Sonority Based Syllabic Hyphenation of Macedonian and Serbian

Scientific paper DOI: 10.18485/judig.2025.1.ch5

Katerina Zdravkova, *0000-0002-9674-3081*
Jana Kuzmanova[1] *0009-0009-5015-4183*

Abstract

Hyphenation is a technique of joining or separating syllables of a word by adding a hyphen. This is typically achieved by implementing the syllabic segmentation. We have attempted to achieve the same task based on the sonority of phonemes, considering that sonority within a syllable increases to the syllable nucleus and decreases at syllable boundaries. For Macedonian, this principle was extended by an additional rule stating that a syllable begins when a monotonically non-decreasing sonority ends. In Serbian, monotonic non-decreasing starts at the beginning of a new syllable. By defining appropriate sonority weights for both languages and defining a very simple splitting strategy, the accuracy of the syllabic hyphenation has surpassed the rule-based approach. It reached 97.59% for Macedonian and 98.68% for Serbian. We intend to further improve it by taking into account PoS tags for Macedonian and to fine-tune the sonority weights for Serbian, hoping to achieve an accuracy that exceeds 99%.

Keywords: Macedonian, Serbian, syllabic hyphenation, phoneme sonority, ChatGPT.

1. Introduction

Hyphenation is a technique of joining or separating the syllables of words that improves their legibility, readability, aesthetics and visual balance of printed or displayed texts [1]. The separation is done at the end

[1] Ss Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, Skopje, {katerina.zdravkova; jana.kuzmanova}@finki.ukim.mk

of a line, using a hyphen to indicate where the break occurs. Since this process combines the concepts of both syllabic and hyphenation, the term for this process can be generalized as syllabic hyphenation [2]. It follows the natural syllable boundaries within a word in a way that respects their linguistic structure [3]. In the languages with a highly predictable syllabic structure, hyphenation provides the most natural and linguistically logical way to break words. In the languages with more complicated compound structures or where typography is heavily involved, non-syllabic hyphenation might be used in specific contexts [4].

The hyphenation of Macedonian language was embedded into Korektor-M, the first software package for spell checking and hyphenation of Macedonian texts. Created by Dragan Mihajlov and Dejan Gjorgjevikj more than 25 years ago, Korektor-M was a favourite tool of many newspapers that used it for automatic proofreading. With the development of Microsoft Office, it became useless. However, although new office suites offer spell checking option, word hyphenation of Macedonian language is not yet enabled.

The first attempt to create a new system for Macedonian hyphenation was done in 2023 [5]. It consisted of two mutually separate parts supporting syllabification and morpheme segmentation [5]. The segmentation was entirely based on rules that are quite general and ambiguous (https://makedonskijazik.mk/). The accuracy of the approach was below our expectations. Therefore, an alternative approach based on phoneme sonority was proposed [6], which was presented on JuDig conference (https://judig.jerteh.rs/). It is explained in more detail in this paper.

The hyphenation of Serbian language has a much longer tradition. It started with the research done by Krstev [7] and was then extended as part of word length counting [8]. The resources used for measuring the length of syllables in Serbian [9] were the bases for evaluating the proposed hyphenation approach. Similarly to Macedonian, the rule based approach was rather complex [10]. Exceeding the 98% accuracy obtained during this study became our main challenge.

2. Sonority Based Syllabic Hyphenation

Phonetically, syllables are sequences of sounds containing one peak of prominence [5]. Phonologically they are units of stress placement. According to the Sound Sequencing Principle, sonority within a syllable rises to the nucleus of the syllable and then falls in sonority [11].

The sonority of Macedonian phonemes depends on their basic classification: vowels, sonorants, voiced and voiceless consonants (Table 1). The sonority of Serbian phonemes is more extensive and includes three additional classes: plosives, fricatives and affricates, in each of which phonemes can be voiced or voiceless (Table 2).

Table 1: Sonority of Macedonian phonemes

Phoneme class	**Phonemes**	**Sonority weight**
Vowels	*а е и о у*	12
Sonorant *r*	*р*	6
Sonorants	*ј л љ м н њ*	4
Voiced consonants	*б в г д ѓ ж з ѕ џ*	2
Voiceless consonants	*к п с т ќ х ц ч ш*	1

Table 2: Sonority weights of Serbian phonemes

Phoneme class	**Phonemes**	**Sonority weight**
Vowels	*а е и о у*	12
Sonorant *r*	*р*	8
Sonorants *l*, *m* and *n*	*л м н*	6
Other sonorants	*в ј љ њ*	5
Plosive voiced	*б г д*	4
Plosive voiceless	*к п т*	3
Fricative voiced	*з ж*	3
Fricative voiceless	*с х ф ш*	2
Affricates voiced	*ђ ћ џ*	2
Affricates voiceless	*ц ч*	1
Special signs	FC S	0

The syllable nuclei in both languages are the five vowels. Their sonority weight is set to 12. In Macedonian, a nucleus can be the sonorant *р* (Latin transcription: *r*) appearing within a consonant group (*крст*, *вр-ста*, *пр-вен-ство*) or at the end of the word (*ма-са-кр*). In Serbian language, apart from the sonorant *р* (*тврд*, *црв*, *тр-ка*), the sonorants *л* and *н* can also become syllable nuclei (for example, *би-ци-кл, де-ба-кл*, *Вл-та-ва*, *Њу-тн*).

Macedonian language has a special sound ‘, which appear in words such as: *‘рж*, *‘рѓа*, and *‘рбет*, mainly at the beginning of the word and succeeded by the sonorant *р*.

Vowel groups (for example, *ау* in *а-у-ро-ра*, *еа* in *и-де-ал*, and *ио* in *а-ви-он*) in both languages are separated by a fictive consonant FC with a sonority weight of 0. Word delimiters are assigned as S, with the sonority weight of 0.

The syllabic roles of the phonemes are determined by calculating the triplet difference:

$$\mathrm{TD}(p_i) = w(p_i) - w(p_{i-1}) - w(p_{i+1}),\ i = 1,\ldots,n \qquad (1)$$

The triplet difference of the nuclei has always a positive value, while the consonants from the onset and the coda are always negative (Table 3 and Table 4). The only exclusion to this rule in Macedonian are the words starting with a voiced consonant succeeded by a voiceless consonant, such as *вчера* and *вчудовиден*. The problem was resolved by adding a restriction that a single voiced consonant cannot form a separate syllable.

In Serbian language, the high sonority weight of the sonorant *р*, when it is between two vowels, reduces the triplet difference of the vowel that immediately follows the fictive consonant, as in *а-у-ро-ра*, making it negative. This relatively rare inconsistency was solved in the second iteration of the syllabic hyphenation of Serbian, in which the weight of the non-syllabic *р* was reduced to 6.

The determination of syllable boundaries depends on a monotonicity of sonority weights. In Macedonian, a new syllable starts when monotonic non-decreasing ends. According to this constraint, whenever the sonority of two adjacent Macedonian consonants is identical, in that case the second one belongs to a new syllable (Table 3). In Serbian, a new syllable begins at a phoneme where the sonority weight series begins to monotonically increase (Table 4).

Table 3: Syllabic hyphenation of the word идеално in Macedonian

	S	И	Д	Е	FC	А	Л	Н	О	S
phonemme sonority	0	12	2	12	0	12	4	4	12	0
triplet difference		10	-22	10	-24	8	-12	-12	8	

Table 4: Syllabic hyphenation of the word идеално in Serbian

	S	И	Д	Е	FC	А	Л	Н	О	S
phonemme sonority	0	12	4	12	0	12	6	6	12	0
triplet difference		8	-20	8	-24	6	-12	-12	6	

3. Development of Syllabic Hyphenation of Macedonian and Serbian

The syllabic hyphenation was developed in Python. The algorithm used in all experiments consists of the same general steps with adjustments for each language. The steps in the base algorithm are the following:

1. Prepare the word – mark the beginning, end and put a marker between each two consecutive vowels.
2. Find the nuclei of the word.
3. Iterate over the characters in the word and form the syllables.

The steps to find the nuclei of the word are similar for both languages. First, the triplet difference is calculated for each group of three letters in the word, and then the nuclei are detected based on this calculation. Every vowel that has a nonnegative triplet difference with its preceding and following letter is detected as a nucleus, as well as the sonorant *р* surrounded by two consonants. For Serbian, the same rules are used, but additionally, the phonemes л and н can be nuclei if they have a positive triplet difference with the preceding and following phonemes.

The main difference between the experiments is in the way syllables are formed. In all cases, the formation of the syllable begins by adding phonemes until a nucleus is detected. In the baseline algorithm, this is followed by adding phonemes as long as the sonority of the current phoneme is higher than that of the following. If the sonority is not greater than that of the following phoneme, and the following phoneme is a consonant, then the first of the two phonemes is also added to the current syllable. This finalizes the formation of the syllable, and the algorithm continues by considering the next nucleus and again adding phonemes, starting one after the last added phoneme until this next nucleus is found.

The updated algorithm for Macedonian adds an additional rule to handle suffixes that shouldn't be split. These suffixes don't always appear at the end of a word – they can also appear in the middle, for example in declined forms of the word. This is done by simply looking at the phonemes surrounding the current phoneme. These are detected after the step of splitting two consonants with the same sonority. The check happens when the first phoneme of the wanted suffix is added to the current syllable. If the following 2, 3 or 4 phonemes, along with the current, form one of the suffixes that should be kept together (*ство*, *ствен*, *ски* and variations), then the current syllable is shortened and its last added phoneme is removed. In the next iteration, the first phoneme of the group is added to the syllable along with the other consonants that would usually be split, since they all come before the nucleus of the following syllable.

The algorithm for Serbian makes only one change to this algorithm. Here, the step where two consonants with the same sonority are split is skipped, and as soon as a consonant with a sonority that's not greater than that of the following is found, it directly checks if the word contains the groups that should be kept together, which are the same as for Macedonian.

4. Evaluation of Sonority Based Syllabic Hyphenation

The syllabic hyphenation of both languages started with the baseline model, which was our benchmark according to which we modified the approach. In Macedonian, it consisted of revoking the segmentation of the suffixes *ски*, *ство* and *ствен* and their inflections. Namely, according to the orthographic rules they should remain within one syllable, which contradicts our constraint of splitting the adjacent phonemes with identical sonority weights into two syllables. In Serbian, the adjustments were done by decreasing the sonority weight of non-syllabic *р*.

The Macedonian language assessment sample consisted of 1310 words within average 3.09 syllables. The accuracy of the baseline algorithm was 89.26%, mainly due to the frequent occurrence of words with the suffixes *ски*, *ство* and *ствен* and their inflections. By adjusting this, as explained in the previous section, the accuracy reached 96.63%. However, it affected the syllabification of the nouns: *гус-ки*, *мас-ки*, *прас-ки*, in which *ски* is not a morpheme. The only solution for effectively solving the problem of separable or non-separable suffix ski is the presence of the PoS tag word. This annotation overcomes the hyphenation problem.

An additional problem was caused by the ambiguous hyphenation of words with at least two adjacent consonants, such as the words: *автократ*, *Белград, декември* and *софтвер*. According to our approach, they were hyphenated as: *ав-ток-рат*, *Белг-рад*, *де-кемв-ри* and *соф-твер*, instead of the hyphenation suggested by the linguists: *ав-то-крат*, *Бел-град*, *де-кем-ври* and *софт-вер*. None of these automatic segmentations is incorrect as far as syllables are concerned. Namely, *крат*, *град* and *софт* are morphemes and they can be subject of further segmentation, depending on the context. Still, when it comes to hyphenation, it is better to leave them as a whole. Introduction of morphemes will be one way to improve the system.

We were curious to find out how ChatGPT copes with the same task. Without any previous training, ChatGPT performed much below our expectations. Namely, many of the words were transliterated into Latin script (for example, Au-gust instead of ав-густ and cada-stre for ка-та-стар), presenting non-existent words. Many words had a wrong spelling, for example *абонент* was hyphenated as *а-бо-нет* instead of *а-бо-нент*, *авангарда* as *а-ва-ган-да* instead of *а-ван-гар-да* or *атрибут* with *а-ти-бут* instead of *ат-ри-бут*. We then instructed ChatGPT with the rules for syllabic hyphenations and with the explanation to take care of the spelling and the Cyrillic script. Even with these instructions, the accuracy of hyphenation reached a modest 61.71%. Interestingly, after our explanations of what went wrong, this large language model gave examples confirming that it agrees with our suggestions, but repeated the same mistakes again, sometimes more than three times in a row.

The Serbian sample consisted of 3020 manually syllabified Serbian words. We first started with the same approach used for the Macedonian language, by simply modifying the sonority weights of the phonemes. After obtaining an accuracy of 75.34%, the constraint for the segmentation when the monotonic non-decreasing ends was adjusted with the constraint that the new syllable starts when the monotonic increase starts. With this adjustment, the accuracy of the baseline algorithm reached 97.59%, which was quite satisfactory. Most of the mistakes were related to words with non-sonoric *р*: *ар-ми-ја*, *мор-нар* and *у-пор-но* instead of *а-рми-ја*, *мо-рнар* and *у-по-рно*. Additionally, the inseparable Macedonian suffix *ски* has caused problems in the Serbian language as well, such as in the words: *гра-дског*, *љу-дски* and *сов-је-тска*. Apart from the fact that it was not segmented, it also collected the consonant that precedes it into a syllable. Several words containing the phoneme *к* (*лак-ше* instead of *ла-кше*, or *пот-сме-хом* instead of *по-тсме-хом*) were also not correctly syllabified. The last problem was the consonant group *пшт*, which should remain as a whole: *о-пште*, *са-о-пшти*, and *у-о-пште*.

By modifying the sonority of non-sylabic *р*, i.e. *р* which is not between two consonants from 8 to 6, the accuracy reached 98.68%, exceeding the rule-based syllabification accuracy [10]. Fine tuning of the sonority weights will be our next attempt to further improve the accuracy.

Similarly to Macedonian, we again tested ChatGPT. The accuracy of Serbian hyphenation was slightly better, reaching an acceptable 79.70% accuracy. Transliteration existed in a more subtle way: *ба-јо-net,* and *бу-dem*.

Spelling mistakes also occurred: *ве-чи-ном* and *бу-даш* instead of *ве-ћи-ном* and *бу-деш*. Consonants with the same sonority were separated: *ак-тив*, *бол-ни-це* and *бе-леж-ник* instead of *а-ктив*, *бо-лни-це* and *бе-ле-жник*.

When we informed ChatGPT that regardless of the training it had successfully completed, it had poorly realized the hyphenation of these two Slavic languages, the chatbot suggested that we try to repeat the same experiment with the professional version of Get Pro, the price of which is 200 US$ per month. For now, we decided to leave this offer, especially because our system is still in the development phase.

5. Conclusions and Further Work

Novacula Occami or Occam's razor [12] teaches us that the simplest solution is often the best one. In this study, we have seen that the most straightforward approach to syllabic hyphenation for Macedonian and Serbian, based on the sonority of the languages, is both practical and efficient. By prioritizing simplicity, we have achieved a robust model for hyphenation that respects the core principles of syllabic structure without unnecessary complexity.

The main reason contributing to the high accuracy of syllabic hyphenation is the phonemic orthography of both languages where one phoneme is represented by one grapheme in the Cyrillic script. The languages are closely related but have different syllabification patterns. In Macedonian, hyphenation is based more on the phonemic sonority, where sounds of the same sonority are separated, while in Serbian, the syllable boundary tends to preserve a stronger connection between phonemes, regardless of their phonetic features.

The approach we proposed is extremely simple and at the same time, very efficient. We intend to further improve it by taking into account the PoS tags for the Macedonian language and the exclusions for Serbian, hoping to reach an accuracy of over 99%. The simplicity of the model allows for better generalization and faster fine-tuning, while still capturing the essential patterns of syllabification in both languages.

References

[1] Walker, S. (2014). Typography & language in everyday life: Prescriptions and practices. Routledge.

[2] Häikiö, T., Bertram, R., & Hyönä, J. (2016). The hyphen as a syllabification cue in reading bisyllabic and multisyllabic words among Finnish 1st and 2nd graders. Reading and Writing, 29, 159-182.

[3] McCaskill, M. K. (1990). Grammar, punctuation, and capitalization: a handbook for technical writers and editors (Vol. 7084). National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Division.

[4] Schlücker, B. (2023). Compounding and linking elements in Germanic. In Oxford Research Encyclopedia of Linguistics.

[5] Mitreska, M., & Zdravkova, K. (2023, May). Syllable and Morpheme Segmentation of Macedonian Language. In 2023 46th MIPRO ICT and Electronics Convention (MIPRO) (pp. 1113-1118). IEEE.

[6] Zdravkova, K., & Kuzmanova, J. (2024). Sonority Based Syllabification of Macedonian and Serbian. Technical editors, 11.

[7] Krstev, C. (1991). Serbo-Croatian hyphenation: a TEX point of view. TUGboat, 12(2), 215-223.

[8] Vitas, D., Pavlović-Lažetić, G., & Krstev, C. (2007). About word length counting in Serbian (pp. 301-317). Springer Netherlands.

[9] Radojičić, M., Lazić, B., Kaplar, S., Stanković, R., Obradović, I., Mačutek, J., & Leššová, L. (2019). Frequency and length of syllables in Serbian. Glottometrics, 45, 114-123.

[10] A. Kovač, M. Marković M: A Mixed-principle Rule-based Approach to the Automatic Syllabification of Serbian. Contributions to Contemporary History / Prispevki za Novejšo Zgodovino. 2019.

[11] G. Clements: The sonority cycle and syllable organization, Phonologica, 63-76, 1988.

[12] Domingos, P. (1999). The role of Occam's razor in knowledge discovery. Data mining and knowledge discovery, 3, 409-425.

Слоговна подела речи македонског и српског језика применом звучности

Катерина Здравкова, Јана Кузманова

Сажетак

Овај рад разматра процес слоговне поделе речи македонског и српског језика применом оригиналног приступа који се заснова на звучности фонема. Његова главна претпоставка је да звучност фонема расте од приступа према језгру, да би опет пала према одступу. Да би се конструисао аутоматски систем за ефикасну и тачну слоговну поделу речи, македонске и српске фонеме су категорисане у неколико типова, а звучност ових фонема је дефинисана да би се одредиле границе слогова.

Оба језика користе самогласнике и слоготворни сонант р као носиоце слога. Скуп језгара у српском језику проширен је и сонантима л и н који могу бити слоготворни. Самогласницима и сонантима је додељена највиша звучност. Сугласници су подељени у две категорије: звучни и безвучни. Српски језик препознаје три додатне категорије: праскави, струјни и сливени, при чему звучност опада од праскавих према сливеним сугласницима.

Да би се одредила језгра и границе слогова, увели смо методу троструке разлике која израчунава разлику у звучности између суседних фонема. Она је позитивна само када је фонема језгро слога. Да би ово правило важило и у случају узастопних самогласника, између њих се додаје фиктивни консонант FC. Звучност FC и границе речи је 0.

Модели слоговне цртице су развијени у програмском језику Python, са прилагођавањима за македонски и српски језик на основу звучности.

Основни модел за македонски језик је постигао тачност од 88,70%. Погрешне поделе су се првенствено односиле на специјалне суфиксе ски, ство и ствен, који према правопису увек припадају истом слогу. Након прилагођавања, тачност подела у македонском језики се попела на 97,59%. Проблеми су укључивали руковање сложеним суфиксима и групама сугласника.

У српском језику, највећи изазов је било неслоготворно р, чија је звучност била смањена на 6. Овим побољшањем, коначна прецизност је достигла 98,68%.

Паралелно са сопственим приступом, упоредили смо резултате користећи ChatGPT. Прво смо га пажљиво обучили, а после првих

неуспешних покушаја, поновили смо обуку објашњавајући грешке. Упркос овим покушајима, ChatGPT је показао лошу тачност поделе речи на слогове (61,71% за македонски и 79,70% за српски), посебно због убацивања транслитерираних слова у неким речима, а каткад је правио и правописне грешке.

Приступ заснован на звучности који предлажемо показао се као једноставан, али ефикасан метод за поделу речи на слогове, који даје високу прецизност за оба језика. Очекујемо да би даља побољшања могла укључити информације у вези врсте речи за македонски језик и фино подешавање звучности за српски језик. Крајњи циљ ових модификација је да тачност подела за оба језика надмаши 99%.