A Rule-Based Syllabifier for Serbian

Aniko Kovač,* Maja Marković†

* 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

Abstract

In this paper, we present an automatic rule-based syllabification algorithm for Serbian based on prescriptive rules from traditional grammar. We explore the problems and limitations of the existing rule set and present the statistical data related to the distribution of syllables and their structure in Serbian.

1. Introduction

Syllables have been considered — although not unequivocally (cf. Koehler, 1996) — 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 language technology in the 1990s (Iacoponi and Savy, 2011), most notably in the areas of speech recognition (SR) and text-to-speech synthesis (TTS).

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.

2. The goal of the paper

In this paper, we present a rule-based automatic syllabifier for Serbian. We based our starting set of rules on *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. However, as the formulation of some of these descriptions proved to be redundant, we devised an algorithm for syllabification aimed to produce an output consistent with the rules prescribed in *Gramatika srpskoga jezika*, rather than a verbatim implementation of the formalized rules, 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).

The goal of the paper is threefold: i) to develop a system for automatic rule-based syllabification for Serbian based on the formalization of existing rule descriptions, 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.

3. The descriptive rule set

Stanojčić and Popović (2005) establish syllables as speech units of the language which can be produced with a single articulatory movement. While there is no consensus on a universal definition of the syllable or what principles should govern the segmentation of words into syllables, there is general agreement that each syllable consists of a syllable-carrying element called *nucleus* which can be preceded by zero or more consonants constituting the *onset* and followed by zero or more consonants making up the *coda*.

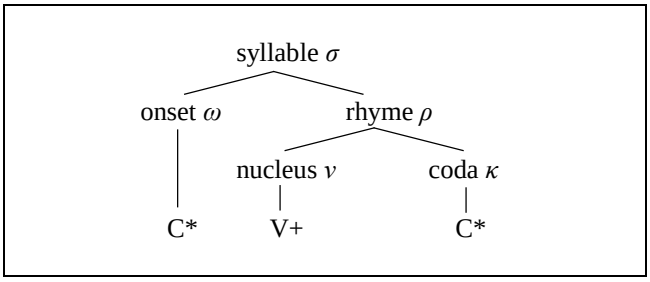

Figure 1: Tree diagram of syllable structure

In accordance with this, Stanojčić and Popović state that syllables in Serbian can be made up of a single phoneme, provided that that phoneme is a vowel. In syllables consisting of multiple phonemes — the nucleus in combination with consonants in the onset and/or coda — the sonorants /r/, /l/ and /n/ can also act as syllable carrying nuclei in Serbian.

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* lj *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, lj *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*]).

The initial member of a consonant cluster in the rule descriptions presented above is understood as the first consonant following a vowel based on the general rule presented under (1). However, a more precise definition would be that the initial member of a consonant cluster is the first consonant following a syllable nucleus — which in the case of Serbian also includes the sonorants /*r*/, /*l*/ and /*n*/ in certain positions. 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*]).

Stanojčić and Popović (2005: 32) 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*].

3.1. A note on modifications of the original rule set

In addition to our expansion of the general rule presented under (1) to include the syllable bearing sonorants /*r*/, /*l*/ and /*n*/ (1*), the rule descriptions in Stanojčić and Popović (2005) needed to be further modified in the following cases.

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 (1). Because of this, these rules were disregarded in our syllabification algorithm.

During our early testing of the verbatim implementation of the rule descriptions of Stanojčić and Popović (2005), we 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*/, /*lj*/ 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. However, other authors (e.g. 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, lj *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. As this approach also respects the limitations put forward by the Sonority Hierarchy — even though this version of our syllabifier is not based on the Sonority Sequencing Principle (SSP) — we follow the treatment of nasals by Kašić (2014) in our implementation.

3.1.1. Alveolar sonorant nuclei

One of the most problematic areas of the rules put forward by Stanojčić and Popović (2005) was their treatment of syllable bearing alveolar sonorants under (7) and (8).

We decided against the treatment of /*r*/ as a syllable nucleus following a vowel in compounds as specified in rule description (7c) as taking morpheme boundaries into consideration would not be a phonological, but rather a morphological criterion of syllabification. We also decided to treat the alveolar sonorant /r/ as non-syllabic before the vowel /*o*/ that is realized as /*l*/ in some members of the paradigm (7d) following Kašić (2014) who states that /*r*/ is no longer systematically treated as a separate syllable in these instances, and that it is pronounced as non-syllabic in words such as *umro [died]*, *groce [throat]* and *otro*

[wiped]. This means that these words should no longer be segmented as *um-r-o*, *gr-o-ce* and *ot-r-o* as suggested in Stanojčić and Popović (2005), but rather as *um-ro*, *gro-ce* and *ot-ro*.

We have also amended rule (7) for syllable bearing /r/, by further specifying it to exclude /r/ followed by the sequence *je* from being treated as a syllable nucleus as this would be in violation of the rule description under (6) which puts the syllable boundary before a sonorant cluster in words from the ijekavski dialect thus keeping the consonant cluster together.

In order to formalize the rule description under (8) of Stanojčić and Popović (2005) which gave no formal criteria defining when /l/ and /n/ were syllable carriers, we drew on generalizations based on their examples for syllable bearing /l/ (*Stlp*, *Vlča glava*, *Žlne*, *Vlava*, *Plzen*) and /n/ (*Idn*, *Ibn-Saud*) and implemented rule (8*) in analogy to the rules defined for the syllable carrying alveolar /r/.

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

However, this resulted in outcomes such as: *Be-rn*, *Ka-rl*, *erla-jn*, *Kla-jn*, *kasa-rn-skim*, *Linko-ln*, *Va-jl-dom* etc. In these examples, the sonority of /l/ and /n/ identified as syllable nuclei is lower than the sonority of a consonant in their immediate context — /r/ and /j/ are more sonorous than /n/ and /l/, and /l/ is more sonorous than /n/. Because of this, native speakers do not perceive as there being a syllable constituted around /l/ and /n/ in these contexts. [1] According to Zec (2000), alveolar sonorants can be syllable carriers 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 need to further specify rule (8*) as follows.

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

Interestingly, this principle applied to the syllable bearing /r/ could also account for our extension of rule (7) keeping the consonant cluster of the ijekavica dialect unsegmented in initial position — because /j/ is more sonorous than /r/, and then /r/ should not be treated as a syllable nucleus initially in words such as *rjeka [river]*. However, our rule extension has a more general scope than the sonority rule as it also accounts for medial clusters (e.g. in *isko-rje-nilo [eradicated]*).

4. Our algorithm[2]

Our syllabification algorithm consists of the following steps:

i. Identify vowels in the word and mark their positions as positions capable of constituting syllable nuclei.

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.

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, 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.
   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.
   c. In all other cases mark the syllable boundary after the syllable nucleus.

5. Results

In this section, we present the statistical distribution data for syllables in Serbian based on our 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 3,607,450 word-forms in *SrpLemKor*. From a total of 4,681,713 entities 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 947,666 (20.24%) entities tagged *PUNCT* (punctuation), *SENT* (sentence separator full-stops), *RN* (Roman numerals), *NUM* (numbers), *ABB* (abbreviations) and *?* (non-Serbian words and other uncategorized entries). An additional 551 (0.01%) entities that contained the characters *w* and *q* 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 12,910 (0.28%) entities were removed as they were solely made up of consonant clusters with no available syllable nucleus candidate.

5.1. Syllable type distributions in Serbian

In the 3,607,450 word-forms from *SrpLemKor*, a total of 8,147,679 syllables were identified. Table 1 presents the

[1] We thank Miloš Košprdić for his insight and helpful discussion on this topic.

[2] 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.

syllable type distribution based on our syllabification algorithm.

Syllable structure	No. of instances	Percent
CV	5034567	61.791
CCV	1009791	12.394
V	863631	10.6
CVC	771143	9.465
CCVC	215267	2.642
VC	131021	1.608
CCCV	69577	0.854
CCCVC	21151	0.26
CVCC	17210	0.211
CCVCC	6487	0.08
CCCCV	4292	0.053
VCC	1526	0.019
CCCCVC	708	0.009
CVCCC	705	0.009
CCCVCC	391	0.005
VCCC	66	0.001
CCVCCC	32	0
CCCCVCC	23	0
CCCVCCC	14	0
CCCCCV	3	0
CCCCCVC	2	0
Other	73	0.001
Total	8147679	100

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

These results show the distribution of syllables in a somewhat noisy data. We found that 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, we found one instance of the structure CCCCVCCC from the German word *Fleischmarkt* [*meat market*], one example of the structure CCCCCCVC from the German *Nachtschatten* [*nightshade*], a single entry CCCCCCCV from the German word *Storchschnabel* [*Crane's-bill*], one instance of the structure CCCCCCVCC from the English *healthystuff*, 4 examples of the structure VCCCCC from two occurrences of the German words *Peitscht* [*lashes*], one instance of *staruch* (typo or possibly Polish [*old man*]) and one instance of the English word *knights*. We also found 10 instances of the structure VCCCC from the German *Ernst* [*seriousness*], *Deutsch* [*German*], and strings such as *ikvbv*, which we assume stand for unfiltered acronyms, and strings we could not associate with any meaning such as *ehmc* and *rhutm*. We have also identified one example of the sequence CVCCCCCCCC to stand for the onomatopoeic vulgarism *mršššššššš* [*go away*].

Besides these, we found 6 types of syllable structure that differed from the structures found by Meštrović et al. (2015) for Croatian. The structures CCCCCVC (e.g. *mo-na-rhstvom [with the monarchy])*, CCCCV (e.g. *se-rbska [Serbian]*, *ca-rstva [kingdoms]*, *sta-ra-te-ljstva [custody]*) and CCCCVC (e.g. *se-rbskom [Serbian]*, *de-jstvom [with effect]*, *vo-đstvom [leadership]*, *spo-rtskim [sport]*, *a-lpskog [alpine]*) represented Serbian entities and are in accordance with the syllabification rules, but present some theoretical issues which we discuss in section 6. In the case of the structure CCCCCV, we separated the counts to include *se-rbstvo [Serbian]* as a problematic but valid entry, but exclude counts resulting from typos (e.g. *ri-va-ststva*, *su-žnjstva*, *šttske*) and foreign words (e.g. *ba-ckstre-et*) which were counted as "Other". The structure CCCCVCC found in foreign origin names (e.g. *Go-ldštajn, Rot-hchild, Ar-mstrong*), and the structure CVCCCC, a result of typos (e.g. *slav-janskh, cr-no-gorskg*), were also counted under "Other" in Table 1.

5.2. 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.

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	612244	50.784	1398930	58.244	1486143	69.499	1537250	64.002
CCV	54417	4.514	376527	15.676	351099	16.419	227748	9.482
V	301295	24.991	379122	15.785	62176	2.908	121038	5.039
CVC	128321	10.644	121162	5.045	155947	7.293	365713	15.226
CCVC	35434	2.939	44923	1.87	47315	2.213	87595	3.647
VC	64037	5.312	57451	2.392	6210	0.29	3323	0.138
CCCV	177	0.015	20012	0.833	24708	1.155	24680	1.028
CCCVC	1490	0.124	3715	0.155	3950	0.185	11996	0.499
CVCC	4666	0.387	0	0	0	0	12544	0.522
CCVCC	1638	0.136	0	0	0	0	4849	0.202
CCCCV	9	0.001	19	0.001	750	0.035	3514	0.146

VCC	1100	0.091	0	0	0	0	426	0.018
CCCCVC	4	0	0	0	46	0.002	658	0.027
CVCCC	568	0.047	0	0	0	0	137	0.006
CCCVCC	104	0.009	0	0	0	0	287	0.012
VCCC	42	0.003	0	0	0	0	24	0.001
CCVCCC	12	0.001	0	0	0	0	20	0.001
CCCCVCC	1	0	0	0	0	0	22	0.001
CCCVCCC	1	0	0	0	0	0	13	0.001
CCCCCV	0	0	0	0	0	0	3	0
CCCCCVC	0	0	0	0	0	0	2	0
Other	36	0.003	1	0	16	0.001	20	0.001

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

Based on *SrpLemKor*, the most frequent monosyllabic syllable structures in Serbian are CV (51%), V (24%) and CVC (11%). The most frequent syllable structures in the initial position of polysyllabic words are CV (58%), V (16%) and CCV (16%). In medial positions in polysyllabic words, the most frequent syllable structures are CV (70%), V (16%) and CVC (7%). The most frequent syllable structures in the final position of polysyllabic words are CV (64%), CVC (15%) and CCV (10%).

It is interesting to note the asymmetry that the syllable structures CVCC, CCVCC, VCC, CVCCC, CCCVCC, VCCC, CCVCCC, CCCCVCC and CCCVCCC 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. The rare (and problematic) structures CCCCCV, CCCCCVC occurred only in the final positions of polysyllabic words.

5.3. Syllable nuclei statistics in Serbian

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

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	2166178	26.586	327721	27.183	604299	25.160	585064	27.360	649094	27.025
o	1747318	21.446	167750	13.914	671083	27.940	385403	18.023	523082	21.778
i	1725046	21.172	228055	18.916	394426	16.422	599859	28.052	502706	20.930
e	1620813	19.893	300701	24.942	430654	17.930	393488	18.401	495970	20.649
u	797667	9.790	178664	14.820	234319	9.756	155017	7.249	229667	9.562
r	88233	1.083	1966	0.163	66435	2.766	19383	0.906	449	0.019
n	1411	0.017	411	0.034	602	0.025	50	0.002	348	0.014
l	1014	0.012	328	0.027	44	0.002	96	0.004	546	0.023

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

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 position are also /*a*/ and /*o*/, and /*o*/ and /*a*/ represent the most frequent nuclei in the initial position of polysyllabic words, in monosyllabic words, the most frequent nuclei are /*a*/ and /*e*/, while in the medial positions in polysyllabic words, the most frequent nuclei are /*i*/ and /*a*/.

6. Discussion

In the previous section, we mentioned that the 3,607,450 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, acronyms, 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%.

While our syllabifier is suitable for the segmentation of words into syllables following the set of provided rule descriptions, we argue that the prescriptive rules themselves need revising as they seem to violate basic phonetic and phonotactic principles of the language.

In their automatic syllabification system for Croatian based on the Onset Maximization Principle, Meštrović et

al. (2015) limit possible onsets in medial and final clusters to those onsets which occur in word-initial positions with some extensions to the allowed onsets following the principle of analogy by place of articulation and taking into account voiced and voiceless consonant pairs. While we remain uncertain whether initial occurrences should be used as a criterion for medial and final onsets, we are interested in exploring the possibility of an onset maximization segmentation based on Meštrović et al. (2015), but limited by the prescriptive rules used in the syllabifier presented in the paper. For example, this would mean that some questionable onsets such as /*pn*/ in *va-pno* [*lime*], which they allow for because /pn/ constitutes a valid onset in the word *pneumatski* [*pneumatic*], would be disallowed and segmented as *vap-no* in such a system because of rule (5) that defines a syllable boundary between a plosive and subsequent consonant that is not one of the sonorants /j/, /v/, /l/, /lj/ and /r/.

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 pre-lexical processing in Serbian.

One of the main problems that we have identified with a syllabifier based on the set of prescriptive rules presented in section 3 is that even with the revised rule set, the results are often problematic when taking into account the viewpoint that the structure of syllables should be in accordance with the Sonority Sequencing Principle. Namely, if we assumed that syllables are structured in such a way that there is a rising sonority of elements in the onset leading up to the nucleus, examples such as some of the problematic cases presented in section 5 (e.g. *se-rbska [Serbian]*, *de-jstvom [with effect]*) clearly violate the Sonority Hierarchy as alveolar sonorants have a higher sonority level than plosives and fricatives.

One way in which we attempted to remedy this was to introduce a limit of onset length to three-syllable clusters, which is the maximum length of non-syllabic consonant clusters word initially in Serbian (Kašić, 2014). While this — in combination with rules (5) and (6) — would indeed resolve the issues in the examples we encountered — they would be segmented as *serb-ska* and *dej-stvom* — medial clusters with a syllabic consonant would still present a problem. For example, the word *najstrpljiviji* [*most patient*], which contains a syllabic /r/ at the beginning of the hypothesized three-syllable maximum onset, would result in a boundary at *najst-rpljiviji* which is incorrect when taking into account the syllabic status of /r/. It would be interesting to see whether an added rule to separate elements with sonority violations might amend the existing rule set and resolve these problems, and compare the results stemming from this rule to the results of a rule limiting the range of possible onsets.

7. Conclusion

In this paper, we presented a rule-based syllabifier modelled after the rule descriptions found in Stanojčić and Popović (2005) and extended by rule specifications from Kašić (2014) 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. 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 improve our system by developing an onset-maximization-based syllabifier as well as a sonority-based syllabifier for Serbian, and then test a combination of these with the prescriptive rules to see if we can create a hybrid system that will produce outputs consistent with the intuition of native speakers of Serbian.

We also believe that, while phonological criteria present a basis for syllabification, in the future we might also need to test whether subsequent approaches coincide with morphological boundaries, or whether the phonological rules need to be amended to respect morphological boundaries as well.

In addition to these issues, 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.

8. References

Dianne C. Bradley, 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.

Andrew Hunt. 1993. Recurrent Neural Networks for Syllabifiation. *Speech Communication* 13(3–4):323–332.

Walter Daelemans and Antal van den Bosch. 1992. Generalization performance of backpropagation learning on a syllabification task. In: *Connectionism and natural language processing: Proceedings of the third Twente Workshop on Language Technology, TWLT3*, pages 27–38, Enschede, the Netherlands. https://pure.uvt.nl/portal/files/760578/generalization.pdf

Luca Iacoponi and Renata Savy. 2011. Sylli: Automatic Phonological Syllabification for Italian. In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, pages 641–644, Florence, Italy. http://eden.rutgers.edu/~li51/php/papers/interspeech2011.pdf.

Zorka Kašić. 2014. *Opšta lingvistika 2 (Fonologija)*. Lecture Materials, Faculty of Philosophy, University of Belgrade.

Kenneth J. Koehler. 1996. Is the syllable a phonological universal? *Journal of Linguistics*, 2:207–208.

Peter Ladefoged and Keith Johnson. 2014. *A Course in Phonetics*. Wadsworth Publishing.

Christian Landsiedel, 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)*, pages 5256–5259, Prague, Czech Republic. http://ieeexplore.ieee.org/abstract/document/5947543.

Yannick Marchand, Connie R. Adsett, and Robert I. Damper. 2009. Automatic syllabification in English: A comparison of different algorithms. *Laguage and Speech* 52(1):1–27.

Jacques Mehler, Jean Yves Dommergues, and Uli Frauenfelder, Juan Segui. 1981. The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 20(3): 298–305.

Ana Meštrović, Sanda Martinčić-Ipšić, Mihaela Matešić. 2015. Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik. *Govor*, 32:3–34.

Zoran Popović. 2010. Taggers Applied on Texts IN Serbian. *INFOtheca* 11(2):21a–38a.

Živojin Stanojčić and Ljubomir Popović. 2005. *Gramatika srpskoga jezika*. Zavod za udžbenike i nastavna sredstva Beograd.

Ivelin Stoianov, 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*, pages 77–95.

Miloš Utvić. 2011. Annotating the Corpus of Contemporary Serbian. *INFOtheca*, 12(2):36a–37a.

Draga Zec. 2000. O strukturi sloga u srpskom jeziku. *Južnoslovenski filolog*, 56(1-2):435–448.