Универзитет у Београду

Quantitative analysis of syllable properties in some Slavic languages

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

Supported by the bilateral Slovak-Serbian grant APVV SK-SRB-2016-0021 (all authors) and by the grant VEGA

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)?
- Wro – cław? Wroc – ław? Wrocł – aw?

Two (relatively widely?) accepted syllabifiction 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

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 three 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 (minor issues with the Serbian results)
- 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
- similar mathematical models as those for words (Zipf- and Poisson-like distributions)?
- some language-specific issues
- typology of Slavic languages based on syllables frequencies?

Rank-frequency distribution of syllables

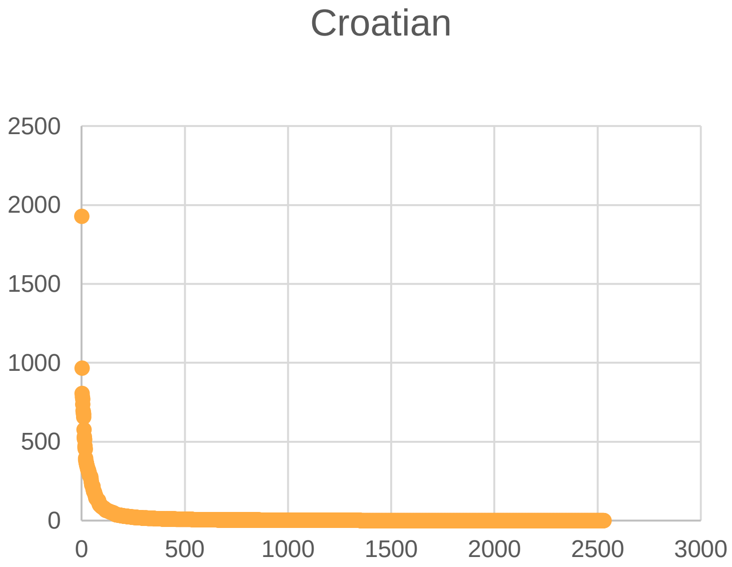
Croatian (30 graphemes), N = 43865

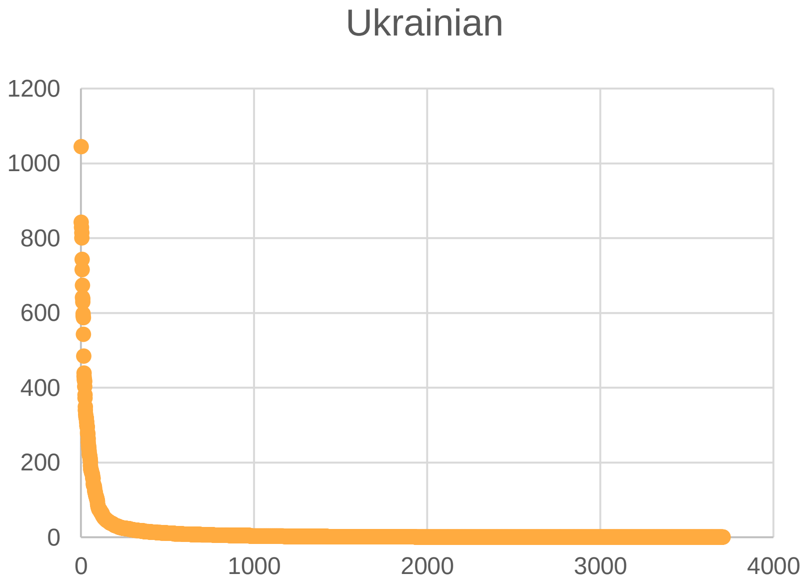
1	1928
2	967
3	806
4	784
5	769
…	
2531	1

Ukrainian (34 graphemes), N = 47064

1	1045
2	843
3	829
4	815
5	801
…	
3709	1

Rank – frequency distributions - figures





Rank – frequency distribution - models

- no discrete model achieves an acceptable fit
- continuous models
- $y = ae^{-c}$
  - CRO: a=930.81, c=0.0296, $R^2 = 0.8974$
  - UKR: a=817.70, c=0.0258, $R^2 = 0.9671$
- Zipf-like functions do not model Croatian data well
- “too high” first frequency is the reason

Distribution of syllable length

Croatian

1	3463
2	25080
3	11737
4	2424
5	188

hyperpoisson distribution
a=0.4632
b=0.0640
C=0.0041

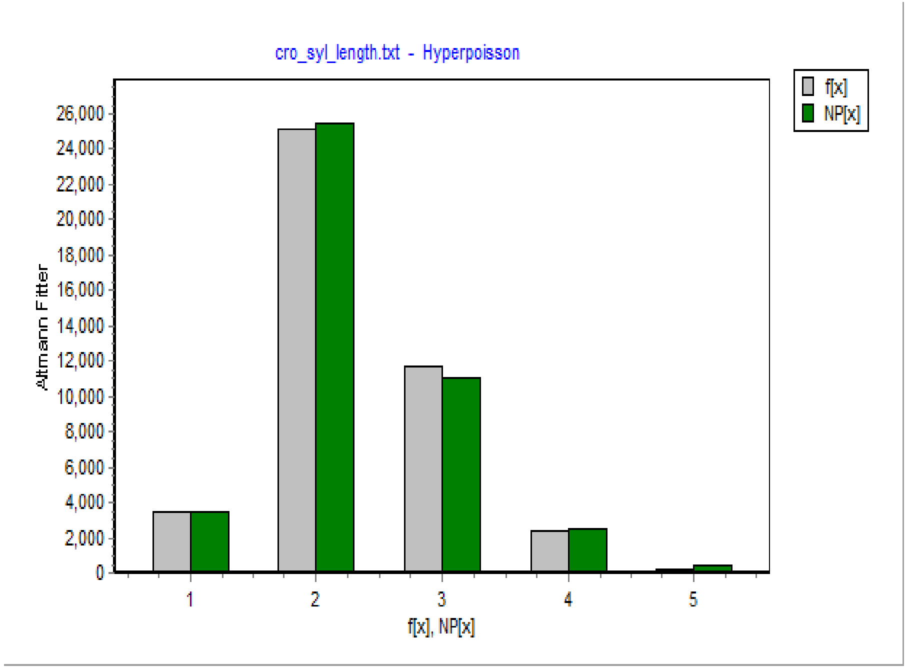
Ukrainian

1	2427
2	26961
3	12688
4	2183
5	132

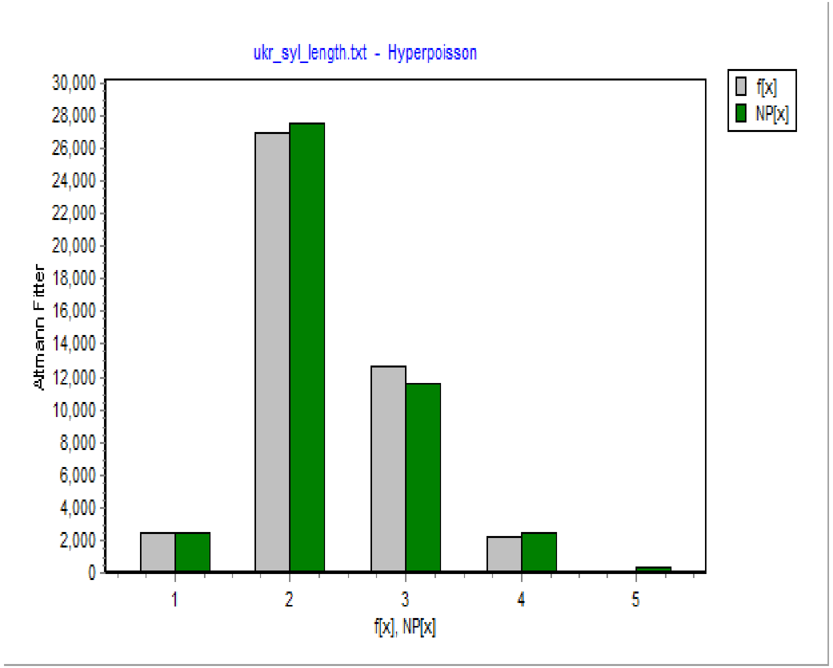
hyperpoisson distribution
a=0.4370
b=0.0393
C=0.0075

Distribution of syllable length - figures

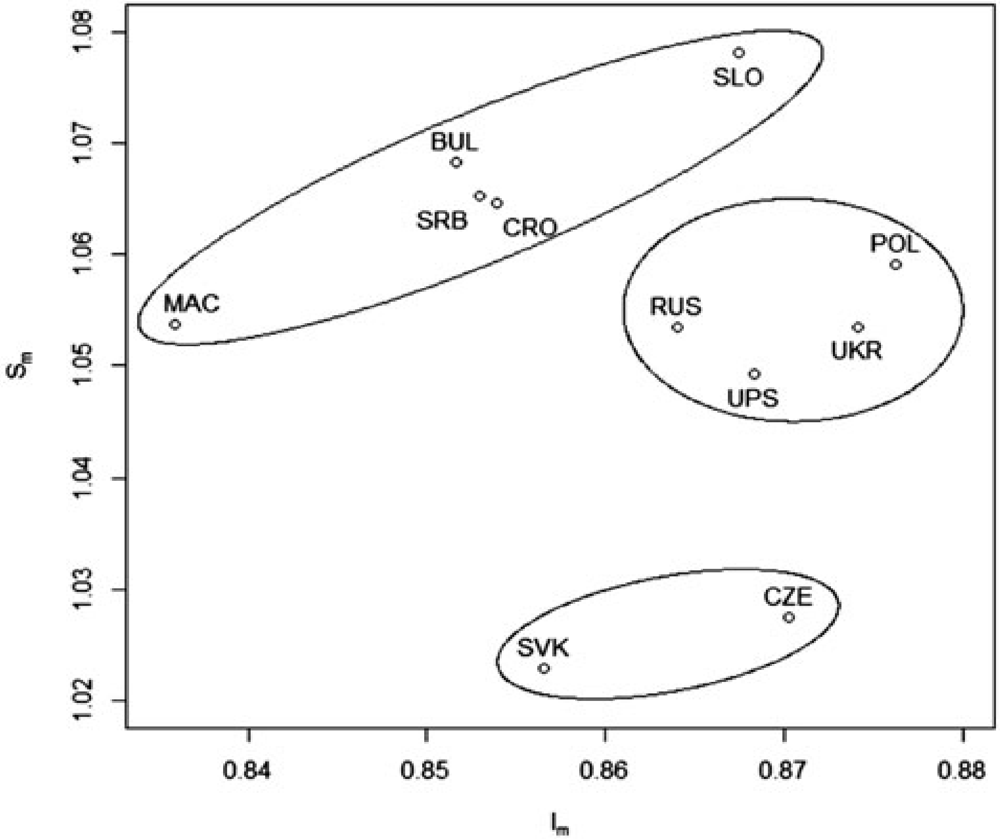
Croatian



Ukrainian



Data-based typology of Slavic languages (graphemes)



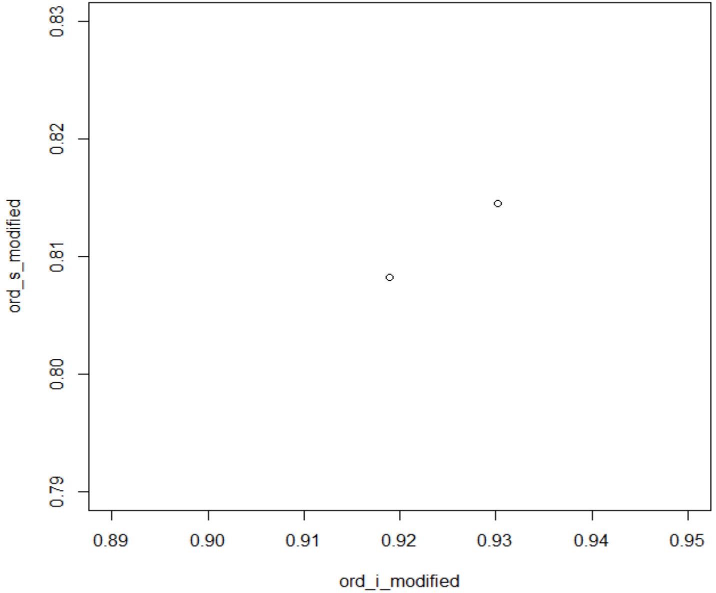
Data-based typology of Slavic languages (graphemes)

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

Data-based typology of Slavic languages (syllables)?

- Coordinations on modified Ord graph
  - CRO: 0.9189, 0.8082
  - UKR: 0.9302, 0.8145

CRO left, UKR right



Conclusions

- start of a systematic investigation of syllables in Slavic languages
- rank-frequency distribution – unclear
- syllable length distribution – similar to word length
- studies on typology based on syllable frequencies opened

Dziękuję za uwagę!

Thank you for your attention!