1.01 UDC: 004.934:821.163.41

Aniko Kovač,[*] Maja Marković[**]

A Mixed-principle Rule-based Approach to the Automatic Syllabification of Serbian

IZVLEČEK

MEŠANI PRISTOP K AVTOMATSKEMU ZLOGOVANJU V SRBŠČINI NA PODLAGI NAČEL IN PRAVIL

V tem prispevku predstavljamo mešani pristop k avtomatskemu zlogovanju v srbščini na podlagi načel in pravil, ki temelji na predpisnih pravilih tradicionalne slovnice v kombinaciji z načelom zaporedja glede na zvočnost (Sonority Sequencing Principle). Proučujemo težave in omejitve obeh uveljavljenih pristopov, ki temeljita na zbirki pravil in zvočnosti; vpeljujemo algoritem, ki uporablja oba načina za doseganje natančnejše členitve besed na zloge, ki bi bila skladnejša z intuicijo rojenih govorcev; in predstavljamo statistične podatke, povezane z razporeditvijo zlogov in njihovo strukturo v srbščini.

Ključne besede: zlog, pristop na podlagi pravil, zvočnost, računalniško jezikoslovje, fonologija

ABSTRACT

In this paper we present a mixed-principle rule-based approach to the automatic syllabification of Serbian, based on prescriptive rules from traditional grammar in combination with the Sonority Sequencing Principle. We explore the problems and limitations of the existing rule set and sonority-based approaches, introduce an algorithm that utilizes both means in an attempt to produce a more accurate segmentation of words into syllables that is

[*] Department of Language Science and Technology, Saarland University Campus A2 2, 66123 Saarbrücken, Germany, anikok@coli.uni-saarland.de

[**] Department of English Language and Literature, Faculty of Philosophy, University of Novi Sad, Dr Zorana Đinđića 2, 21000 Novi Sad, Serbia, majamarkovic@ff.uns.ac.rs

better aligned with the intuition of the native speakers, and present the statistical data related to the distribution of syllables and their structure in Serbian.

Keywords: syllable, rule-based approach, sonority, computational linguistics, phonology

Introduction

Syllables have been considered — although not unequivocally (cf. Koehler 1966) — to be one of the basic units in phonology constituting the minimal units of pronunciation, and to play a role in prosody, phonotactics, and phonological processing (Ladefoged and Johnson 2014). The role of the segmentation of words into syllables and their distributional properties began to see an increase in importance in speech technologies in the 1990s (Iacoponi and Savy 2011), most notably in the areas of speech recognition (SR) and text-to-speech synthesis (TTS).

Syllable segmentation today plays a role in speech technologies on the segmental level — conditioning the length of segmental units such as consonants and vowels — as well as on the prosodic level — governing rhythmical alternations (Bigi and Petrone 2014). Syllable segmentation is also a key component in hyphenation (e.g. Kaplar et al. 2018), although it should be noted that, at least in Serbian, hyphenation is governed by a partially diverging set of rules from those governing syllabification[1]. Syllable distribution data is also of crucial importance for psycholinguistic experiments, as syllable frequency has been shown to play a role in the processing of words (e.g. Barber et al. 2004; Cholin et al. 2006; Cholin and Levelt 2009). Developing an automatic system of syllabification allows for the segmentation of large-scale language corpora needed for the development of automatic systems or the extraction of relevant data related to frequency syllable distributions, which would otherwise require a large number of trained annotators and would be a resource and cost heavy undertaking.

The two generally distinguishable approaches to automatic syllabification are rule-based versus data-driven approaches (Marchand et al. 2009). While data-driven approaches have taken over many aspects of natural language processing, and there are a number of data-driven models of syllable segmentation using artificial neural networks (e.g. Daelemans and van den Bosch 1992; Hunt 1993; Stoianov et al. 1997; Landsiedel et al. 2011), the unavailability of segmented data for Serbian makes rule-based approaches the only viable option for automatic syllabification in Serbian.

To the best of our knowledge, there is a single publicly available attempt at developing a rule-based syllabifier for Serbian by Kaplar et al. (2018). In this paper we lay out a number of problems and limitations with the ruleset used in their syllabification system and why relying on the existing set of prescriptive rule descriptions from traditional grammar is insufficient to capture and describe a syllabification system that

[1] For example, hyphenation rules ban the segmentation after a syllable consisting of a single vowel at word onset, while this segmentation is allowed and expected according to the rules of syllabification.

is aligned with the intuition of native speakers of Serbian. A relatable attempt at automatic syllabification was developed by Meštrović et al. (2015) for Croatian, the key difference between their work and ours being in the principle behind the syllabification algorithm which in their case relied solely on the onset maximization principle — limiting possible syllable onsets to valid onsets at the beginning of words. Taking into account Morelli's (1999) limitations on possible syllable onsets in Serbo-Croatian, the onset maximization principle employed by Meštrović et al. could be considered a comparatively liberal system. In order to attempt to constrain our syllabifer, we are decided on a different approach that will not rely on onset maximization, but rather a combination of a number of alternative principles.

In this paper we present a mixed-principle rule-based approach to the syllabification of Serbian. Our starting set of rules is based on the *Gramatika srpskoga jezika* by Stanojčić and Popović (2005), a prescriptive textbook for Serbian grammar that presents a set of rule descriptions for the segmentation of words into syllables. In a previous version of our syllabification algorithm (Kovač and Marković 2018), we made a number of changes to the rule descriptions of Stanojčić and Popović (2005) as the formulation of some of the descriptions proved to be redundant, some were example-based and not specific enough for a formal implementation, and we also expanded them with three added modifications related to the treatment of nasals and the alveolar sonorant /*r*/ based on Kašić (2014) and the treatment of alveolar sonorants /*l*/ and /*n*/ based on Zec (2000). In this paper we extend our previous algorithm to include a module for validating the structure of syllables in terms of their compliance with the Sonority Sequencing Principle (SSP), thus further fine-tuning the accuracy of our segmentation, and resolving a number of problems noted in our earlier implementation.

The goal of the paper is threefold: i) to improve our system for automatic rule-based syllabification for Serbian based on the formalization of existing rule descriptions by the addition of the sonority sequencing validation module, ii) to provide an analysis of the outcomes of the automatic syllabification process in order to address possible theoretical considerations and serve as a basis for the development of future syllabifiers, and iii) to present statistical data related to the distribution of syllables and their structure in Serbian.

Prescriptive Rule Descriptions

Our starting set of rules was based on the formalization of the rule descriptions governing the segmentation of words into syllables from the *Gramatika srpskoga jezika* by Stanojčić and Popović (2005). Being a prescriptive textbook on Serbian grammar used at a high school level by all student profiles, we expected these rules to constitute the common knowledge base shared by the majority of native speakers.

Regarding syllable boundaries, Stanojčić and Popović (2005, 37) establish the following general rule (1).

(1) *In words made up of multiple phonemes, consonants, sonorants and vowels, the syllable boundary comes after the vowel and before the consonant (e.g.* či-ta-ti *[*to read*]).*

In addition to this general rule, they list the following rules — (2), (3), (4), (5) and (6) — that further specify medial syllable boundaries depending on consonant manner of articulation.

(2) *Medially, in a consonant cluster which has an affricate or fricative sound in its initial position, the syllable boundary will be before that consonant cluster (e.g.* po-šta *[*post*],* ma-čka *[*cat*]).*
(3) *The syllable boundary will be before a consonant cluster if, in a consonant cluster found medially in a word, the second position in the cluster is occupied by one of the sonorants /*v*/, /*j*/, /*r*/, /*l*/ or /*ʎ*/ preceded by any other consonant besides a sonorant (e.g.* sve-tlost *[*light*]).*
(4) *If a consonant cluster consists of two sonorants, the syllable boundary will be between them so that one sonorant belongs to the preceding, and one sonorant belongs to the following syllable (e.g.* lom-ljen *[*broken*]).*
(5) *If a consonant cluster consists of a plosive in its initial position and some other consonant except the sonorants /*j*/, /*v*/, /*l*/, /*ʎ*/ and /*r*/, the syllable boundary will be between the consonants (e.g.* lep-tir *[*butterfly*]).*
(6) *If in a cluster of two sonorants, the second position is occupied by the sonorant /*j*/ from je corresponding to the ijekavica dialect to /*e*/ in the ekavica dialect, the syllable boundary will be before that group (e.g.* čo-vjek *[*man*]).*

Stanojčić and Popović (2005, 32) also introduce the rule descriptions (7) and (8) to define when the sonorants /*r*/, /*l*/, and /*n*/ constitute syllable nuclei.

(7) *The sonorant /*r*/ can be a syllable carrier in standard Serbian when:*
   a. it is found medially between two consonants (e.g. tr-ča-ti *[*to run*]),*
   b. it is found initially before a consonant (e.g. r-va-ti se *[*to wrestle*]),*
   c. it is found after a vowel in compounds (e.g. za-r-đa-ti *[*to rust*]),*
   *d. before /*o*/ that is realized as an /*l*/ in other members of the paradigm (e.g.* o-tr-o *(m.) from* o-tr-la *(f.) [*wiped*]).*
(8) *The other two alveolar sonorants, /*l*/ and /*n*/ can be syllable carriers in dialectal toponyms (e.g.* Stlp, Vlča glava, Žlne*) or foreign toponyms (e.g.* Vltava, Plzen*) but also in other personal names (e.g. English* Idn *or Arabic* Ibn-Saud*), and in the word* bicikl *[*bicycle*].*

Revising the Existing Rule Set

The development of our syllabification algorithm has been an iterative process testing the existing rule set and making changes as needed. While other authors (e.g. Kaplar et al. 2018) used the rule descriptions of Stanojčić and Popović (2005) directly to implement a software architecture for syllabification in Serbian, we have found a number of problems with this approach.

The definition of the rule description under (1) causes the initial member of a consonant cluster in the rule descriptions under (2)–(6) to be understood as the first consonant following a vowel. However, given that the sonorants */r/*, */l/*, and */n/* can also constitute syllable nuclei in Serbian in certain positions, as presented under rule descriptions (7) and (8), a more precise definition would be that the initial member of a consonant cluster is the first consonant following an element that constitutes a syllable nucleus. The general rule under (1) should be then revised as follows.

(1*) *In words made up of multiple phonemes, consonants, sonorants and vowels, the syllable boundary comes after the vowel or sonorants /*r*/, /*l*/, and /n/ in syllable bearing positions and before the consonant (e.g.* či-ta-ti *[*to read*],* tr-ča-ti *[*to run*]).*

In addition to our expansion of the general rule presented under (1) to include the syllable bearing sonorants, while formalizing the rule descriptions via finite-state automata, rules (2) and (3) proved to be redundant as they produced identical outcomes to the general rule under (1*). Because of this, these rules were disregarded in our syllabification algorithm.

During our early testing of the verbatim implementation of the rule descriptions, we also noticed that the existing rule descriptions treated a consonant cluster consisting of a nasal in initial position followed by a consonant that is not one of the sonorants */j/*, */v/*, */l/*, */ʎ/*, and */r/* as a part of the following syllable onset, producing outcomes such as: *gu-ngula* [*commotion*], *mo-mci* [*guys*], *ka-ncelarije* [*offices*], *su-nce* [*sun*], etc. Contrary to Stanojčić and Popović (2005), authors such as Kašić (2014) argue that nasals should be treated analogously to plosives during syllabification because there is a complete occlusion in the oral cavity during their production. If this principle were to be employed, rule (5) should be revised as follows.

(5*) *If a consonant cluster consists of a plosive or nasal in its initial position and some other consonant except the sonorants /j/, /v/, /l/, /ʎ/,* and */r/, the syllable boundary will be between the consonants.*

Following rule (5*), the examples above would then be segmented as: *gun-gula* [*commotion*], *mom-ci* [*guys*], *kan-celarije* [*offices*], *sun-ce* [*sun*], etc. Even though in the earlier implementation of our syllabifier (Kovač and Marković 2018) we did not want to employ the Sonority Sequencing Principle (SSP), we opted for the treatment of nasals by Kašić (2014) in our implementation, which respected the limitations put forward by the Sonority Hierarchy, and was more in line with native speaker intuition.

The Sonority Hierarchy

Sonority Theory accounts for the organization of segments into well-formed sequences, both within the syllable and across syllabic boundaries. This organization is driven by principles of sonority, a property that is used as the basis of ranking all sounds along a scale, from less sonorous to more sonorous ones. Although there is a general consensus that segments are ranked by their inherent sonority, the notion of sonority itself is not unambiguously described in the phonetic and phonological literature. Among the phonetic approaches, Ladefoged (1982) defines sonority as the perceptual salience or loudness of a sound, and Bloch and Trager (1942; according to Goldsmith 1995) define it as the amount of airflow in the resonance chamber. For others, sonority is dependent on multiple phonetic parameters (Ohala and Kawasaki 1984; Ohala 1990; Butt 1992). In the phonological literature, sonority is generally defined as a multi-valued feature (Foley 1972; Hankamer and Aissen 1974; Selkirk 1984), although there are also authors who argue that it is derivable from the more basic binary features of phonological theory (Clements 1990). Other questions that are often addressed are whether sonority scales are universal or language-specific, allowing freedom to languages in assigning sonority values, and how fine-grained distinctions sonority scales should capture. For example, Clements' universal sonority scale includes only four major classes of consonants (Clements 1990), ranked from least sonorous to most sonorous, as in (i):

(i) O < N < L < G
(O = obstruents, N = nasals, L = liquids, G = glides)

Selkirk (1984, 112) proposes a much more detailed scale, which divides all sounds into 11 groups, assuming more subtle differences in sonority values. Selkirk also states that the sonority indices may not be as important in themselves as the sonority relations that they express. Selkirk's scale of sonority in consonants is given in (ii):

(ii) p, t, k < b, d, g < f, θ < v, z, ð < s < m, n < l < r

Sonority scales serve as the basis of constructing segment sequences within syllables. The universal cross-linguistic generalization is that in the sequence of segments, the one ranking highest on the sonority scale constitutes the peak of the syllable, i.e. it is the syllabic nucleus. As for the other segments around the nucleus, they are organized so that the more sonorous ones are closer to the nucleus, and less sonorous ones are more distant. This generalization is referred to as Sonority Sequencing Principle (SSP). Thus a syllable with an ascending sonority slope in the onset and a descending slope in the coda, such as, for example *blunt*, is a well-formed syllable, whereas **lbutn* is prohibited, due to the violation of the SSP. Adopting thee SSP often solves the problems of syllabic consonants, since they generally occur in environments where they constitute a sonority peak, as in the Serbian word *pr-vi*.

The Need for Sonority

Apart from the segmentation of nasals analogously to plosives following Kašić (2014) that relied on principles of the SSP, in our initial attempt at the formalization of the rule description under (8) of Stanojčić and Popović (2005) we had to rely on sonority to define the criteria for when the alveolar sonorants /*l*/ and /*n*/ act as syllable nuclei.

As Stanojčić and Popović gave no formal criteria defining the contexts of syllable bearing /*l*/ and /*n*/, our initial attempt to draw on generalizations based on their examples for syllable carrying /*l*/ (*Stlp, Vlča glava, Žlne, Vlava, Plzen*) and /*n*/ (*Idn, Ibn-Saud*). In analogy to the rules descriptions under (7a) and (7b) and our added rule (7c*) defining the contexts in which the alveolar phoneme /*r*/ can act as a syllable nucleus, we implemented rule (8*) to define the conditions under which the phonemes /*l*/ and /*n*/ can act as syllable bearing nuclei.

(8*) *The other two alveolar sonorants, /*l*/ and /*n*/, can be syllable carriers if they are found:*
a) medially between two consonants,
b) initially before a consonant, or
c) finally after a consonant.

However, the formulation under (8*) allowed for outcomes such as: *Be-rn, Ka-rl, erla-jn, Kla-jn, kasa-rn-skim, Linko-ln, Va-jl-om,* etc. in which the phonemes /*l*/ and /*n*/ identified as syllable nuclei have a lower sonority level than the consonants in their onset or coda. Because the phonemes /*r*/ and /*j*/ are more sonorous than the phonemes /*l*/ and /*n*/, and the lateral phoneme /*l*/ is more sonorous than the nasal phoneme /*n*/, native speakers do not perceive the elements of lower sonority as syllable nuclei in these contexts. Zec (2000) states that alveolar sonorants can be syllable bearing elements in Serbian only in contexts in which there is no segment of a higher level of sonority in their immediate vicinity. Because of this, we needed to further specify rule (8*) to take sonority constraints into consideration as follows.

(8**) *The other two alveolar sonorants, /*l*/ and /*n*/, can be syllable carriers if they are found:*
a) medially between two consonants of lower sonority,
b) initially before a consonant of lower sonority, or
c) finally after a consonant of lower sonority.

It turns out that this principle can also account for the behavior of the syllable bearing /*r*/ in Serbian. In fact, it does not only provide a general account for consonantal syllabic nuclei in Serbian that subsumes the rules under (7) and (8**) it also accounts for our extension of rule (7) that keeps the the consonant cluster /*rje*/ of the ijekavica dialect unsegmented in initial position[2]. Because the phoneme /*j*/ has a higher level of sonority than /*r*/, the phoneme /*r*/ should not be treated as a syllable nucleus initially in words such as *rjeka* [*river*].

In our previous implementation of the syllabifier (Kovač and Marković 2018), we attempted to limit our reliance on the Sonority Sequencing Principle to the cases above. However, during the evaluation of our algorithm, we encountered a number of syllable structures that were unexpected due to their absence from the onset maximization approach to syllabification developed for Croatian by Meštrović et al. (2015). Namely, we encountered the syllable structure CCCCCVC in *mo-na-rhstvom* [*with the monarchy*], the structure CCCCCV in the words *se-rbska* [*Serbian*], *ca-rstva* [*kingdoms*], and *sta-ra-te-ljstva* [*custody*], and the structure CCCCVC in *se-rbskom* [*Serbian*], *de-jstvom* [*with effect*], *vo-đstvom* [*leadership*], *spo-rtskim* [*sport*], and *a-lpskog* [*alpine*].

The way we attempted to remedy this issue was to limit the syllable onset length three-syllable clusters, which is the maximum length of non-syllabic consonant clusters word initially in Serbian (Kašić 2014). While this constraint, in combination with rules (5) and (6), resolved the issues in the examples we encountered — with this limitation, they are segmented as *mo-narh-stvom* [*with the monarchy*], *serb-ska* [*Serbian*] (three-syllable onset limitation + rule (5)), *car-stva* [*kingdoms*], *sta-ra-telj--stva* [*custody*], *serb-skom* [*Serbian*], *dej-stvom* [*with effect*], *vođ-stvom* [*leadership*], *sport-skim* [*sport*], *alp-sko*g [*alpine*] — some medial clusters with a syllabic consonant still remained a problem. For example, in the word *najstrpljiviji* [*most patient*], which contains a syllabic /*r*/, the syllable boundary that would be placed between /*na*/ and /*jstr*/ — *na-jstr-pljiviji* — which does not coincide with native speaker intuition. The Sonority Sequencing Principle seems like a perfect solution for this cases, as it would require the structure of a syllable to follow a sonority scale, with the syllable nucleus being the most sonorous element, while sonority would gradually decrease towards the periphery of the syllable (Zec 2000). With this added sonority requirement, the phoneme /*j*/, being more sonorous than /*s*/ and /*t*/, would have to constitute a part of the previous syllable where it would be of a lower sonority when compared to its neighbouring syllable bearing vowel, and the syllable boundary would be *naj-str-pljiviji* which is in line with native speaker intuition.

As a final check following rules (1)–(8**), we add rule (9) that has the ability to shift the syllable boundary in order to avoid a violation of the sonority hierarchy.

(9) *If the syllable structure resulting from rules (1)–(8**) does not conform to the Sonority Sequencing Principle, move the boundary so that the phoneme violating the sonority sequence is shifted into the neighboring syllable.*

[2] It should be noted that while sonority sequencing accounts for the non-syllabic treatment of /*r*/ before /*je*/ in initial position, our rule extension is still needed as it has a more general scope than the sonority rule and accounts for segmentation in medial positions as well (e.g. in words such as *isko-rje-nilo* [*eradicated*]).

An Adapted Sonority Hierarchy

In our sonority sequencing module, we relied on a combination of Selkirk's (1984) sonority scale, the sonority apertures for Serbian described by Subotić et al. (2012), and some notes on sonority sequencing in Serbian from Zec (2000). Our sonority scale is shown under (iii).

(iii) p, t, k < b, d, g < ts, tʃ, tɕ < f, ʃ, h < v, z, ʒ < s < m, n, ɲ < l, ʎ < j, r < a, e, i o, u

The highest sonority group in our implementation was made up by the vowels of Serbian. As vowels constitute syllable nuclei and there can only be a single vowel per syllable, we did not need to make a distinction between three sonority apertures of vowels (i, u < e, o < a) as it is the case in the hierarchy of Subotić et al. (2012). Following Selkirk (1984), we divided sonorants into three sonority classes, and following Zec (2000), we treated liquids as more sonorous than nasals, and, within liquids, the phoneme /r/ as more sonorous than laterals. For the needs of our implementation, we treated the phoneme /r/ and glide /j/ as a single sonority group, although from a theoretical standpoint /j/ would be considered as more sonorous out of the two given its semi-vowel nature. We opted for treating /s/ as an element of higher sonority than voiced fricative despite its voiceless nature following Selkirk (1984), and expanded Selkirk's hierarchy with the addition of affricates between voiceless fricatives and voiced plosives as a parallel to the aperture order presented by Subotić et al. (2012).

It is important to note that there are sequences which clearly do not conform with the SSP in a number of languages, and which may undermine the relevance and power of the sonority hierarchy. A very common pattern, found across a number of unrelated languages, is the possibility of an /s/ + plosive sequence in the syllable onset, which would be in clear violation if we were to adopt the sonority scale outlined above. In Serbian, there is a known ambiguity in syllable segmentation in the case of continuant fricative phonemes. For example, the word *postaviti* [*to set*] can be syllabified as both *po-sta-vi-ti* and *pos-ta-vi-ti* (Gvozdanović 2011). We therefore adopt the view put forward in Morelli (1999), who argues that fricatives and plosives may be treated as a single class with respect to sonority in these cases — since splitting them into separate classes would make wrong typological predictions — and add an exception to our sonority sequencing module that leaves fricative + plosive sequences as a viable sequence in the syllable onset.

Our Algorithm[3]

Our mixed-principle syllabification algorithms consists of the following steps:

[3] Our implementation of the algorithm can be found at https://github.com/versi-regular/rule-based_syllabifier_sr, licensed under the GNU General Public License v3.0. It was developed using Python 3.x and processes 10380 tokens/s on average estimated on a 4,681,713 token corpus processed on an Intel® Core™ i5-3210M CPU @ 2.50GHz with 8.00 GBs of DDR3L-1600 SODIMM, including pre-processing, clean-up, and transliteration.

I. Identify vowels in the word and mark their positions as positions capable of constituting syllable nuclei (based on (1)).
II. If a word contains the letters *l*, *n* or the letter *r* not followed by the sequence *je* in the center of a consonant cluster consisting of elements of lower sonority or at the beginning or a word followed by a consonant of lower sonority, or the letters *l* or *n* at the end of a word preceded by a consonant of lower sonority, treat those positions in the word as capable of constituting syllable nuclei (based on (1*), (7), and (8**)).
III. For each position identified as capable of constituting a syllable nucleus:
   A. If it is followed by a sequence of two sonorants, mark the syllable boundary between the two sonorants (based on (4)), except if the second sonorant is *j* and it is followed by *e*. If the second sonorant is *j* followed by *e*, mark the syllable boundary before the sonorant cluster (based on (6)).
   B. If it is followed by a sequence of a plosive or nasal and a plosive, fricative, affricate or nasal, mark the syllable boundary between the two consonants (based on (5*)).
   C. In all other cases mark the syllable boundary after the syllable nucleus (based on (1*)).
IV. Run a recursive sonority check (based on (9)):
   A. If the word consists of more than one syllable, convert the syllable structures identified by the previous steps into sonority group values.
   B. For each syllable, check if there is a violation of the SSP at the edges of the syllable ignoring the check at the onset on the word-initial syllable and the check in the coda of the word-final syllable.
   C. If a violation found is a sequence of a fricative followed by a plosive in the onset, ignore the violation.
   D. If there is a violation, remove the letter from the edge of the syllable, and add it onto the neighboring syllable.
   E. Repeat until no violation is found.

Syllable Distribution Data

In this section, we present the statistical distribution data of syllables in Serbian based on our updated syllabification process applied to the Serbian Lemmatized and PoS Annotated Corpus *SrpLemKor* (Popović 2010; Utvić 2011). We chose *SrpLemKor* for our analysis, because its annotation allowed us to filter out numbers, Roman numerals, abbreviations and non-Serbian words or suffixes in compounds (at least to some extent) and thus reduce noise in the data.

The following results show the syllable distribution statistics based on 3,648,543 non-unique word-forms (word tokens) from *SrpLemKor*. From a total of 4,681,713 entities (punctuation and word tokens) in our version of the corpus, 113,679 (2.43%)

entities of texts #260, #4505 and #4517 were excluded because the files contained faulty encoding. Based on corpus tags, we excluded 919,161 (19.63%) entities tagged PUNCT (punctuation), SENT (sentence separator full-stops), RN (Roman numerals), NUM @card@ (Arabic numerals), ABB (abbreviations) and ? (non-Serbian words and other uncategorized entries). An additional 815 (0.02%) entities that contained the characters w, q and x were removed in an attempt to further reduce noise stemming from foreign words, as not all foreign words were tagged as such in the corpus. In the process of syllabification, an additional 12,877 (0.28%) entities were removed as they were solely made up of consonant clusters with no available syllable nucleus candidate.

Syllable Type Distributions in Serbian

In the 3,648,543 word-forms from *SrpLemKor*, a total of 8,196,771 syllables were identified. Table 1 presents the syllable type distribution based on our mixed-principle syllabification algorithm.

Table 1: Syllable structure distribution of syllables in the *SrpLemKor* corpus

Syllable structure	No.of instances	Percent
CV	5030622	61.37321636
CCV	938275	11.44688561
CVC	913603	11.14588903
V	852854	10.40475573
CCVC	218126	2.661121068
VC	141980	1.7321455
CCCV	56168	0.685245446
CVCC	20339	0.248134296
CCCVC	14362	0.175215338
CCVCC	6274	0.076542336
VCC	2234	0.027254635
CCCCV	780	0.009515942
CVCCC	731	0.008918146
CCCVCC	170	0.002073987
CCCCVC	84	0.001024794
VCCC	67	0.000817395
CCCCVC	36	0.000439197
Other	66	0.000805195
Total	8196771	100

These results show the distribution of syllables in a somewhat noisy data. We found there are still foreign words annotated as non-foreign in the corpus constituting some of the less-frequent syllable structures listed as "Other" in Table 1. For example, an instance of the syllable structure VCCCCC was found to correspond to the segmentation of the German word *Pe-itscht* [*lashes*], the syllable structure CCCCVCCC was identified in the German word *Fle-i-schmarkt* [*meat market*], and the structure CCCCCVC was found in the German word *Gle-i-chschal-tung* [*co-ordination*]. The structure CCCCCCVC was found in the German word *Na-chtschat-ten* [*nightshade*] and in the toponym *CRYSLER*. The syllable structure CCVCCCC was found in the source transcription of the last name *Pe-tritsch* and in the English word *knights*. The syllable structure CCCVCCC was identified to be a part of the German words *Wol-fsmilch* [*spurge*] and *E-in-ge-schickt* [*sent in*] and to correspond to the English word *string*. The syllable structure CCCCCCV was identified in the German words *We-i-hna-chtsbra-e-u-che* [*Christmas trees*], *Stor-chschna-bel* [Crane's bill], while the structure CCCCCV was found in the words *Re-chtsge-schi-chte* [*history of law*] and *Um-gan-gsspra-che* [*vernacular*], as well as in the sequences *šttske* and *su-žnjstva*. The syllable structure CCCCVCC was found in the German word *Ze-it-schrift* [*magazine*], and in multiple occurrences of the source spelling of the last names *Schmidt* and *Rot-hchild*. The structure VCCCC was found in the German words *Deutsch* [*German*], *Ernst* [*seriousness*], in the sequence *der-demnaechst* [*soon*], and in the strings *ikvbv* and *EHCmc*. As can be seen from the examples above, besides foreign origin words, noise in the data can also be found in typos and strings we did not manage to identify. Another example of such string was *ngBpJKTnQ* identified as the structure VCCCCCCCC. Most structures identified as CVCCCC were the result of typos, e.g. serbsk, kra-levstv, pod-dan-stv, carstv, slav-jansk, ju-go-slo-venskg, cr-no-gorskg, but also foreign origin names, e.g. *Hirsch, Herbst, Lokotsch,* and *Worlds* in additions to strings such as *majnds and Gorrrr*. In addition to these, one occurrence of the syllable structure CVCCCCCCCC that stood for the onomatopoeic vulgarism *mrššššššš* [*go away*].

We also found 2 syllable structures that differed from the structures found by Meštrović et al. (2005) for Croatian. The structure CCCCVC was identified in the words *vo-đstvom* [*with leadership*], *za-ko-no-da-vstvom* [*with legislature*], *mo-nar-hstvom* [*with monkhood*], *lu-ka-vstvom* [*with slyness*], *be-zzglob-na* [*without wrists*], and in the paradigm members of the word *po-sthlad-no-ra-to-vski* [*post-cold-war*]. It also occurred in the Russian word *Zdra-vstvuj* [hello], in the German-origin word *Ha-up-tstrum-fi-rer* [*mid-level commander*], in the German *Ra-u-schmit-tel* [*intoxicant*] and *Li-e-be-spflan-ze* [*love plant*] and in the misspelled Serbian words *pri-ja-tljskih* [*friendly*] and *kvdrat* [*square*]. The structure CCCCV was found in the words *bi-vstvu* [*existence*], *va-zdu-ho-plo-vstvo* [*aviation*], *kra-lje-vstva* [*kingdoms*], *zdra-vstve-noj* [*health*], *vo-đstvo* [*leadership*], *ču-vstva* [*feeling*], *pre-i-mu-ćstva* [*advantages*], and *mo-gu-ćstvu* [*possibility*]. It also occurred in German words such as *Pfin-gstro-se* [*peony*], *Ke-u-schhe-it* [*chastity*], *Schne-e-glo-ec-kchen* [*snowdrop*], *Schne-e-ro-se* [*Chrismas rose*], *Ge-i-sskle-e* [*cystus*], *Vol-ksbra-uch* [*popular custom*], *Vol-ksgla-u-ben* [*popular belief*], *Schri-ften* [*regulations*], *Schlu-e-ssel-blu-me* [*cowslip*], and more. We discuss the implications of these for our syllabification algorithm in the Discussion section below.

Syllable Type Positional Distributions in Serbian

We also examined the syllable type frequencies with respect to their position in a word. Four positional frequencies are presented in Table 2: syllable type frequencies in monosyllabic words, and syllables type frequencies in the initial position, in medial positions, and in the final position of polysyllabic words.

Table 2: Syllable structure distribution of syllables in the *SrpLemKor* corpus categorized by position

Syllable structure	Monosyllabic words		Polysyllabic words					
	MONO		INITIAL		MEDIAL		FINAL	
	No.of instances	Percent	No.of instances	Percent	No.of instances	Percent	No.of instances	Percent
CV	612214	50.382	1356771	56.064	1476732	68.956	1584905	65.49
CCV	62244	5.122	372181	15.379	305247	14.254	198603	8.21
CVC	129337	10.644	178859	7.391	211979	9.898	393428	16.26
V	301295	24.795	369133	15.253	61241	2.860	121185	5.01
CCVC	35428	2.916	50383	2.082	53397	2.493	78918	3.26
VC	64038	5.270	67539	2.791	7123	0.333	3280	0.14
CCCV	174	0.014	19754	0.816	20260	0.946	15980	0.66
CVCC	5368	0.442	1052	0.043	695	0.032	13224	0.55
CCCVC	1490	0.123	3976	0.164	4427	0.207	4469	0.18
CCVCC	1635	0.135	206	0.009	17	0.001	4416	0.18
VCC	1125	0.093	162	0.007	18	0.001	929	0.04
CCCCV	14	0.001	21	0.001	381	0.018	364	0.02
CVCCC	579	0.048	3	0.000	1	0.000	148	0.01
CCCVCC	105	0.009	0	0.000	0	0.000	65	0.00
CCCCVC	1	0.000	0	0.000	25	0.001	58	0.00
VCCC	45	0.004	0	0.000	0	0.000	22	0.00
CCCCVC	11	0.001	0	0.000	0	0.000	25	0.00
Other	38	0.003	0	0.000	7	0.000	21	0.00

Based on *SrpLemKor*, the most frequent monosyllabic syllable structures in Serbian are CV (50%), V (25%) and CVC (11%). The most frequent syllable structures in the initial position of polysyllabic words are CV (56%), CCV (15%) and V (15%). In medial positions in polysyllabic words, the most frequent syllable structures are CV (69%), CCV (14%) and CVC (10%). The most frequent syllable structures in the final position of polysyllabic words are CV (65%), CVC (16%) and CCV (8%). It is interesting to note the asymmetry that the syllable structures CCCVCC, VCCC, and CCCCVC occurred only in monosyllabic words and in the final position of polysyllabic words, while the syllable structure CCCCVC occurred in all positions except the initial position in polysyllabic words.

Syllable Nuclei Statistics in Serbian

The distribution of different syllable nuclei in Serbian based on the *SrpLemKor* corpus is presented in Table 3.

Table 3: Syllable nuclei statistics and positional frequencies of syllables in the *SrpLemKor* corpus

Nucleus	TOTAL		Monosyllabic words		Polysyllabic words					
			MONO		INITIAL		MEDIAL		FINAL	
	No.of instances	Percent	No.of instances	Percent	No.of instances	Percent	No.of instances	Percent	No.of instances	Percent
a	2177498	26.566	330629	27.209	604764	24.990	585787	27.353	656318	27.120
e	1646579	20.088	304442	25.054	447662	18.498	394573	18.425	499902	20.657
i	1730439	21.111	230637	18.980	394735	16.311	600823	28.056	504244	20.836
l	939	0.011	326	0.027	32	0.001	77	0.004	504	0.021
n	1261	0.015	409	0.034	544	0.022	33	0.002	275	0.011
o	1753091	21.388	168126	13.836	671752	27.758	385687	18.010	527526	21.798
r	88021	1.074	1898	0.156	66250	2.738	19560	0.913	313	0.013
u	798943	9.747	178674	14.704	234301	9.682	155010	7.238	230958	9.544

Based on the positional nucleus distribution data, it can be seen that overall /a/ and /o/ constitute the most frequent nuclei in Serbian. However, there is some positional variation. While the most frequent nuclei in final, medial, and initial position of polysyllabic words are also /a/ and /o/, in monosyllabic words, the most frequent nuclei are /a/ and /e/.

Discussion

While our mixed-principle rule-based syllabification algorithm is suitable for the segmentation of words into syllables following the ruleset we devised based by the combination of prescriptive rule descriptions and the employment of the Sonority

Sequencing Principle, there are still some practical and theoretical considerations to be addressed.

While reporting on the syllable distribution data, we mentioned that the 3,648,543 word-forms extracted from *SrpLemKor* used for the calculation of statistical data related to the distribution of syllables and their structure in Serbian still contained some noise such as foreign words, typos, and possibly random character strings. Based on 500 random samples taken from the syllable output data checked by a human evaluator, the estimate of the amount of such noise in the data is <2%. Given the nature of corpus-based data, this noise should not significantly impact the reliability of the distributional information.

From a theoretical standpoint, in formulating our algorithm, we disregarded the three-syllable consonant cluster limitation put forward by Kašić (2014) in favor of exploring the limitations of the sonority module. The occurrence of the two syllable types CCCCVC and CCCCV, which were not present in the onset-maximization-based syllabification algorithm for Croatian (Meštrović et al. 2015), shows that in a limited number of instances this constraint is needed to exclude syllable clusters that are in accordance with the SSP and prescriptive rule descriptions, but seem contrary to native speaker intuition about syllable boundaries. In addition to this, there is the ambiguity in syllable segmentation in the case of continuant fricative phonemes (Gvozdanović 2011) with the continuant constituting either the first place in the onset of the syllable or the last place in the coda of the previous syllable, e.g. the possibility to syllabify *postaviti* [*to set*] as *po-sta-vi-ti* and *pos-ta-vi-ti*, would require a larger-scale study examining the intuition of native speakers on syllabification to make an assumption about contemporary tendencies in the segmentation in these contexts.

In order to verify the syllabic status of different clusters, it would be interesting to conduct a series of monitoring studies modeled after Mehler et al. (1981), who have shown that reaction times to a word are faster if the word is primed by a sequence corresponding to a syllable in the word when compared to priming with a string that does not constitute a syllable. Bradley et al. (1993) argue that these effects produce mixed results in some languages which contain a large number of ambisyllabic segments, so these studies may also reveal whether and to what extent syllables play a role in prelexical processing in Serbian.

Conclusion

In this paper we presented a mixed-principle rule-based syllabifier modelled after the rule descriptions found in Stanojčić and Popović (2005), extended by rule specifications from Kašić (2014) and Zec (2000), and complemented by a sonority sequencing module based on Selkirk (1984), Subotić et al. (2012), and Zec (2000).

An implementation of the existing prescriptive rules for the segmentation of words into syllables allowed us to gain an insight into the problem areas of the rule descriptions, and propose a number of revisions and amendments to the existing rules. The sonority sequencing module revealed the need for an additional onset-length limitation constraint, and pointed out the limitations of sonority in ambiguous consonant clusters that would require further exploration and validation by native speaker intuition. We have also gained an insight into the distribution of different syllable structures and syllable nuclei following this approach, which will be useful for comparison with the performance of alternative syllabification systems.

In the future, we plan to compare our system to an onset-maximization-based syllabifier for Serbian in combination with the prescriptive rules to see if we can create an alternative system that will produce outputs consistent with the intuition of native speakers of Serbian. It would be interesting to see a systematic comparison of our current approach and the onset-maximization approach with data gathered from the intuition of contemporary native speakers of Serbian.

We also believe that, while phonological criteria present a basis for syllabification, in the future we will also need to test whether and to what extent approaches based solely on phonological criteria result in syllable boundaries that coincide with morphological boundaries. Our assumption is that phonological rules will need to be amended by morphological criteria to result in syllabification that respects morphological boundaries as well.

In addition to these, the question of the treatment of foreign origin words and transcribed foreign words might be an additional point to consider. As an extension of a syllabifier, a language detection algorithm might be employed to properly segment the former, while the latter might not need special treatment as the process of transcription should in itself contain a degree of phonological adaptation.

Acknowledgment

This research was supported by the Serbian Ministry of Education and Science under the projects Development of Dialogue Systems for Serbian and Other South Slavic Languages (TR-32035) and Languages and Cultures in Time and Space (ON-178002).

Sources and Literature

Literature:

- Barber, Horacio, Marta Vergara, and Manuel Carreiras. 2004. “Syllable-frequency Effects in Visual Word Recognition: Evidence from ERPs.” *Neuroreport* 15 (3): 545–48.
- Bradley, Dianne C., Rosa M. Sánchez-Casas, and José E. García-Albea. 2007. “The Status of the Syllable in the Perception of Spanish and English.” *Language and Cognitive Processes* 8 (2): 197–233.

- Bigi, Brigitte, and Caterina Petrone. 2014. "A Generic Tool for the Automatic Syllabification of Italian." In *Proceedings of The First Italian Conference on Computational Linguistics, CLiC-it*, 73–77. Pisa: Pisa University Press. http://siti.fileli.unipi.it/projects/clic/proceedings/Proceedings-CLICit-2014.pdf.
- Butt, Matthias. 1992. "Sonority and the Explanation of Syllable Structure." *Linguistische Berichte* 137: 45–67.
- Cholin, Joana, Willem J. M. Levelt, and Niels O. Schiller. 2006. "Effects of Syllable Frequency in Speech Production." *Cognition* 99 (2): 205–35.
- Cholin Joana, and Willem J. M. Levelt. 2009. "Effects of Syllable Preparation and Syllable Frequency in Speech Production: Further Evidence for Syllabic Units at a Post-lexical Level." *Language and Cognitive Processes* 24(5): 662–84.
- Clements, George N. 1990. "The Role of the Sonority Cycle in Core Syllabification." In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by John Kingston, John and Mary E. Beckman, 282–333. Cambridge: Cambridge University Press.
- Daelemans, Walter, and Antal van den Bosch. 1992. "Generalization Performance of Backpropagation Learning on a Syllabification Task." In *Connectionism and Natural Language Processing: Proceedings of the 3rd Twente Workshop on Language Technology, TWLT3*, 27–38. Enschede: University of Twente, Department of Computer Science. https://pure.uvt.nl/portal/files/760578/generalization.pdf.
- Foley, James. 1972. "Rule Precursors and Phonological Change by Meta-rule." In *Linguistic change and generative theory*, edited by Robert P. Stockwell and Ronald K. S. Macaulay, 96–100. Bloomington: Indiana University Press.
- Goldsmith, John A. 1995. *The Handbook of Phonological Theory*. London: Blackwell Publishers.
- Gvozdanović, Jadranka. 2011. "Phonological Domains." In *Sandhi Phenomena in the Languages of Europe,* edited by Henning Andersen, 27–54. Berlin: Mouton de Gruyter.
- Hankamer, Jorge, and Judith Aissen. 1974. "The Sonority Hierarchy." In *Papers from the Parasession on Natural Phonology*, edited by Anthony Bruck, Robert Allen Fox, and Michael W. La Galy, 131–45. Chicago: Chicago Linguistic Society.
- Hunt, Andrew. 1993. "Recurrent Neural Networks for Syllabification." *Speech Communication* 13 (3–4): 323–32.
- Iacoponi, Luca, and Renata Savy. 2011. "Sylli: Automatic Phonological Syllabification for Italian." In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, 641–44. Florence: International Speech Communication Association. http://eden.rutgers.edu/~li51/php/papers/interspeech2011.pdf.
- Kaplar, Sebastijan, Marija Radojičić, Ivan Obradović, Biljana Lazić, and Ranka Stanković. 2018. "Solution for Quantitative Analysis of Texts in Serbian Based on Syllables." In *ICIST 2018 Proceedings* 2, 315–20. Belgrade: Society for Information Systems and Computer Networks. http://www.eventiotic.com/eventiotic/library/paper/429.
- Kašić, Zorka. 2014. "Opšta lingvistika 2 (Fonologija)." Lecture Materials, Faculty of Philosophy, University of Belgrade.
- Koehler, Klaus J. 1966. "Is the Syllable a Phonological Universal?" *Journal of Linguistics* 2: 207–208.
- Kovač, Aniko, and Maja Marković. 2018. "A Rule-Based Syllabifier for Serbian." In *Proceedings of the Conference on Language Technologies and Digital Humanities 2018*, 140–46. Ljubljana: Ljubljana University Press.
- Ladefoged, Peter, and Keith Johnson. 2014. *A Course in Phonetics*. Belmont: Wadsworth Publishing.
- Ladefoged, Peter. 1982. *A Course in Phonetics*. New York: Harcourt Brace Jovanovich.
- Landsiedel, Christian, Jens Edlund, Florian Eyben, Daniel Neiberg, and Björn Schuller. 2011. "Syllabification of Conversational Speech Using Bidirectional Long-Short-Term Memory Neural Networks." In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5256–9. Prague: IEEE. http://ieeexplore.ieee.org/abstract/document/5947543.

- Marchand, Yannick, Connie R. Adsett, and Robert I. Damper. 2009. "Automatic Syllabification in English: A Comparison of Different Algorithms." *Language and Speech* 52 (1): 1–27.
- Mehler, Jacques, Jean Yves Dommergues, Uli Frauenfelder, and Juan Segui. 1981. "The Syllable's Role in Speech Segmentation." *Journal of Verbal Learning and Verbal Behavior* 20 (3): 298–305.
- Meštrović, Ana, Sanda Martinčić-Ipšić, and Mihaela Matešić. 2015. "Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik." *Govor*, 32: 3–34.
- Morelli, Frida. 1999. "The Phonotactics and Phonology of Obstruent Clusters in Optimality Theory." PhD diss., University of Maryland.
- Ohala, John, and Haruko Kawasaki. 1984. "Prosodic Phonology and Phonetics." *Phonology Yearbook*, 1: 113–27.
- Ohala, John. 1990. "The Phonetics and Phonology of Aspects of Assimilation." In *Papers in Laboratory Phonology I*, edited by John Kingston, John and Mary E. Beckman, 258–75. Cambridge: Cambridge University Press.
- Popović, Zoran. 2010. "Taggers Applied on Texts in Serbian." *INFOtheca* 11 (2): 21a–38a.
- Selkirk, Elisabeth O. 1984. "On the Major Class Features and Syllable Theory." In *Language Sound Structure*, edited by Mark Aronoff and Richard T. Oehrle, 107–36. Cambridge: MIT Press.
- Stanojčić, Živojin, and Ljubomir Popović. 2005. *Gramatika srpskoga jezika*. Belgrade: Zavod za udžbenike i nastavna sredstva Beograd.
- Stoianov, Ivelin, John Nerbonne, and Huub Bouma. 1997. "Modelling the Phonotactic Structure of Natural Language Words with Simple Recurrent Networks." In *Computational Linguistics in the Netherlands 1997: Selected Papers from the Eight Clin Meeting*, 77–95. Amsterdam: Rodopi.
- Subotić, Ljiljana, Dejan Sredojević, and Isidora Bjelaković. 2012. *Fonetika i fonologija: Ortoepska i ortografska norma standardnog srpskog jezika*. Novi Sad: Filozofski fakultet Univerziteta u Novom Sadu.
- Utvić, Miloš. 2011. "Annotating the Corpus of Contemporary Serbian." *INFOtheca* 12 (2): 36a–37a.
- Zec, Draga. 2000. "O strukturi sloga u srpskom jeziku." *Južnoslovenski filolog* 56 (1–2): 435–48.

Aniko Kovač, Maja Marković

A MIXED-PRINCIPLE RULE-BASED APPROACH TO THE AUTOMATIC SYLLABIFICATION OF SERBIAN

SUMMARY

In this paper we present a mixed-principle rule-based approach to the automatic syllabification of Serbian based on prescriptive rule descriptions from traditional grammar found in Stanojčić and Popović (2005), extended by rule specifications from Kašić (2014) and Zec (2000), and complemented by a sonority sequencing module based on Selkirk (1984), Subotić et al. (2012), and Zec (2000).

Syllable segmentation plays a role in speech technologies – most notably in the areas of speech recognition and text-to-speech synthesis – at both the segmental and prosodic levels. It is also one of the governing factors in hyphenation, and syllable frequency distribution data is used in psycholinguistic experiments as a covariate. The unavailability of segmented data for Serbian makes a rule-based approach to automatic syllabification the only viable option as there is no data available for training a data-driven neural network model, and the segmentation of large-scale language corpora by trained annotators would be a resource and cost heavy undertaking.

Our goal in this paper is threefold: i) we extend and improve an earlier version of our syllabification algorithm by introducing a sonority sequencing validation module which resolves a number of issues present in the earlier version of our syllabifier, ii) we provide a detailed analysis of the outcomes of the automatic syllabification process in order to address possible theoretical considerations and serve as a basis for the development of future syllabifiers, and iii) we present the statistical data related to the distribution of syllables and their structure in Serbian to be used in psycholinguistic experiments.

The implementation of the existing set of prescriptive rules for the segmentation of words into syllables in Serbian allowed us to gain an insight into problem areas of the rule descriptions, and propose a number of revisions and amendments to the existing rules. The sonority sequencing module revealed the need for an additional onset-length limitation constraint, and pointed out the limitations of sonority in ambiguous consonant clusters – such is the case with continuant fricative phonemes that seem to be able to occupy either the first place in the onset of a syllable or the last place in the coda of a previous syllable – that would require further exploration and validation by native speaker intuition.

The data on the distribution of different syllable structures and syllable nuclei following this approach will be useful for comparison with the performance of alternative syllabification systems. In the future, it would be interesting to see a systematic comparison of our current approach to alternative approaches such as an onset-maximization approach evaluated on segmentation data gathered from the native speakers of Serbian.

Aniko Kovač, Maja Marković

MEŠANI PRISTOP K AVTOMATSKEMU ZLOGOVANJU V SRBŠČINI NA PODLAGI NAČEL IN PRAVIL

POVZETEK

V tem prispevku predstavljamo mešani pristop k avtomatskemu zlogovanju v srbščini na podlagi načel in pravil, ki temelji na opisih predpisnih pravil tradicionalne slovnice (kot jih navajata Stanojčić in Popović 2005), razširjenih z opredelitvami pravil (kot jih navajata Kašić (2014) in Zec (2000)) in dopolnjenih z modulom za zaporedje glede na zvočnost (na podlagi del avtorjev Selkirk 1984; Subotić et al. 2012; Zec 2000).

Členitev na zloge ima pomembno vlogo v govornih tehnologijah – zlasti na področjih prepoznavanja govora in pretvorbe besedila v govor – na segmentalni in prozodični ravni. Je tudi eden od vodilnih dejavnikov pri deljenju besed. Podatki o frekvenčni porazdelitvi zlogov se uporabljajo v psiholingvističnih poskusih kot sočasna spremenljivka. Pristop k avtomatskemu zlogovanju, ki temelji na pravilih, je edina smiselna izbira, saj za srbščino ni na voljo segmentiranih podatkov, iz katerih bi se model nevronske mreže lahko učil. Projekt, pri katerem bi usposobljeni komentatorji razčlenjevali obsežne jezikovne korupse, pa bi bil zelo zahteven in drag.

Naš prispevek ima tri cilje: i) razširiti in izboljšati predhodno različico našega algoritma za zlogovanje z vpeljavo modula za potrjevanje zaporedja glede na zvočnost, ki odpravlja vrsto težav iz predhodne različice našega zlogovalnika; ii) predstaviti podrobno analizo rezultatov avtomatskega postopka zlogovanja, da bi spodbudili morebitne teoretične razmisleke in zagotovili podlago za razvoj prihodnjih zlogovalnikov; in iii) predstaviti statistične podatke, povezane s porazdelitvijo in strukturo zlogov v srbščini, ki jih bo mogoče uporabiti pri psiholingivstičnih poskusih.

Uporaba uveljavljene zbirke predpisnih pravil za členitev besed na zloge v srbščini nam je omogočila, da smo dobili podroben vpogled v težavna področja pri opisih pravil in predlagali vrsto sprememb in popravkov uveljavljenih pravil. Modul za zaporedje glede na zvočnost je razkril potrebo po dodatni omejitvi dolžine vzglasja in izpostavil omejitve zvočnosti pri dvoumnih soglasniških sklopih (na primer priporniki, ki očitno lahko zavzemajo prvo mesto na začetku zloga ali zadnje mesto na koncu predhodnega zloga), ki bi jih bilo treba dodatno raziskati in potrditi s pomočjo intuicije rojenega govorca.

Podatke o porazdelitvi različnih zlogovnih struktur in jeder, pridobljene s tem pristopom, bo mogoče uporabiti za primerjavo z delovanjem drugih sistemov za zlogovanje. Zanimivo bi bilo opraviti sistematično primerjavo našega pristopa z drugimi pristopi, na primer pristopom maksimizacije vzglasja, ovrednotenim na podlagi podatkov o členitvi, pridobljenih od rojenih govorcev srbščine.