A Rule-Based Syllabifier for Serbian
WHAT DID WE SET OUT TO DO?

The Goal

i) Develop a system for automatic rule-based syllabification for Serbian

ii) Provide an analysis of the outcomes to address theoretical considerations and serve as a basis for the development of future syllabifiers

iii) Present syllable distribution data for Serbian
Our Approach

- Rule-based vs. data-driven
- Existing rule descriptions:
  - Gramatika srpskoga jezika by Stanojčić and Popović (2005)
  - Kašić (2014)
  - Zec (2000)
HOW DID WE SEGMENT?

The Rules

(1) In words made up of multiple phonemes, consonants, sonorants and vowels, the syllable boundary comes after the vowel and before the consonant.

či-ta-ti [to read]

(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.

po-šta [post]

(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.

sve-tlost [light]
HOW DID WE SEGMENT?

The Rules

(1) In words made up of multiple phonemes, consonants, sonorants and vowels, the syllable boundary comes after the vowel and before the consonant.

či-ta-ti [to read]

(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.

po-šta [post]

(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.

sve-tlost [light]

tr-ča-ti [to run]

r-va-ti se [to wrestle]
The Rules

(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.

\( \text{či-ta-ti [to read]} \)

(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.

\( \text{po-šta [post]} \)

(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.

\( \text{sve-tlošt [light]} \)

A Rule-Based Sylabifier for Serbian
(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.

\[ \text{lom-ljen [broken]} \]

(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.

\[ \text{lep-tir [butterfly]} \]

(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.

\[ \text{čo-vjek [man]} \]
HOW DID WE SEGMENT?

The Rules

(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.

\( \text{lom-ljen [broken]} \)

(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.

\( \text{lep-tir [butterfly]} \)

(6) If in a cluster of two sonorants, the second position is occupied by the sonorant \( j \) from \( je \) corresponding to the \( \text{ijekavica dialect} \) to \( e \) in the \( \text{ekavica dialect} \), the syllable boundary will be before that group.

\( \text{čo-vjek [man]} \)
HOW DID WE SEGMENT?

The Rules

(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.

\[ \text{lom-ljen [broken]} \]

(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.

\[ \text{lep-tir [butterfly]} \]

(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.

\[ \text{\v{c}o-vjek [man]} \]
HOW DID WE SEGMENT?

The Rules

(7) The sonorant r can be a syllable carrier in standard Serbian when:

a. it is found medially between two consonants,  
   tr-ča-ti [to run]

b. it is found initially before a consonant,  
   r-va-ti se [to wrestle]

c. it is found after a vowel in compounds,  
   za-r-da-ti [to rust]

d. before o that is realized as an l in other members of the paradigm.  
   o-tr-o (m.) from o-tr-la (f.) [wiped]
HOW DID WE SEGMENT?

The Rules

(7) The sonorant r can be a syllable carrier in standard Serbian when:

a. it is found medially between two consonants,
   \[ \text{tr-ča-ti [to run]} \]

b. it is found initially before a consonant,
   \[ \text{r-va-ti se [to wrestle]} \]

c. it is found after a vowel in compounds,
   \[ \text{za-r-da-ti [to rust]} \]

d. before o that is realized as an l in other members of the paradigm.
   \[ \text{o-tr-o (m.) from o-tr-la (f.) [wiped]} \]
HOW DID WE SEGMENT?

The Rules

(7) The sonorant r can be a syllable carrier in standard Serbian when:

a. it is found medially between two consonants, \( \text{tr-ča-ti} \) \([\text{to run}]\)

b. it is found initially before a consonant, \( \text{r-va-ti se} \) \([\text{to wrestle}]\)

except if it is followed by the sequence \( \text{je} \).

c. it is found after a vowel in compounds, \( \text{za-r-da-ti} \) \([\text{to rust}]\)

d. before o that is realized as an l in other members of the paradigm, \( \text{o-tr-o} \) \((m.)\) from \( \text{o-tr-la} \) \((f.)\) \([\text{wiped}]\)
HOW DID WE SEGMENT?

The Rules

(8) The other two alveolar sonorants, l and n can be syllable carriers in:

a. dialectal toponyms,  
   Stš, Vlča glava, Žlne

b. foreign toponyms,  
   Vltava, Plzen

c. personal names, and in  
   English Idn or Arabic Ibn-Saud

d. the word  
   bicikl [bicycle].
The Rules

(8) The other two alveolar sonorants, l and n can be syllable carriers in:

a. dialectal toponyms, e.g., Stlp, Vlča glava, Žlne

b. foreign toponyms, e.g., Vltava, Plzen

c. personal names, and in English, e.g., Idn or Arabic Ibn-Saud

d. the word, e.g., bicikl [bicycle].
HOW DID WE SEGMENT?

The Rules

(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.

Střp, Vlča glava, Žìne,
Vltava, Plzen
English idn or Arabic Ibn-Saud
bicikl [bicycle]
(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.

Bern not Be-rn
Klajn not Kla-jn
Linkoln not Linko-ln

Stlp, Viča glava, Žlne,
Vltava, Plzen
English idn or Arabic Ibn-Saud
bicikl [bicycle]
HOW ABOUT THE DATA?

The Results

- 3,607,450 word-forms in *SrпLemKor* (Popović, 2010; Utvić, 2011)

- Most frequent syllable types: CV (62%), CCV (12%), V (11%), and CVC (9%)

- Positional distribution data for different syllable types in monosyllabic words for the initial, medial, and final positions of polysyllabic words

- Asymmetries of syllable structures occurring only in monosyllabic words and the final position of polysyllabic words: CVCC, CCVCC, VCC, CVCCC, CCCVCC, VCCC, CCVCCC, CCCCCVCC, and CCCVCCC

- Syllable nuclei statistics including their overall and positional frequencies in monosyllabic and polysyllabic words
HOW ABOUT THE DATA?

The **Results**

• ~2% noise in the data

• 6 syllable structures not found by an onset-maximization syllabifier in Croatian (Meštrović et al., 2015)

```
CCCCCVC mo-na-rhstvom
CCCCV se-rbska, ca-rstva
CCCCVC de-jstvom
CCCCCV se-rbstvo
CCCCVCC Go-idštajn, Rot-hchild, Ar-mstrong
CVCCCC cr-no-gorskg
```
Conclusions

• We developed a rule-based syllabifier for Serbian based on prescriptive rule descriptions.
• In the process, we discovered the shortcomings and inaccuracies of the existing prescriptive rule set.
• This approach still has some issues that should be resolved.
• A combination of onset maximization following (Meštrović et al., 2015) and the rule descriptions might provide an accurate capture of native speaker intuition.
References

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