala (Eds), *Text, Speech and Dialogue 2004 – Lecture Notes in Artificial Intelligence 3206* (pp. 179-186). Berlin: Springer-Verlag.

Pešikan, M., Jerković, J., and Pižurica, M. (2010). *Pravopis srpskoga jezika*. New and expanded edition. Novi Sad: Matica srpska.

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. *Proceedings of Web as Corpus Workshop.* Louvain-la-Neuve.

Sinclair, J. (2005). Corpus and text: Basic principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 1-16). Oxford: Oxbow Books.

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Utvić, M. (2013). *Izgradnja referentnog korpusa savremenog srpskog jezika*. Unpublished PhD thesis, University of Belgrade.

Vitas, D., Krstev, C., Obradović, I., Popović, L., and Pavlović-Lažetić, G. (2003). An overview of resources and basic tools for the processing of Serbian written texts. *Proceedings of the Workshop on Balkan Language Resources and Tools, 1st Balkan Conference in Informatics*. Thessaloniki.

Vitas, D., Pavlović-Lažetić, G., and Krstev, C. (2006). About word length counting in Serbian. In P. Gryzbek (Eds), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues* (pp. 301-317). Dordrecht: Springer.

Vujičić Stanković, S., Krstev, C., and Vitas, D. (2014). Enriching Serbian WordNet and electronic dictionaries with terms from the culinary domain. In H. Orav, C. Fellbaume and P. Vossan (eds), *Proceedings of the Seventh Global WordNet Conference* (pp. 127-132). University of Tartu, Estonia.

Maja P. Miličević

Sažetak
POLUAUTOMATSKA IZRADA UPOREDIVIH ĆIRILIČNIH I LATINIČNIH ŽANROVSKIH KORPUSA SRPSKOG JEZIKA

Elektronski korpusi već više decenija predstavljaju značajan izvor podataka za lingvistička istraživanja. Međutim, u proučavanju srpskog jezika iz korpusne perspektive javlja se specifičan problem vezan za upotrebu dva različita pisma. Veći korpusi, poput referentnog Korpusa savremenog srpskog jezika i Internet korpusa srWaC, za kodiranje tekstova koriste latinično pismo, čime izostaju resursi za proučavanje pojava za koje je potrebno uporediti ćirilične i latinične tekstove. Jedan mogući način prevazilaženja ovog problema jeste samostalna izrada korpusa tekstova sa Interneta upotrebom softverskog paketa BootCaT. Konkretno, u radu se prikazuje mogućnost upotrebe ovog paketa za paralelnu izradu žanrovski orijentisanih korpusa na ćiriličnom i latiničnom pismu, pri čemu se porede četiri različite metode, zasnovane na ključnim rečima, unigramima, bigramima i trigramima. Procedura na kojoj softver počiva, zasnovana na zadavanju (nizova) reči na osnovu kojih se poluautomatski biraju Internet stranice čiji sadržaj će biti preuzet i uključen u korpus, omogućuje unošenje istih (nizova) reči na svakom pismu zasebno. Proceduru u radu ilustrujemo na primeru izrade korpusa recepata. Rezultati pokazuju da se dobijeni ćiriličini i latinični korpusi međusobno razlikuju: dok su bigrami i trigrami podjednako pouzdani indikatori žanra u oba slučaja, samo se za latinično pismo podjednako dobar rezultat dobija i putem ključnih reči. Zaključuje se da je ciljni žanr na Internetu dominantno predstavljen latiničnim tekstovima, dok ćirilični upiti zasnovani na istim rečima kao rezultat često daju čitave knjige posvećene receptima ili tekstove opštijeg tipa, poput odrednica sa Vikipedije. Ovakav rezultat ukazuje na delimičnu funkcionalnu podelu između pisama srpskog jezika.

(Footnotes)

[1] Keywords: 1, bake, spoon.nom, cream. nom/acc, put; minutes.gen, milk.gen, oil. nom/acc, sugar.gen, cook. Unigrams: put, on-top, of, water.gen, no; preparation. nom, put, beat, then, for. Bigrams: “and on”, “leave that”, “and bake”, “cools down”; “minutes.gen on”, “when it”, “100 g”, “30 minutes.gen”. Trigrams: “salt and pepper”, “in greased.acc pan.ACC”, “chocolate.gen for cooking.acc”; “paper. ins for baking.acc”, “in preheated.loc oven.loc”, “and garlic.nom/acc”.