# Solution for quantitative analysis of texts in Serbian based on syllables

Sebastijan Kaplar*, Marija Radojičić*, Ivan Obradović**, Biljana Lazić**, Ranka Stanković**
* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
** University of Belgrade, Faculty of Mining and Geology, Belgrade, Serbia
kaplar@uns.ac.rs, marija.radojicic@uns.ac.rs, ivan.obradovic@rgf.bg.ac.rs, biljana.lazic@rgf.bg.ac.rs, ranka.stankovic@rgf.bg.ac.rs

***Abstract*— The aim of this paper is to present an approach to quantitative analysis of syllable distribution in Serbian texts. The approach is based on creating a software that allows automatic processing of texts in Serbian, and which will have as a result the text with words divided in syllables, according to the syllabification rules for Serbian. Also, the program should give a detailed quantitative analysis of the processed text, which included the analysis of syllables by type, length, frequency, words, and text. In this paper some of the initial results are presented.**

## I. INTRODUCTION

Slavic languages are Indo-European languages spoken by the people who mostly inhabit Eastern Europe, the Slavs. Slavic languages descend from Proto-Slavic, their parent language, which was derived from Proto-Indo-European, the ancestor language for all Indo-European languages. During their common existence, a number of isoglosses in phonology, morphology, lexis, and syntax developed for the Slavic languages [1]. The approach described in this paper is based on the fact that all Slavic languages share the same root. This means that some, if not the majority of the rules, can be applied to several Slavic languages.

The main goal is to create a software solution for processing syllables in several Slavic languages. The initial goal was to automatize processing of syllables in Serbian, which is important for hyphenation and as additional information in dictionaries.

In order to achieve this goal, the software solution was developed. It relies on the syllabification rules for Serbian language. The software can extract and give detailed information about types of syllables and their distribution, which will be presented in this paper.

## II. MOTIVATION

As previously mentioned, the idea was to automatize the processing of syllables in Serbian which is important for hyphenation. At the same time, syllabification can be useful as additional information in dictionaries. Developed software can extract and give detailed information about types of syllables and distribution of syllables. The software is specially adapted as a tool for detailed analysis of syllables, which have an important role in creating a mathematical model for Slavic languages.

This software can handle text in two alphabets used in Serbia, Latin and Cyrillic, with text in ekavian and iekavian pronunciation. According to literature there is a set of recognized syllable types in Serbian [2],[3],[4], which the software can recognize and give detailed information on them. Also, the software can calculate the total number of different syllables and different syllables per type. The initial analysis was performed on the novel **"The Master and Margarita"**.

## III. RESEARCH QUESTIONS

Natural language is challenging to process, and each language is specific, with a complex nature and set of rules that are used. The goal of this research was to perform a quantitative analysis of the syllables in Serbian by type, length, and frequency. Syllables have not been mathematically modelled systematically, and the main reason is the problem with their definition (i.e., with word syllabification). Syllabification can be performed algorithmically, where the focus is on the models for syllable frequency, type and length, and which prepares the data, among other things, for further statistical tests. Based on the analysis of those results, a mathematical model that describes syllables in Serbian is planned, which will be followed by models in other Slavic languages.

## IV. METHODOLOGY

A syllable is a sound unit composed of an uninterrupted sound sequence which is pronounced with a single opening of the mouth. The software presented is based on syllabification rules proposed in Serbian grammar written by Stanojčić and Popović [5]. Each syllable in Serbian has one vowel (a, e, i, o, u) or sonorant (r, l, n), which is the syllable "nucleus". Also, a syllable can be composed of only one vowel, for example, prepositions *u* and *o*.

A general rule defines syllables ending in a vowel (open syllable type – example *ma-ma*), or in a consonant (closed syllable type – example *ot-vor*).

In addition, there are five rules related to phonetic nature of the syllables:

- If there is a group of consonants, and the first one is fricative (z, ž, f, s, š or h) or affricate (dž, đ, c, č or ć), the syllable ends before that group. Examples: *o-sta-li-ma, dvo-ri-šte, ko-nji-čka*.
- If there is a group of consonants, where the first consonant is not a sonorant (v, r, j, l, lj, n, nj or m) and the second is a sonorant v, j, r, l or lj, the syllable ends before that group. Examples: *u-pla-še-no, sve-tlo-šću, ne-str-plji-vo*.
- If there are two sonorants next to each other, they belong to separate syllables. It means that the

syllable border is between them. Examples: *raz-um-lji-vo, cr-ve-na.*

- If a plosive consonant (b, d, g, p, t or k) is followed by some other consonant, except (j, v, l, lj or r), the syllable border is between them. Examples: *pro-šap-ta, pred-sed-ni-ka.*
- If there is a group composed of two sonorants, where the sonorant j that belongs to ijekavian "je" (corresponding to ekavian e) is on the second position, the syllable closure is before that group. Examples: *čo-vjek, go-rje-ti.*

In addition, there are semantic rules for syllabification in Serbian. These rules have priority over the aforementioned rules. Examples can be found in compound words, for example, the verb *razljutiti*. Phonetic syllabification would be *ra-zlju-ti-ti,* and semantic syllabification, which has a higher priority, is *raz-lju-ti-ti*. In this case the prefix *raz* is one syllable.

## V. SOLUTION

### A. Discussion

A software solution was developed which can process and perform analysis on textual materials. The novel **"The Master and Margarita"** was analysed. It have been chosen, together with the novel **"How the Steel Was Tempered"** which will be analysed later, having in mind the existence of translations in other Slavic languages for the purpose of comparison of results. These texts in Serbian were given as an input to the program, which then performed and applied the aforementioned rules.

Text pre-processing was performed on the textual file in order to clear the words from clutter (i.e. accidental space, tab, comma...). The words were then split into syllables, based on the rules for Serbian. The process of obtaining syllables was not easy to perform, because it had to follow all the rules for Serbian, and the algorithm went through several iterations before the final one was accepted. When the syllables extracted, their type was determined, their length and frequency calculated, and data for further statistical processing prepared.

### B. Architecture

The software architecture consists of three main components (Figure 1). The first component is a text pre-processor, the second and the most important one is the text analyzer, and the third component produces the output of the analyzed text, which will serve as the input for the mathematical model.

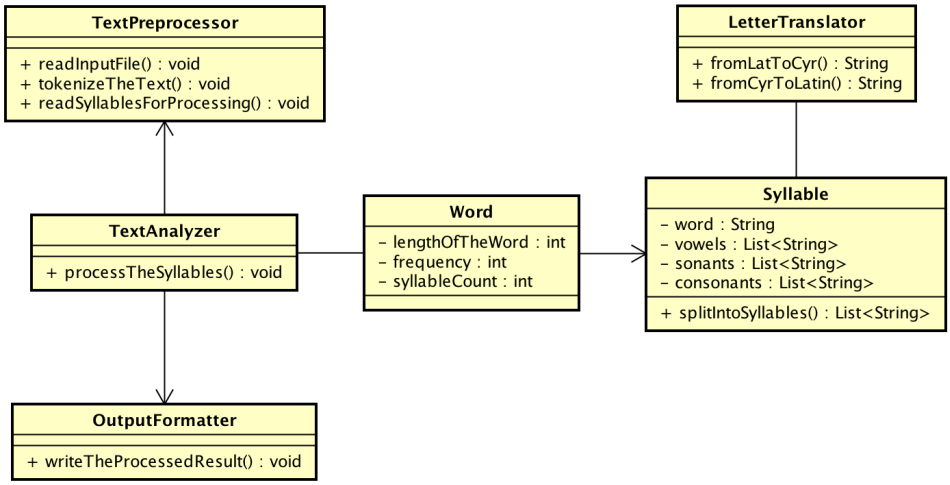



Figure 1. Simplified software architecture

The text pre-processor has several tasks to perform in order to prepare the text for the text analyzer. . During the process of reading the input file, which can be in the Latin or the Cyrillic alphabet, the text pre-processor has three main tasks:

1. To form tokens (which are basically words) and to clean the text form clutter
2. If the text is in Latin, to convert it to Cyrillic for further processing
3. To prepare the structures that will be populated by the text analyzer (second component) during its work.

Forming correct tokens is crucial for the next steps in the application, and needs to be done with great care. If the tokens are formed incorrectly, the application will not produce a satisfactory result, and the number of mistakes will be large. In order to produce a satisfactory result, a lot of attention was given to token extraction and to cleaning of the input text from clutter.

In order to enhance the analysis, every token (word) was translated from Latin to Cyrillic, assuming that it wasn't already written in Cyrillic. This transcription is very useful for software manipulation because it minimizes ambiguity. For example, the letter *lj* in Serbian is written with two characters in Latin, as

opposed to Cyrillic where there is only one-character љ, which corresponds to the rule in Serbian, where one letter is one sound.

While forming tokens, it was necessary to organize them in a way that enables the calculation of the number of times each token appeared, its length, number of vowels, etc. This information was used in the statistical analysis that was performed on the results obtained. Tokens were organized using dictionaries (i.e. Maps) in such a way that the key of the map is the token itself, and the value associated with it is the data structure that has all the necessary information. When the same token reappears, the number of repetitions in dictionary is increased by one.

### C. *Text analysis*

During token iteration and the forming of data structures, one other process takes place, namely, text (token) analysis. Every token is analyzed before it is saved into the dictionary, in order to obtain the necessary information for further data manipulation.

Token syllabification takes place in the text analyzer component. Tokens are split by vowels for the first, rough form of syllabification. After that, multiple syllabification rules are applied on each token, in order to correct the errors in the initial division into syllables, and produce the correctly split token. In Figure 2 the control flow of the program is presented.

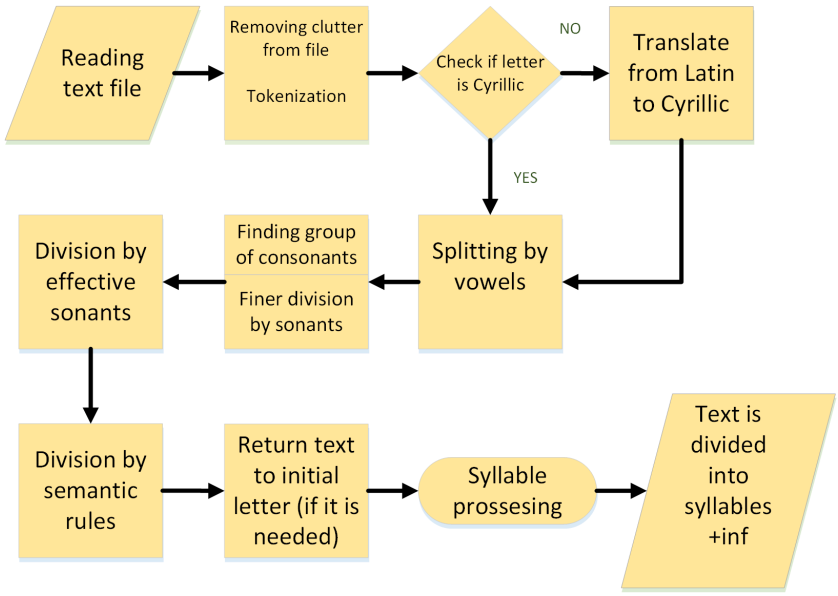
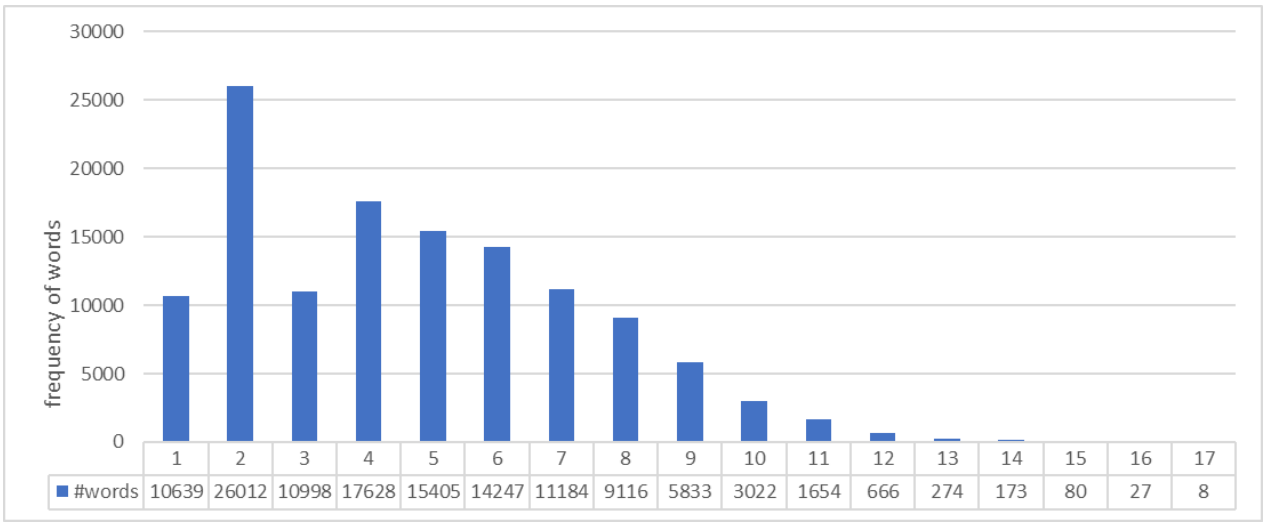



Figure 2. Program flow control

## VI. RESULTS

In this part results for the novel "The Master and Margarita will be presented. This text contains 126966 words, with a total of 22658 different word forms. Word length was from 1 to 17 letters. Words with the length of 2 were the most frequent (Figure 3).




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| #words | 10639 | 26012 | 10998 | 17628 | 15405 | 14247 | 11184 | 9116 | 5833 | 3022 | 1654 | 666 | 274 | 173 | 80 | 27 | 8 |

Figure 3. Word frequency by word length in literature work "The Master and Margarita"

The most frequent words in the text were conjunctions, particles, abbreviated form of auxiliary verb “jesam”, and prepositions (Figure 4). Also, the text contains 1337 words that appear in the text only once.

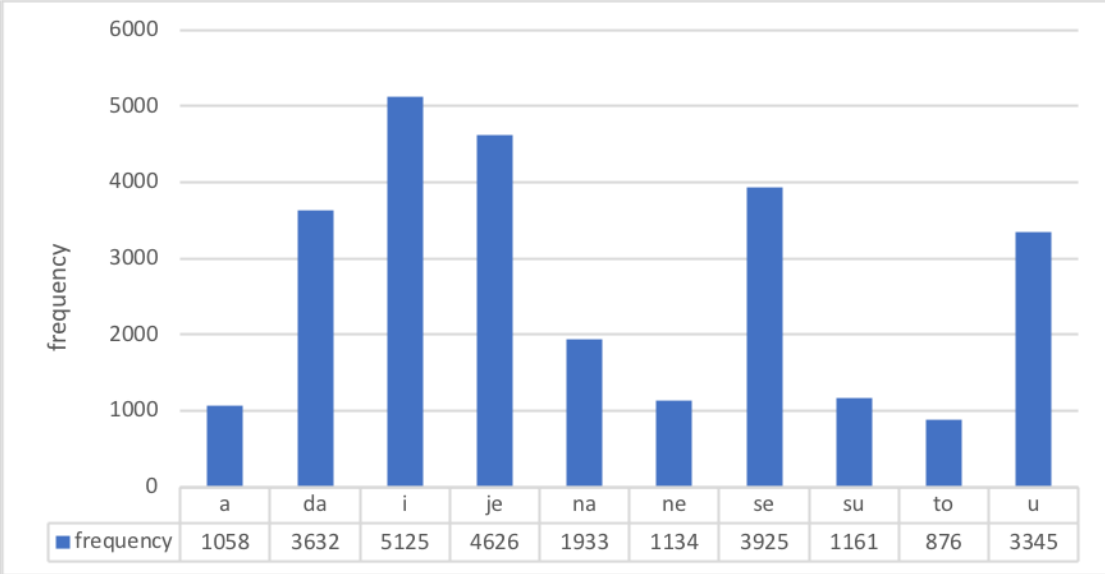



Figure 4. The most frequent words in literature work “The Master and Margarita”

Different words were divided into 72819 syllables. Words contained from one to eight syllables. The most frequent words were those with three syllables, and there were only two words with eight syllables in the text (Figure 5).

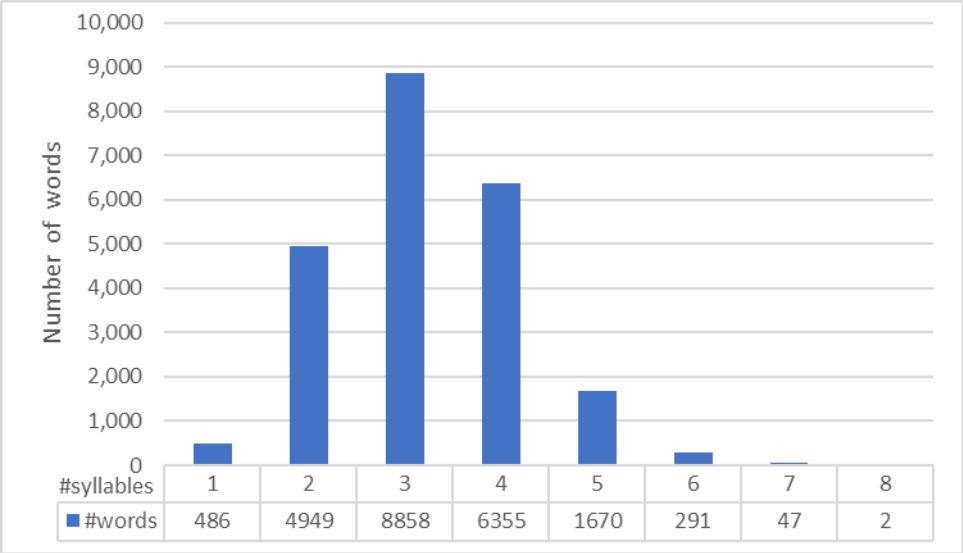



Figure 5. Number of words by number of syllables in literature work “The Master and Margarita”

## *A. Types of syllables*

As previously mentioned, several types of syllables in Serbian were recognized. Out of 72 819 syllables found in the novel “The Master and Margarita”, 71 667 belong to some of the already recognized canonical types (Table 1). There are 2448 different syllables.

| | V | VC | VCC | VCCC |
|---|---|---|---|---|
| **V** | 5132.0 | 1227.0 | 10.0 | 0 |
| **CV** | 43637.0 | 8272. | 114 | 6.0 |
| **CCV** | 10565 | 2021.0 | 31.0 | 0 |
| **CCCV** | 617.0 | 29.0 | 1 | 0 |

Table 1. Number of syllables by types. V denotes vowel and C denotes consonant.

There are also 1152 syllables which do not belong to any of the recognized types. Some of them contain syllabifical "r", and this was a group we analyzed further. Results in Table 2 show the results of this analysis, where letter "r" is treated as a vowel. Within 1152 syllables, there were 74 different syllables, 1117 of them belong to one of the recognized types of syllables, and the 35 remaining "syllables" represent some mistake (wrong syllabification, or incorrectly written word in the text).

| | V | VC | VCC |
|---|---|---|---|
| V | 22 | | |
| CV | 764 | 9 | 3 |
| CCV | 316 | 3 | |
| CCCV | | | |

Table 2. Syllables by type with syllabifical „r". V denotes „r" and C denotes consonant.

Also, we analyzed syllable lengths. The length of syllables in this text were from one to six letters. The analysis of length and frequency of syllables is depicted in Figure 6.

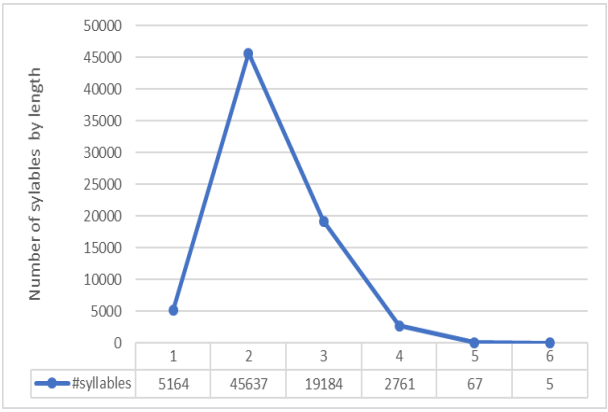



Figure 6. Syllables by their length

The most frequent syllable was syllable „o", and there were 829 syllables which appear only once in text. The most frequent syllables are depicted in Figure 7.

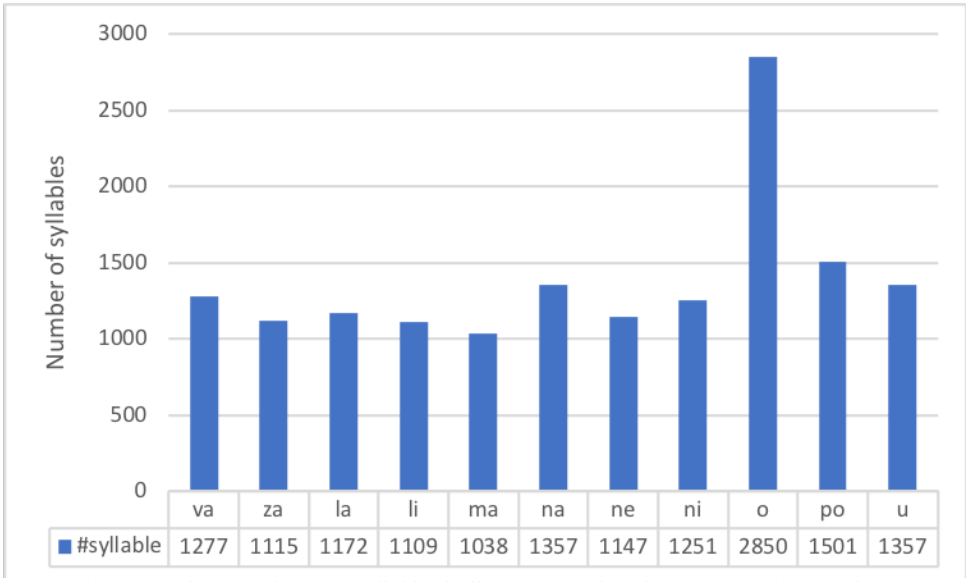



Figure 7. The most frequent syllables in literature work "The Master and Margarita"

## PRECISION

Precision of the software was tested using a ground truth of 1178 words, which contain the most frequent words of different length, with a minimum of three syllables. Out of these 1178 words, the software divided correctly 1160 into syllables (Figure 8). Among the ground truth words, errors did not appear in words of the length of 4,5,10,11,13,15 (Figure 9).

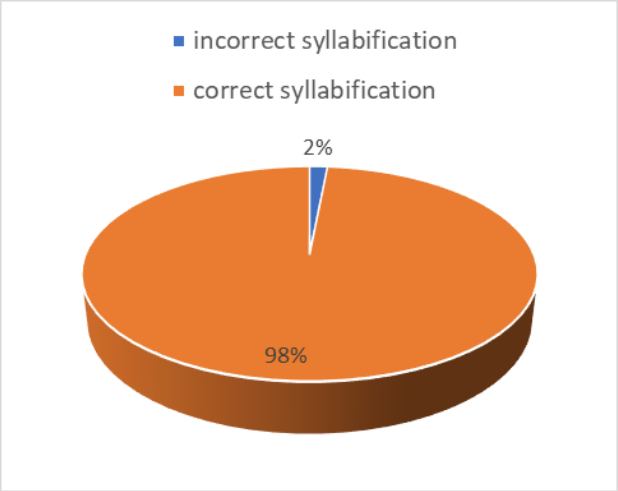


Figure 8. Precision of the software for syllabification

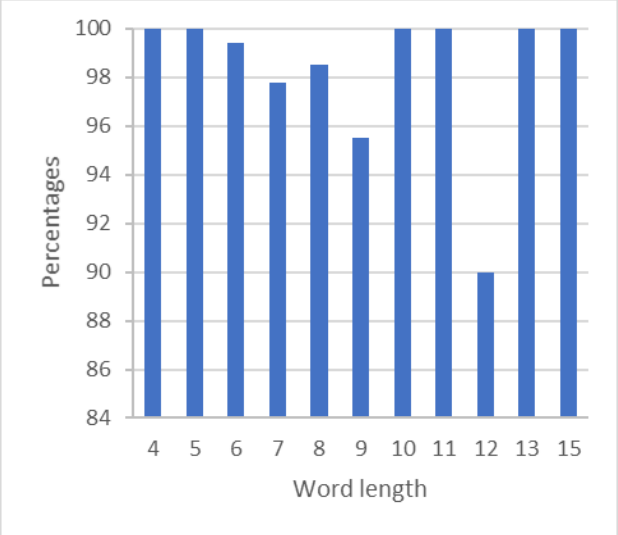


Figure 9. Precision of the software by word length

Another analysis of results of ground truth shows software precision by number of syllables in words. It can be observed that there are no errors in words from ground truth that contain 5 or 6 syllables (Figure 10).

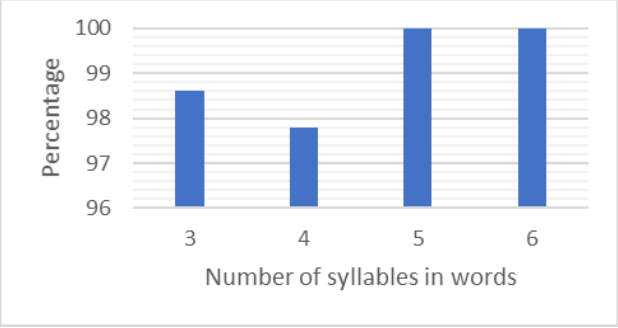


Figure 10. Software precision by number of syllables in words

## VII. Conclusion

This paper presented software for syllabification and quantitative analysis for texts in Serbian. Results are shown for the novel "The Master and Margarita". Precision of the software was tested with ground truth and results are presented in the paper.

An idea for further research is to perform a quantitative analysis for more texts in Serbian, in order to detect patterns in distribution of syllables and to extend ground truth for obtaining higher precision. Also, an idea is to upgrade the software, so that it takes into account the sonority of vowels, sonants and consonants. According to literature, such software could be applicable to several Slavic languages. After the processing of texts in different Slavic languages by this software, and following a quantitative analysis, the distribution of syllables by types will be modelled. The final part of the project would be the creation of a comprehensive mathematical model of syllables for Slavic language.

## Acknowledgment

This research was partially supported by the Serbian Ministry of Education and Science under the grants #III 47003 and the Bilateral Slovak-Serbian project: SK-SRB-2016-0020.

## References

[1] Fortson, Benjamin W. (2010), *Indo-European Language and Culture: An Introduction* (2nd ed.), Malden, Massachusetts: Blackwell.

[2] Krstev, C. (1989) *Programski sistemi za uređivanje teksta – magistarska teza,* Beograd : C. Krstev.

[3] Krstev C. (1985) *Rastavljanje reči srpskohrvatskog jezika na kraju retka*, In: Zbornik radova sa III naučnog skupa "Računarska obrada jezičkih podataka", pp. 289-301.

[4] Obradović, I., Obuljen, A., Vitas, D., Krstev, C., & Radulović, V. (2010). *Distribution of canonical syllable types in Serbian*. In: Text and language: structures, functions, interrelations : quantitative perspectives pp. 145-157.

[5] Stanojčić, Ž., & Popović, Lj. (2008). *Gramatika srpskog jezika.* Beograd: Zavod za udžbenike