Quantitative analysis of syllable properties in Serbian (and some other Slavic languages)

Ján Mačutek
with
Marija Radojičić, Biljana Lazić, Sebastijan Kaplar, Ranka Stanković, Ivan Obradović, Lívia Leššová

Department of Applied Mathematics and Statistics
Comenius University, Bratislava, Slovakia

December 10, 2018

Supported by the bilateral Slovak-Serbian grant APVV SK-SRB-2016-0021

Syllable

- no common accepted definition
  - "scholars ... found it convenient to refer to the syllable, while nobody had done much about defining it" (Haugen, The syllable in linguistic description, 1956)
  - "matters are hardly better now than they were then" (Cairns & Raimy, Handbook of Syllable, 2011, after citing Haugen)
  - "providing a precise definition of the syllable is not an easy task" (Crystal, A Dictionary of Linguistic and Phonetics, 2008)
  - "a unit of speech for which there is no satisfactory definition" (Ladefoged & Johnson, A Course in Phonetics, 2011)

Syllable structure

- nucleus – usually a vowel, sometimes a syllabic consonant
- onset – consonant(s) preceding the nucleus
- coda – consonant(s) following the nucleus
- examples:
  - vuk (wolf, Serbian)
    - v – onset, u – nucleus, k – coda
  - vlk (wolf, Slovak)
    - v – onset, l – nucleus (syllabic consonant), k – coda

Big question

- How to determine syllables, i.e., how to divide a word into syllables, if there is no established syllable definition?
- every vowel "creates" its "own" syllable, but what to do with intervocalic consonant(s)?
- Be – o – grad? Be – og – rad? Be – ogr – rad?

Two (relatively widely?) accepted syllabification principles

- maximal onset principle
  - keep syllables open, i.e., consider intervocalic consonant(s) as onsets so that a syllable ends with a vowel. . . but do not violate a sonority hierarchy
- sonority hierarchy principle
  - syllable nucleus constitutes a sonority peak of a syllable, i.e., sonority decreases towards both edges of a syllable
  - vowels > approximants > nasals > fricatives > affricates > stops

OK. . . but. . .

- even if one accepts these two principles, there remain some problems
- some words in some languages have syllables which are not possible to reconcile with the two principles
- example: rty (lips, Czech) – r is more sonorous that t, but this word is a monosyllable, so there are no possibilities to divide it

Our approach

- with respect to sonority, we distinguish only two classes of consonants (sonorants and others)
- we slightly modify the sonority hierarchy principle (we allow sonority plateaus, i.e. sequences of consonants with the same sonority)
- we keep syllables open unless they violate our version of sonority principle
- the list of sonorous consonants is language-specific, we take it from established linguistic sources

Bilateral Slovak-Serbian project

- official aim of the project - quantitative analysis of syllables in Russian, Serbian, and Slovak
- unofficially – more (perhaps all) Slavic languages
- state of the art – syllabification of Serbian, Croatian, and Ukrainian ready
- Serbian and Croatian – no diphthongs, syllabic consonant – r between two other consonants
- Ukrainian – no diphthongs, no syllabic consonants
- language material – parallel language corpus (Russian novel "Kak zakaljalas' stalj" – "How the steel was tempered" and its translations into 11 other Slavic languages) created by Emmerich Kelih

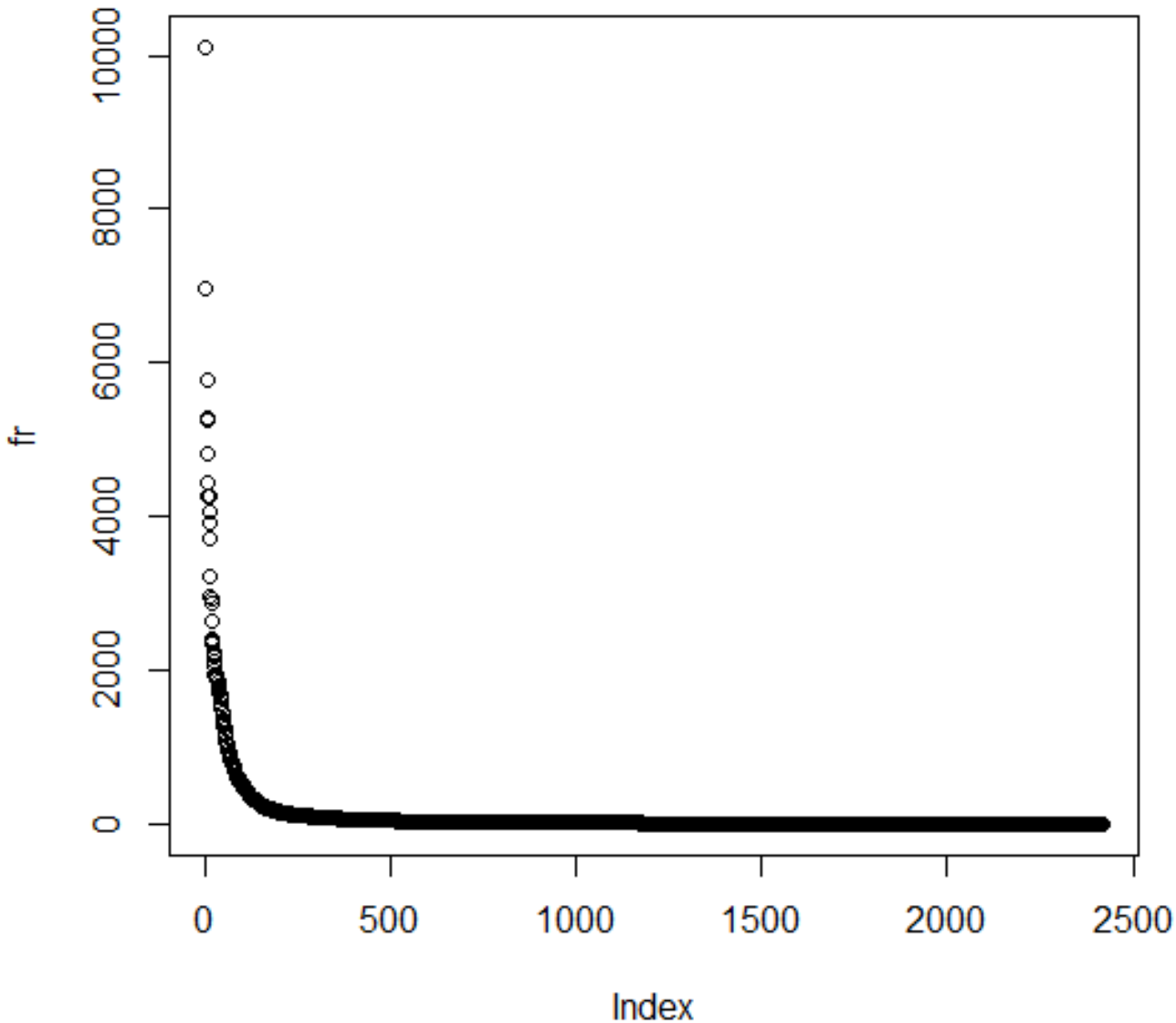
Some results

- rank – frequency distribution of syllables
- distribution of syllable length
- Strauss, Fan, Altmann (2008) - similar mathematical models as those for words (Zipf- and Poisson-like distributions)?
- typology of Slavic languages based on syllables frequencies?

Syllable frequencies in Serbian

1	10103	o
2	6970	je
3	5778	u
4	5291	na
5	5248	da
⋮		
2419	1	

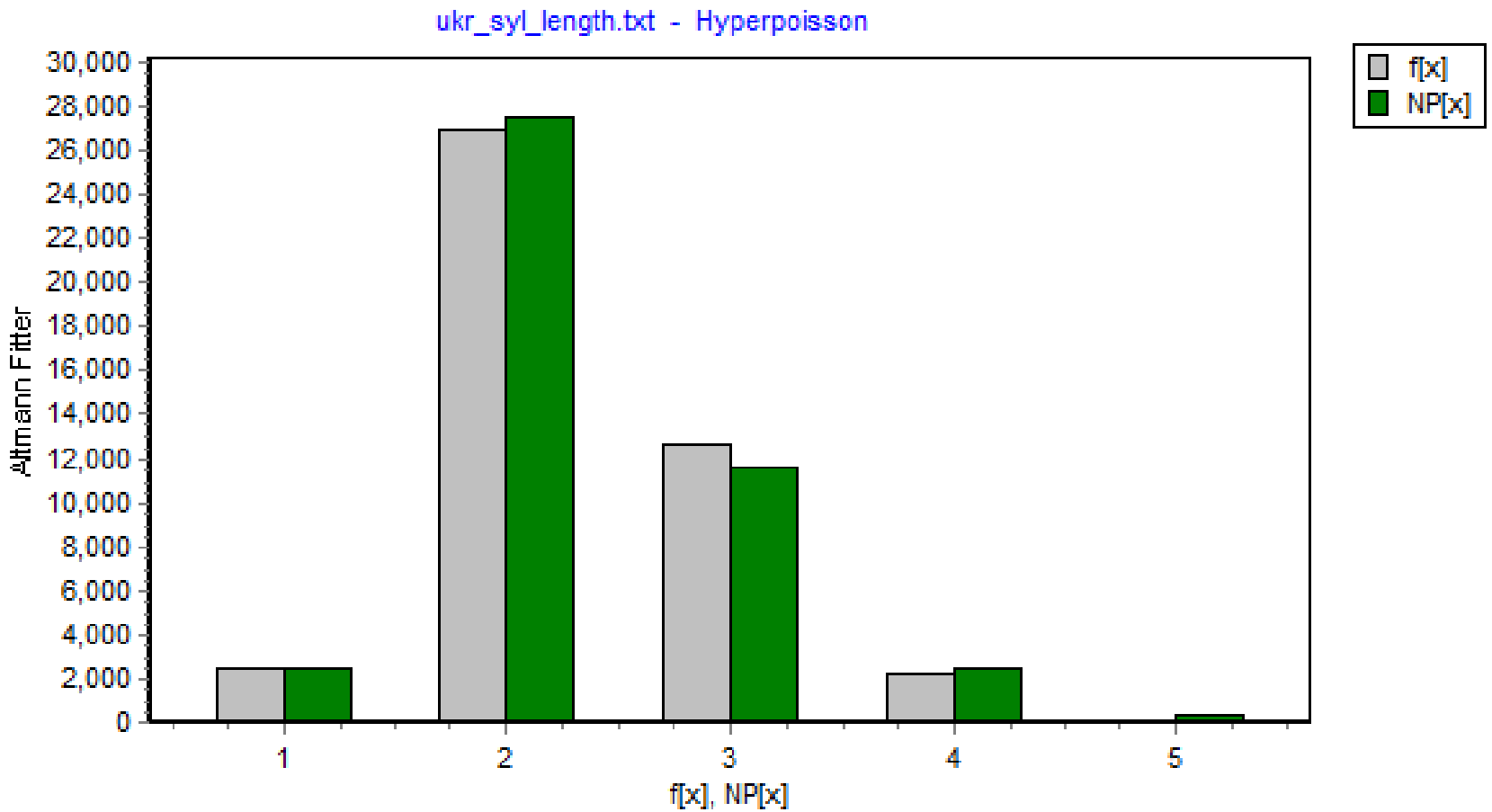
Syllable frequencies in Serbian



Distribution of syllable length

1	23505
2	153939
3	54554
4	6982
5	236
6	3

Syllable length



Least effort principle. . . in other words, we are lazy

- words: the higher frequency of a word, the shorter it is
- is it true also for syllables?

Relation between frequency and length of syllables

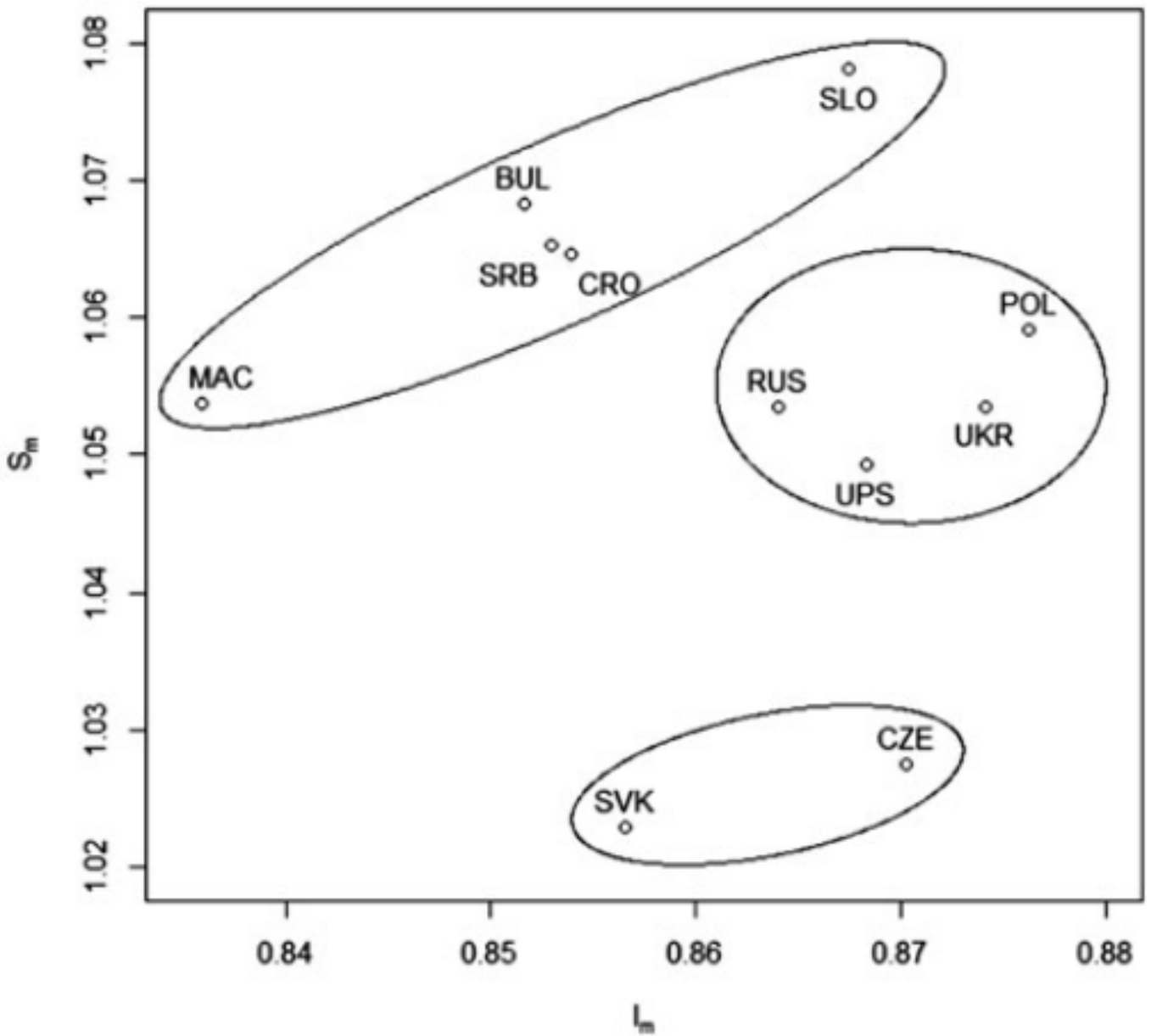
length and mean frequency

1	3917.5
2	785.4
3	37.5
4	9.7
5	6.4
6	1

Syllables pooled so that there are roughly 30000 of them in each group

1.6, 1.86, 2, 2, 2.13, 2.23, 2.84, 3.33

Data-based typology of Slavic languages (graphemes)



Data-based typology of Slavic languages (graphemes)

- Ord graph – uses ratios of mean, variance and skewness
- our modification (Koščová, Mačutek, Kelih 2016, JQL 23, 177-190) - these characteristics are replaced with indices of qualitative variation

Conclusions

- start of a systematic investigation of syllables in Slavic languages rank-frequency distribution – similar to word length syllable length distribution – similar to word length relation between frequency and length – similar to the one for words
- studies on typology based on syllable frequencies opened

Hvala na pažnji!

Thank you for your attention!