Syllable and Morpheme Segmentation of Macedonian Language

M. Mitreska and K. Zdravkova
Faculty of Computer Science and Engineering, Skopje, Macedonia
maja.mitreska@students.finki.ukim.mk, katerina.zdravkova@finki.ukim.mk

Abstract - **Communication is the key to human development. Approximately 5% of the world's population experience some form of hearing disability. Modern assistive devices and technologies can improve the communication skills of hearing impaired people by transcribing the speech into text. The creation of such an application depends on the language specific morphosyntactic properties. It usually starts with the syllabification. The research presented in this paper focuses on the development of an automatic system for rule-based and sonority-based syllable and morpheme segmentation of Macedonian language, which can be easily incorporated into an efficient speech recognition system. The segmentation rules for breaking the words down into syllables and into morphemes were created according to the new orthography of the Macedonian language. For the sonority-based approach, a novel phonological distance measure was introduced capable of efficient syllable clustering. The implementation of the framework is developed in Python using several data structures for optimized performance and CPU usage. Both segmentation strategies were evaluated using the electronic lexicon consisting of more than one million words. A linguistic expert was consulted during the entire process. The consistency of the obtained results promises their sustainability for further speech processing applications.**

Keywords - communication; speech recognition; hearing impairment, word segmentation

I. Introduction

Communication is a key to human development that plays a vital role in everyday activities. It enables social relations and facilitates the exchange of information and wisdom. Since its development about 100000 years ago, human language has facilitated interaction, supporting considerably the birth and progress of civilization [1].

According to the World Health Organization, over 5% of the world's population have some hearing disorder that impacts their unobstructed cognition, education and employment, causing an annual global cost of around one trillion US$ [2]. The projection is that by 2050, nearly 2.5 billion people will have some kind of hearing loss [2].

Assistive technologies for hearing impaired people are categorized as assistive listening devices (ALD), augmentative and alternative communication (AAC) and alert systems (AS) [3]. ALD either amplifies the sound (for example, hearing loops and personal amplified systems), or wirelessly transmit it (FM, infrared and Bluetooth systems) [4].

AAC is intended for people with speech, language or communication disorder [5]. Powered by machine learning (ML) and deep learning (DL), these NLP based applications embrace word prediction, speech recognition and context processing [6]. AS encompasses notification systems, alert messaging and emergency communication systems intended to alert and protect people from risks and threats [7].

A team of young experts and researchers from the AI innovative company iReason (https://ireason.mk/) intensively work on the creation of assistive software modules for students with communication disorders [8]. One of their main goals are voice bot technologies, which rely on speech processing. To support the transcription of the Macedonian spoken language into text, we focused our research on syllable and morpheme segmentation. Syllable segmentation enables the discovery of the phonological structure of words [9]. The accuracy of speech recognition can additionally be improved by morpheme segmentation, which increases "the capability of understanding and producing new word forms" [10]. According to Yang et al. [11], the following four speech processing tasks: phoneme recognition (PR), automatic speech recognition (ASR), keyword spotting (KS), and query by example spoken term detection (QbE) are crucial to establish a universal performance benchmark. The first two are used to transcribe speech into text, the latter to detect the spoken content [11].

The paper presents the recently developed automatic system for syllable and morpheme segmentation of Macedonian language. In the absence of a larger available corpus of already segmented words, the realization of both tasks was rule-based. The segmentation rules, prefixes and suffixes were extracted from new Macedonian orthography [12]. The sonority-based rules are an original contribution of this paper. The main language resource was the annotated electronic lexicon [13], which is available from the CLASSLA CLARIN knowledge centre for South Slavic languages [14].

The paper continues with the second section, which announces the linguistic background. Third and fourth sections introduce rule-based and sonority-based syllable and morpheme segmentation. Fifth section presents the application development. Sixth section is dedicated to evaluation of the results and identification of the major anomalies of both segmentations. The paper concludes with the further improvements and the impact of word segmentation to speech recognition of Macedonian.

II. Linguistic background

Phonetically, syllables are sequences of sounds “containing one peak of prominence” [15]. Phonologically, they are units of accent placement [15]. Syllables must contain one vowel, which is the voiced, central-oral frictionless sound [18]. For example, the words *apple*, *butterfly*, *frog*, *table*, and *tomato* are divided into the following syllables: *ap-ple*, *but-ter-fly*, *frog*, *tab-le*, and *to-ma-to*, each containing exactly one vowel.

Morphemes are defined as ‘the minimal meaning-bearing units of a language’ [16]. They express the internal structure of complex words [17]. Morphemes are composed of two separate classes: bases (according to some linguists, roots) and affixes [16]. Those affixes that precede the base are prefixes, those that follow it are suffixes. For example, the word *unhealthy* consists of the prefix *un*, the base *health* and the suffix *y*.

A. Syllable segmentation

Syllable segmentation, also called syllabification or syllabication can be performed implementing a data-driven or a rule-based approach [18]. Data-driven methodology requires a large corpus of already syllabified words [18], [19]. By implementing various machine learning techniques the accuracy can reach 95% [19]. The absence of such a corpus can be bypassed by implementing the rule-based approach. Instead of learning the syllabification from the training set, a concise set of mutually exclusive rules should be established, enabling the correct segmentation [18]. The accuracy is again very high, proving that rule-based segmentation is not inferior to data-driven [20].

B. Morpheme segmentation

Morpheme segmentation methods are usually classified as rule-based, data-driven and hybrid [21]. Rule-based methods detect morpheme boundaries statistically, using for example, letter variety statistics [22] or conditional random fields [23]. Data-based methods can be supervised, semi-supervised and unsupervised [21]. A very successful supervised morpheme segmentation applied convolutional neural networks [24]. Semi-supervised learning that combines several unsupervised segmentation techniques using conditional random fields proved its efficiency for Finish [25]. A good example of a successful unsupervised segmentation was achieved with Bi-LSTM neural network [26]. Hybrid methods can combine rule-based and statistical approaches coupled with unknown morpheme guessing [27] or rule-based morpheme word representation coupled with unsupervised morphological analysis [28].

C. Macedonian language specific phonetic features

The Macedonian language is a phonetic language and the relationship between spoken sounds (phonemes) and written sounds (graphemes) is very strong [29]. Therefore, the linguistic rules of forming syllables, as well as morphemes are very strict [30].

Macedonian language has five vowels: *а* (*a*), *е* (*e*), *и* (*i*), *о* (*o*) and *у* (*u*) and no diphtongs. Most of the words, like: *воз* (Latinized: *voz* / English: *train*), *вода* (*voda* / *water*), *воздушест* (*vozdushest* / *airy*) and *вообразен* (*voobrazen* / *conceinted*) are syllabized with a sequence of one up to five phonemes, one of them compulsory a vowel: *воз*, *во-да* and *воз-ду-шест* and *во-о-бра-зен*.

The sonorant *r* (in Cyrillic, *р*) can also be a syllable-carrier [31] if found in one of the following contexts:

- It appears in the middle of a consonant group. Such words are: *крв* (*krv* / *blood*), *првак* (*prvak* / *champion*), *здрвен* (*zdrven* / *stiff*);
- It is preceded by an apostrophe, which replaces the hard sign existing in some Macedonian dialects and is followed by a consonant at the beginning of the word, such as: *’рбет* (*‘rbet* / *spine*), *’рж* (*‘rzh* / *rye*), *’ртење* (*‘rtenje* / *germination*);
- It follows a consonant at the end of the word. This situation occurs in a few words of foreign origin, like *масакр* (*masakr* / *masacre*) and *жанр* (*genre*) [12].

The list of prefixes is rather long. They can consist of one syllable only: *без* (*bez*), *ис* (*is*), *по* (*po*), and *се* (*se*), or several syllables: *кусо* (*kuso*), *обез* (*obez*), *прет* (*pret*), and *сино* (*sino*). The list of suffixes is even longer, because the language is highly inflected [32]. They depend on the POS tag. For morpheme segmentation, the suffixes *ски* (*ski*), *ство* (*stvo*) and *ствен* (*stven*), which are nominal or adjectival trigger the corresponding inflections for gender, number and definiteness [33].

Macedonian language is a low-resourced language and it is lacking the manually segmented lexicon. Therefore, the only applicable methods implemented during our research could be rule-based. They comprise two flows: segmentation according to the rules extracted from the orthography and segmentation according to sonority of the phonemes. Both flows were independently done aiming to perform syllable and morpheme segmentations. The following two sections introduce them briefly.

III. Rule-based segmentation

The syllable segmentation was performed by exploiting the phonological distances measures [18] and by implementing a finite set of segmentation rules for the Macedonian [12]. Morpheme segmentation was done implementing the linguistic knowledge originating from the Macedonian orthography [12] in conjunction with the lemmas from the lexicon [13].

A. Macedonian syllabification rules

These are the most simplified rules that were incorporated in the rule-based syllable segmentation of Macedonian language.

1. Each syllable should contain exactly one vowel or the syllable-carrier sonorant *р* (*r*).
2. A syllable can consist of one vowel only independently on its position in the word: *авион* (*avion* / *plane*; *а-ви-он)*, *аурора* (*aurora* / *aurora*, *а-у-ро-ра*, *меана* (*meana* / *tavern*; *ме-а-на)* and *тргнаа* (*trgnaa* / *started*; *трг-на-а*).

3. If a consonant group appears at the beginning of the word, then the syllable consists of all the consonants and the vowel or the embedded sonorant *р* (*r*) it is followed by: *здравство* (*здравство* / *health*; *здрав-ство*), *според* (*spored* / *according*; *спо-ред*), *крвав* (*krvav* / *bloody*; *кр-вав*).
4. If a consonant group appears at the end of the word, then the syllable consists of all the consonants and the vowel or the sonorant *р* (*r*) that precede it: *амбиент* (*ambient* / *ambience*; *ам-би-ент*), *радост* (*radost* / *happiness*; *ра-дост*), *накрст* (*nakrst* / *crosswise*; *на-крст*).
5. If a consonant group appears in the middle of the word, then it is divided in half, with the beginning belonging to the first and the end to the next syllable: *коска* (*koska* / *bone*; *кос-ка*), *тетратка* (*tetratka* / *notebook*; *тет-рат-ка*, *поттик* (*pottik* /*motive*; *пот-тик*). Longer consonant groups, like in *исклучи* (*iskluchi* / *turn off*; *ис-клу-чи)* or *оздрави* (*ozdravi* / *recover*) should be treated separately to reduce incorrect segmentation caused by inconsistent division.
6. An exception to the rule 5 are the nouns that end in *ство* (*stvo*) and the adjectives that end in *ски* (*ski*) and *ствен* (*stven*). These three suffixes always remain undivided: *чувство* (*чувство* / *emotion*; *чув-ство*), *градски* (*gradski* / *urban*; *град-ски*), *единствен* (*edinstven* / *unique*; *један-ствен*). The inflections for: gender, plural, definite, distal and proximal definiteness preserve the rule 6.
7. The following suffixes: *штво* (*shtvo*), *шки* (*shki*), and *чки* (*chki*) can be divided according to the rule 5 or can remain undivided obeying the rule 6.

B. *Macedonian morphemes*

Morphemes in the Macedonian language are crucial for word formation and for inflection [12]. For example, the noun *учител* (*uchitel* / *teacher*) consists of two morphemes: the verb *учи* (*uchi* / *to teach*) and the suffix *тел* (*tel*). The noun *надградба* (*dogradba* / *annex*, *extension*) consists of three morphemes: the prefix *над* (*nad*), the verb *гради* (*to build*) and the suffix *ба* (*ba*).

Morpheme segmentation starts with the prefixes. A specific attention was paid to prefixes consisting of one vowel only, like the vowel *и* (*i*) in the adjective *иреален* (*irealen* / *unreal*; *и-ре-а-лен*).

Considering the rich inflectional paradigm of many Macedonian word categories [33] suffixes were searched throughout all of the word. While the suffix *ствен* (*stven*) remains unchanged during inflection, *ски* (*ski*) and *ство* (*stvo*) alter the final vowel for gender and number.

To support the speech recognition, all the prefixes, word bases and suffixes consisting of more than one syllable were additionally syllabized. This is not part of the morpheme segmentation, but it can significantly facilitate the recognition of longer words, particularly those that are part of an accentual whole.

C. *Phonological distance for syllable segmentation*

In 1965, Levenshtein defined phonological distances aiming to correct the mistakes of binary information caused by reversing, losing or adding of some binary value during transmission [34]. Levenshtein's distance inspired the development of two additional measures: the dialect measurement [35] and the measure for language classification [36].

In 1988, Clements introduced the sonority sequencing principle (SSP), according to which the sonority within a syllable rises to the syllable nucleus (in Macedonian, the vowels and the sonorant *р* (*r*)) and fall in sonority thereafter [37]. Although disputed, this principle was applicable to Macedonian language. SSP was combined with the phonological distances in the novel sonority-based segmentation method, which was implemented for both tasks: syllable and morpheme segmentation. The following section introduces the new method in more detail.

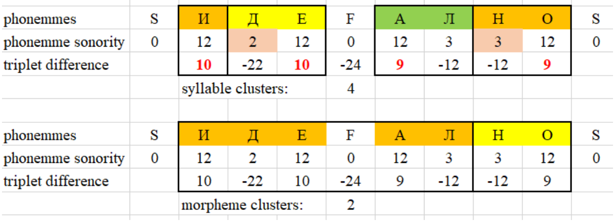
IV. SONORITY-BASED SEGMENTATION

The sonority of the Macedonian phonemes is highest for vowels. It is falling down from sonorants towards voiced and voiceless consonants (see Table 1). The list is expanded with the special delimiters S and F.

TABLE I. SONORITY OF MACEDONIAN PHONEMES

Type	**Macedonian phonemes**	
	List of phonemes	***Weight***
Vowels	*а* (*a*), *е* (*e*), *и* (*i*), *о* (*o*), *у* (*u*)	12
Sonorant *р*	*р* (*r*)	5
Sonorants	*ј* (*j*), *л* (*l*), *љ* (*lj*), *м* (*m*), *н* (*n*), *њ* (*nj*)	3
Voiced consonants	*б* (*b*), *в* (*v*), *г* (*g*), *д* (*d*), *ѓ* (*gj*), *ж* (*zh*), *з* (*z*), *ѕ* (*dz*), *џ* (*dzh*)	2
Voiceless consonants	*к* (*k*), *п* (*p*), *с* (*s*), *т* (*t*), *ќ* (*kj*), *ф* (*f*), *х* (*h*), *ц* (*c*), *ч* (*ch*), *ш* (*sh*)	1
Special	‘, spacing (S), fictive consonant (F)	0

The syllabification starts with the preprocessing of the words, which adds a fictive consonant F to phonologically separate the consecutive vowels (Fig. 1).



phonemmes	S	И	Д	Е	F	А	Л	Н	О	S
phonemme sonority	0	12	2	12	0	12	3	3	12	0
triplet difference		10	-22	10	-24	9	-12	-12	9	
		syllable clusters:			4					

phonemmes	S	И	Д	Е	F	А	Л	Н	О	S
phonemme sonority	0	12	2	12	0	12	3	3	12	0
triplet difference		10	-22	10	-24	9	-12	-12	9	
		morpheme clusters:			2					

Figure 1. Segmentation of a word with two consecutive vowel

If the phonemes are denoted with the letter *p*, and their sonority weights with *w*, then the word can be represented with the string:

$$w(p_i),\ i = 1,\dots,n \qquad (1)$$

n being the number of phonemes and fictive consonants.

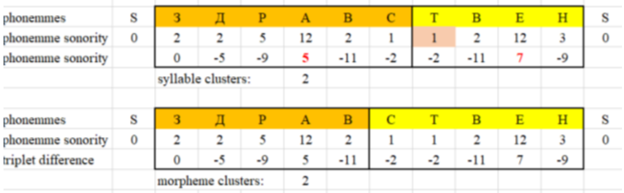
The sonority of the spacing S before $w(p_0)$ and after the word $w(p_{n+1})$ is 0. For each phoneme, a triplet difference *TD* is calculated as:

$$TD(p_i)=w(p_{i\_}) - w(p_{i-1}) - w(p_{i+1}),\ i = 1,\ldots,n \qquad (2)$$

Triplet difference of the vowels is always positive (Fig. 1 and 2). This property was achieved by denoting the weights 12 to vowels, which guarantee that the difference is bigger than 2, even when the vowel is surrounded by two sonorants, like in the word *апарат* (*aparat* / *device*).

Morpheme segmentation was done by going through all the prefixes, the potential bases matched with the lemmas corresponding to nouns and verbs and finally the suffixes extended with their inflections. To support speech recognition, they were additionally syllabized according to both syllable segmentation methods.

Following SSP, the sonority of the phonemes within a syllable is monotonically increasing towards the syllable nucleus. Syllable border is the phoneme after which the strictly monotonic decrease of the sonority ends (Fig. 1 and 2, highlighted with apricot). According to this rule, the adjective *идеално* (*idealno* / *ideal*) has four syllables: *и* (*i*), *де* (*de*), *ал* (*al*) and *но* (*no*). Its morpheme segmentation consists of the base *идеал* (*ideal*) and the suffix *но* (*no*). Syllabized morpheme segmentation of this word is identical with the syllable segmentation. Syllable and morpheme segmentation of two-syllable words with larger consonant groups like *здравствен* (*zdravstven* / *health*) usually differ. While the sonority-based rule suggests the syllables *здравс* (*zdravs)* and *твен* (*tven*), Macedonian orthography and morpheme segmentation are identical: *здрав* (*zdrav / healthy*) and *ствен* (*stven*).



phonemmes	S	З	Д	Р	А	В	С	Т	В	Е	Н	S
phonemme sonority	0	2	2	5	12	2	1	1	2	12	3	0
phonemme sonority		0	-5	-9	5	-11	-2	-2	-11	7	-9	
		syllable clusters:			2							
phonemmes	S	З	Д	Р	А	В	С	Т	В	Е	Н	S
phonemme sonority	0	2	2	5	12	2	1	1	2	12	3	0
triplet difference		0	-5	-9	5	-11	-2	-2	-11	7	-9	
		morpheme clusters:			2							

Figure 2. Segmentation of a word with longer consonant groups

V. APPLICATION DEVELOPMENT

The proposed segmentations were developed in Python for simple integration and further use in various systems, including its embedding into automation tasks.

The framework consists of two main parts, rule-based and sonority-based syllable and morpheme segmentation. Its inter-process communication is illustrated with the pipeline presented in the Fig. 3.

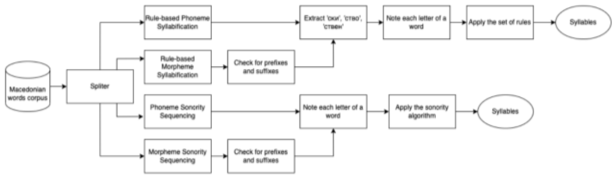

Macedonian words corpus
Spliter
Rule-based Phoneme Syllabification
Rule-based Morpheme Syllabification
Check for prefixes and suffixes
Phoneme Sonority Sequencing
Morpheme Sonority Sequencing
Check for prefixes and suffixes
Extract 'ски', 'ство', 'ствен'
Note each letter of a word
Apply the set of rules
Syllables
Note each letter of a word
Apply the sonority algorithm
Syllables

Figure 3. A pipeline of splitting words into their syllables.

Before the segmentation, the dataset was clustered into ten subsets that consist of words with different phoneme structuring. The division of the lexicon was predominantly made to facilitate the evaluation of both segmentations in detail, with the intention of discovering where the new sonority-based approach yields different results than the rule-based approach.

VI. EVALUATION OF BOTH APPROACHES

Since there is no previous work done in this field, and there are no existing records that contain Macedonian words and their respective syllable segmentation for comparison, an annotated lexicon consisting of 1236537 Macedonian words [13], which was run through the four types of segmentation. For evaluation purposes, the two phoneme-based and two morpheme-based approaches were compared with each other aiming to measure how much the segmentations from one approach matches or differs from the other approach. This evaluation method was considered relevant since one of the approaches is based on the rules of syllabification in the Macedonian language and each output that this method obtains is considered a true match. So, the evaluation was made by matching the outputs of the phoneme rule-based segmentation as a reference output and the outputs from the phoneme sonority-based segmentation.

Table II represents the matches between the two implemented approaches for the 10 clusters. The last one is further divided depending on the suffixes that have a special treatment in Macedonian orthography.

TABLE II. EVALUATION OF SEGMENTATION ACCURACY BY MATCHING BOTH APPROACHES

ID					
	Cluster type	***Number of words***	***Frequency***	***Phoneme***	***Morpheme***
1	Alternating vowels and consonants	154476	12.49%	1.000	0.973
2	Two-phoneme consonant group	428537	34.66%	0.992	0.944
3	Three-phoneme consonant group	178426	14.43%	0.628	0.746
4	Four-phoneme consonant group	43860	3.55%	0.774	0.825
5	Sonorant ‘*р*’	73803	5.97%	0.971	0.927
6	Two consonant groups	380489	30.77%	0.889	0.883
7	Several consonant groups	128129	10.36%	0.857	0.871
8	One vowel group	241637	19.54%	0.914	0.934
9	Two vowel groups	21379	1.73%	0.910	0.956
10.1	Words with ‘*ски*’	21405	1.73%	0.187	0.957
10.2	Words with ‘*ство*’	2511	0.20%	0.089	0.937
10.3	Words with ‘*ствен*’	2273	0.18%	0.000	0.969

The accuracy of the sonority-based approach presents an effective approach that does not rely on a set of fixed rules. More noticeable mistakes can be detected in the third and fourth subsets where the consonant group consists of three or four consonants. Namely, the sonority-based approach usually shifts one of the consonants into the previous syllable, particularly when the sonority of the phonemes is identical. One such word is *авторскана* (*avtorskana* / *that author's*). The phoneme rule-based approach outputs *ав-тор-ска-на*, while the phoneme sonority-based approach outputs *ав-торс-ка-на*.

The same problem can be also noticed in the tenth subset where the segmentation performed with the phoneme approach has low accuracy, however when the segmentation takes into consideration the morphemes, the accuracy significantly increases. This is due to the orthographic rule that the phonemes in the suffixes *ски* (*ski*), *ство* (*stvo*) and *ствен* (*stven*) are never separated, although according to their sonority they must be separated after the phoneme *с* (*s*). A typical example of this case is the word *армиски* (*armiski* / *army*), where the phoneme rule-based approach gives *ар-ми-ски* (*ar-mi-ski*), while the sonority-based approach results in *ар-мис-ки* (*ar-mis-ki*).

The higher accuracy in the morpheme approach is due to the general extraction of the prefixes and suffixes which are most of the time one-syllable morphemes.

An additional interesting case are the words starting in a voiced consonant that are followed by a voiceless consonant. This is quite rare, because in the Macedonian language sonority equalization is usually applied, according to which when two consonants with different sonorities are next to each other, the voiced consonant is transformed into the corresponding voiceless pair [13]. There are some exceptions to this rule that validate the rule-based approach, but may show errors in the sonority-based approach because the segmentation is done by considering and weighing all the letters of the word. The most interesting example of this are the words *вторник* (*vtornik* / *Tuesday*) and *вчера* (*vchera* / *yesterday*), where the rule-based approach produces the segmentations *втор-ник* (*vtor-nik*) and *вче-ра* (*vche-ra*), while the sonority-based segmentations results *в-тор-ник* (*v-tor-nik*) and *в-че-ра* (*v-che-ra*), where the evaluation of syllable number fails. The key reason for this inaccuracy is the fact that the phoneme *в* (*v*) is not a syllable carrier.

Regarding morpheme segmentation, the two approaches give similar results since they pay attention to the detection of prefixes and suffixes, and the difference that is detected is in the additional syllabic definition of the morpheme.

VII. CONCLUSION AND FURTHER WORK

The paper presents two seemingly trivial tasks. They were created using rules and an original sonority-based approach. Both segmentations were exhaustively manually evaluated with hundreds of words automatically segmented and evaluated by the authors and an expert.

The comparison of the results is more than promising. Reaching a segmentation match of 90.12% during syllabification of already morphologically divided words, it will undoubtedly become a valuable contribution to various upcoming language technologies.

A detailed evaluation of the implementation of lexical clustering proves the superiority of syllabized morpheme segmentation compared to the traditional rule-based syllabification.

The authors believe that after fixing the perceived inaccuracies the sonority-based syllabification of morpheme segmentation can be considered the most adequate for word segmentation in the Macedonian language and used for hyphenation within the text processor.

Even without any adjustments, the sonority-based segmentation is more than sufficient to support the Macedonian language speech recognition system that is under construction. Its simplicity will contribute to the efficient recognition of oral language and its translation into text, which is a key component of mobile applications for the hearing impaired. By using such an assistive technology, they will be able to establish uninterrupted two-way communication with people who do not know the Macedonian sign language.

ACKNOWLEDGMENT

Sincere gratitude to Dr. Elka Jačeva Ulčar from the Institute for Macedonian Language, who contributed to the development of the system with her expert advice. This work is partially financed by the Faculty of Computer Science and Engineering (FSCE) at the Ss. Cyril and Methodius University in Skopje.

REFERENCES

[1] WHO, Deafness and hearing loss, retrieved from https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss on 5 January 2023.

[2] S. Miyagawa, S. Ojima, R. C. Berwick, and K. Okanoya. "The integration hypothesis of human language evolution and the nature of contemporary languages." Frontiers in psychology 5, 2014, p. 564.

[3] A. Dhanjal, and W. Singh. "Tools and techniques of assistive technology for hearing impaired people." International conference on machine learning, big data, cloud and parallel computing (COMITCon), pp. 205-210. IEEE, 2019.

[4] C. Nieman, and E. S. Oh. "Hearing loss." Annals of internal medicine 173, no. 11, (2020, ITC81-ITC96.

[5] D. Beukelman, and P. Mirenda. Augmentative and alternative communication. Baltimore: Paul H. Brookes, 1998.

[6] Y. Elsahar, S. Hu, K. Bouazza-Marouf, D. Kerr, and A. Mansor. "Augmentative and alternative communication (AAC) advances: A review of configurations for individuals with a speech disability." Sensors 19, no. 8, 2019, p. 1911.

[7] K. Zdravkova, V. Krasniqi, F. Dalipi, and M. Ferati. "Cutting-edge communication and learning assistive technologies for disabled children: An artificial intelligence perspective." Frontiers in Artificial Intelligence, 2022, p. 240.

[8] K. Mishev, A. Karovska Ristovska, O. Rashikj-Canevska, and M. Simjanoska. "Assistive e-Learning Software Modules to Aid Education Process of Students with Visual and Hearing Impairment: A Case Study in North Macedonia.", ICT Innovations 2021, pp. 145-159. Springer, Cham, 2022.

[9] C. Ewen and H. Van Der Hulst: The phonological structure of words: an introduction. Cambridge University Press, 2001.

[10] M/ Creutz: Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition. Helsinki University of Technology, 2006.

[11] S. Yang, P. Chi, Y.Chuang, C. Lai, K. Lakhotia, Y. Lin, A. Liu et al. "Superb: Speech processing universal performance benchmark." arXiv preprint arXiv:2105.01051, 2021.

[12] Institute of Macedonian Language “Krste Misirkov”, Orthography of the Macedonian language, second edtion, Kultura, 2017, https://pravopis.mk/sites/default/files/Pravopis-2017.PDF

[13] Petrovski, A. Morphological computer dictionary –contribution to Macedonian language resources, PhD thesis, Ss. Cyril and Methodius University in Skopje, 2008, in Macedonian

[14] CLASSLA: Knowledge centre for South Slavic languages, FAQ for Macedonian language resources and technologies, https://www.clarin.si/info/k-centre/faq4macedonian/

[15] J. O'Connor and J. Trim. "Vowel, consonant, and syllable—

[16] J. Krámský: "The word as a linguistic unit." In The word as a linguistic unit. De Gruyter Mouton, 2019.

[17] A. Kovač, and M. Marković: A Rule-Based Syllabifier for Serbian, Proceedings of the Conference on Language Technologies and Digital Humanities, pp. 140-46. 2018.

[18] N. Sanders and S. Chin: Phonological distance measures, Journal of quantitative linguistics 16, no. 1, pp 96-114, 2009.

[19] Y. Marchand, C. Adsett and R. Damper.: Automatic syllabification in English: A comparison of different algorithms." Language and speech 52, no. 1, pp 1-27, 2009.

[20] F. Asahiah: Comparison of rule-based and data-driven approaches for syllabification of simple syllable languages and the effect of orthography." Computer Speech & Language 70: 101233. 2021.

[21] T. Zundi and C. Avaajargal: Word-level morpheme segmentation using transformer neural network, Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 139-143. 2022.

[22] Ç. Çöltekin: Improving successor variety for morphological segmentation., LOT Occasional Series 16, pp. 13-28. 2010.

[23] H. Liu, M. Li, J. Zhang and L. Chen. :Morpheme segmentation using bilingual features, 2012 International Conference on Asian Language Processing, pp. 209-212. IEEE, 2012.

[24] A. Sorokin and A. Kravtsova: Deep convolutional networks for supervised morpheme segmentation of Russian language, Conference on Artificial Intelligence and Natural Language, pp. 3-10. Springer, Cham, 2018.

[25] T. Ruokolainen, O. Kohonen, S. Virpioja and M. Kurimo: Painless semi-supervised morphological segmentation using conditional random fields., Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, pp. 84-89. 2014.

[26] E. Bolshakova and A. Sapin: Bi-LSTM Model for Morpheme Segmentation of Russian Words. Conference on Artificial Intelligence and Natural Language, pp. 151-160. Springer, Cham, 2019.

[27] G. Lee, J. Cha and J. Lee: Syllable-pattern-based unknown-morpheme segmentation and estimation for hybrid part-of-speech tagging of Korean." Computational Linguistics 28, no. 1, pp. 53-70, 2002.

[28] M. Luong, P. Nakov and M. Kan: A hybrid morpheme-word representation for machine translation of morphologically rich languages., arXiv preprint arXiv:1911.08117. 2019.

[29] D. Spasovski, G, Peshanski and G, Madjarov: The influence the training set size has on the performance of a digit speech recognition system in macedonian." In International Conference on ICT Innovations, pp. 205-212. Springer, Cham, 2014.

[30] W. Ghai and N. Singh: Literature review on automatic speech recognition. International Journal of Computer Applications 41, no. 8, 2012.

[31] B. Morén: Consonant–vowel interactions in Serbian: Features, representations and constraint interactions, Lingua 116, no. 8, :pp. 1198-1244, 2006.

[32] K. Zdravkova: Resolving Inflectional Ambiguity of Macedonian Adjectives, LREC 2022 Workshop Language Resources and Evaluation Conference, p. 60. 2022.

[33] M. Bonchanoski and K. Zdravkova: Learning syntactic tagging of Macedonian language, Computer Science and Information Systems 15, no. 3, pp. 799-820, 2018.

[34] V. Levenshtein: Binary codes capable of correcting deletions, insertions, and reversals., Soviet physics doklady, vol. 10, no. 8, pp. 707-710. 1966.

[35] J. Nerbonne and W. Heeringa: Measuring dialect distance phonetically." In Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology. 1997.

[36] T. Dunning: Statistical identification of language. Las Cruces: Computing Research Laboratory, New Mexico State University, 1994.

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