universität wien

Emmerich Kelih

(Vienna)

Institut für Slawistik

On the dependency of word length on text length

Empirical results from Russian and Bulgarian parallel texts

- Institute for Slavic Studies, University of Vienna
- http://www-gewi.uni-graz.at/quanta/
- https//slawistik.univie.ac.at/mitarbeiter/kelih-emmerich// [emmerich.kelih@univie.ac.at]

Aims and objectives of the study:

Is there a systematic relation between word length and text length ?

1. word length studies: state of the art
2. text length as a problem of QL

2.1. TTR in Slavic parallel texts

2.2. linguistic interpretation of TTR

3. systematic relation between word length and vocabulary size

3.1. empirical evidence from Russian and Bulgarian

4. conclusion and major results

State of the art

word length studies:

- discussion of appropriate word definitions
- question of relevant units of measurements
- Menzeraths law (interrelation of word and syllable length)
- study of influence factors: text and discourse type

  genre and authorship
- modelling (discrete probability distributions)

→ what about <u>text length</u>?

text length as a problem of QL:

•min., max. or "optimal" text length for word length studies?

•vocabulary size *V* (number of word types) and text length (N)

•different approaches of modelling the Type-Token-Ratio (TTR)

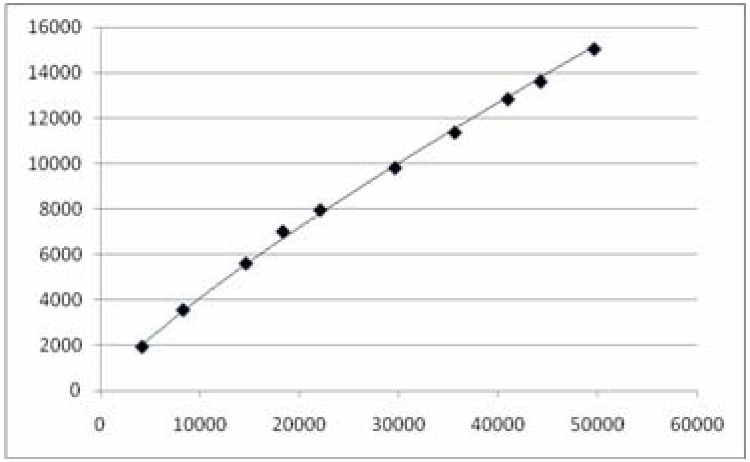
$V = \alpha\sqrt{N}$	Guiraud (1954)
$V = \alpha N^{\beta}$	Herdan (1964)
$V = \frac{\alpha N}{\beta + N}$	Tuldava (1995)
$V = \frac{\alpha N}{1 - \beta + \beta N}$	Köhler/Martináková (1998)

→ ***V*** is a nonlinear function of text length ***N***

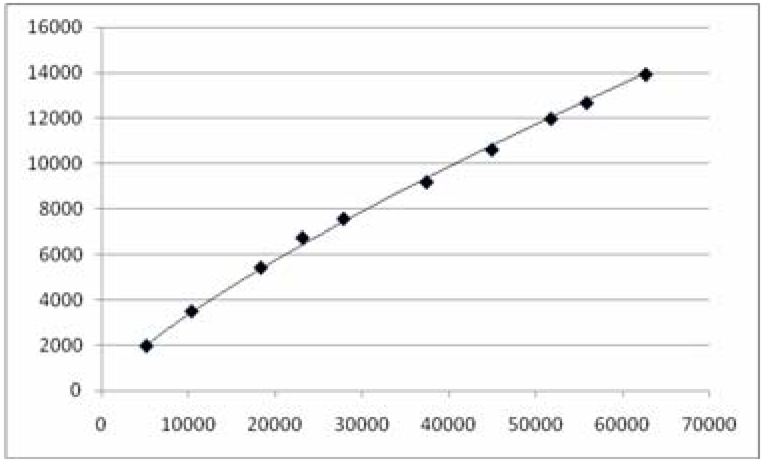
→ TTR cannot be interpreted as indicator of vocabulary richness and

stylistic diversification !

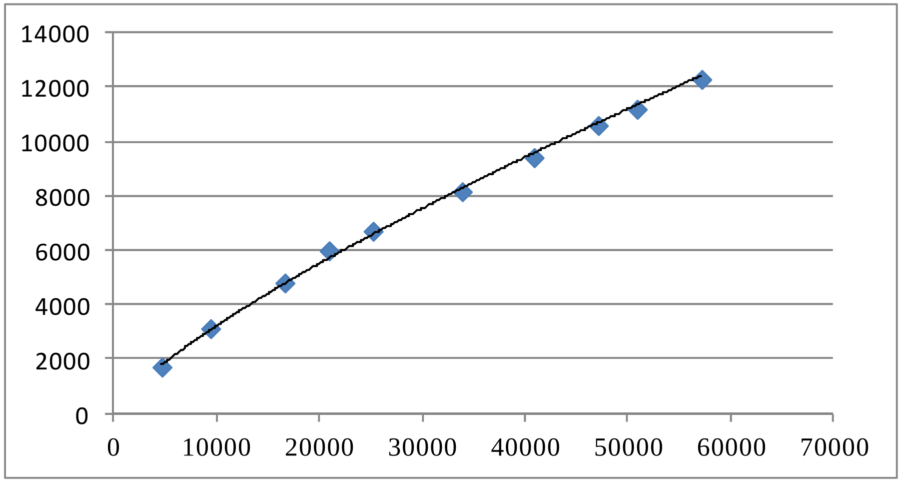
TTR in Slavic parallel texts: “How the Steel was tempered”



Russian: *N* (49672) vs. *V (15053)*



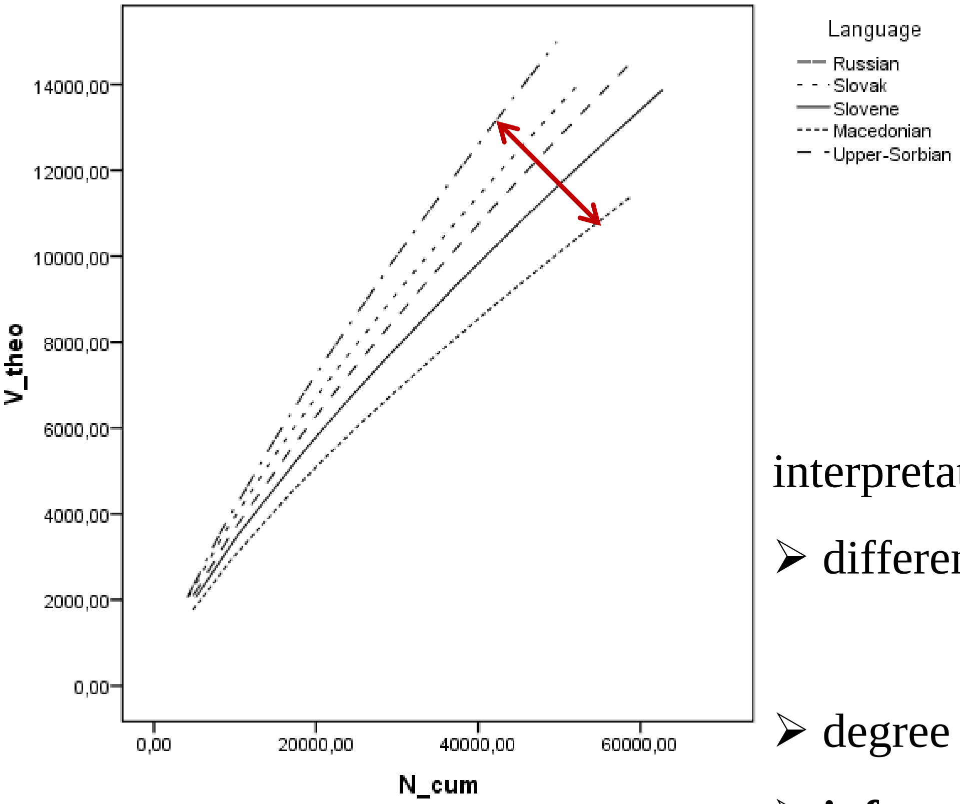
Slovenian: *N* (62646) vs. *V (13940)*



Bulgarian: *N* (57164) vs. *V (12303)*

Language	**N**	**V**	**V/N**
Russian	49672	15053	3,30
Slovenian	62646	13940	4,49
Bulgarian	57164	12303	4,40

TTR (theoretical values) in different Slavic languages (same text)



interpretation of TTR:

- different increase of TTR: morphological "richness"
- degree of analytism and synthetism
- **information flow** within one text

Linguistic interpretation of TTR:

→ the more often a token is repeated, the slower the spread of information.

→ successive introduction of new word form types.

→ a **systematic increase** of synsemantic and autosemantic word forms.

autosemantic word forms: give new semantic information

synsemantic word forms: organisation of the grammatical structure of the text

Hence:

→ With increasing text length (number of types) a successive increase of word length (measured in the number of graphemes/syllables is observable.

Empirical evidence from parallel text corpora:

“Master and Margarita” (1928-1940) by M. Bulgakov

Languages available:

Russian
Ukrainian
Belorussian

Slovak
Czech
Polish

Croatian
Serbian (1)
Serbian (2)
Bulgarian
Macedonian

German

→ Analysis of

1. **Russian** (original): synthetic language
2. **Bulgarian** (translation): analytic language

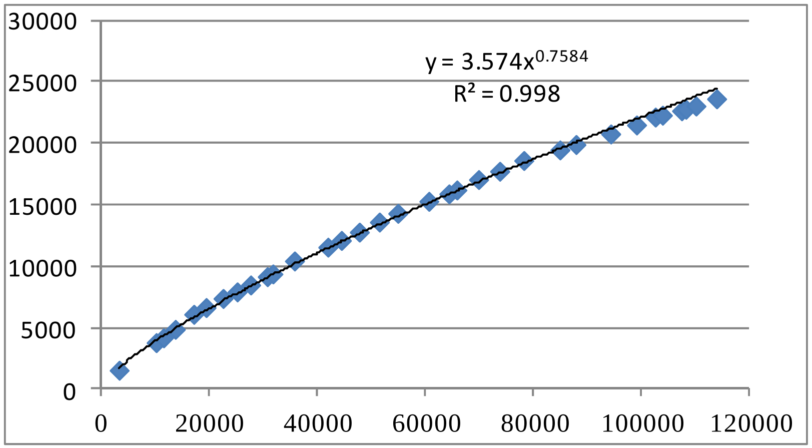
Procedure: Step (1)

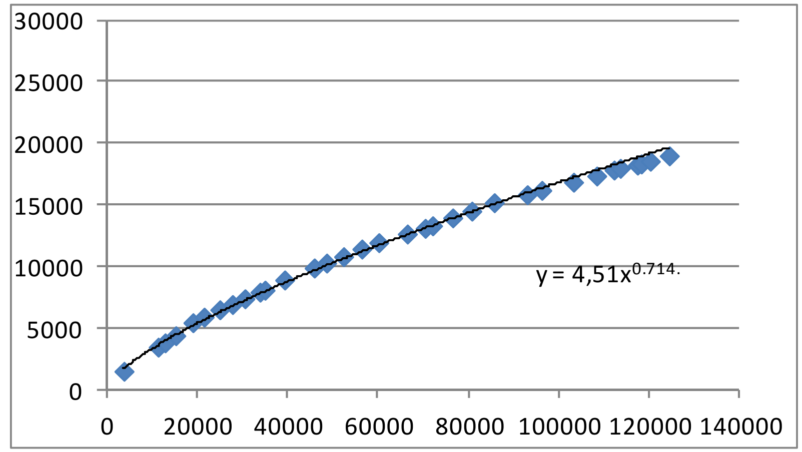
1.text-preprocessing of 33 chapters

2.determination of the number of types and tokens

3.cumulation of the single chapters: 1; 1+2; 1+2+3, 1 …+ 33 = whole text

4.modelling TTR in Russian and Bulgarian





Language	**N**	**V**	**V/N**
Russian	113748	23640	4,81
Bulgarian	124380	18996	6,55

Russian: $V = 3.57 * N^{0,58}$

Bulgarian: $V = 4.52 * N^{0,72}$

Procedure: Step (2)

1. Determination of word length (stepwise chapter per chapter)

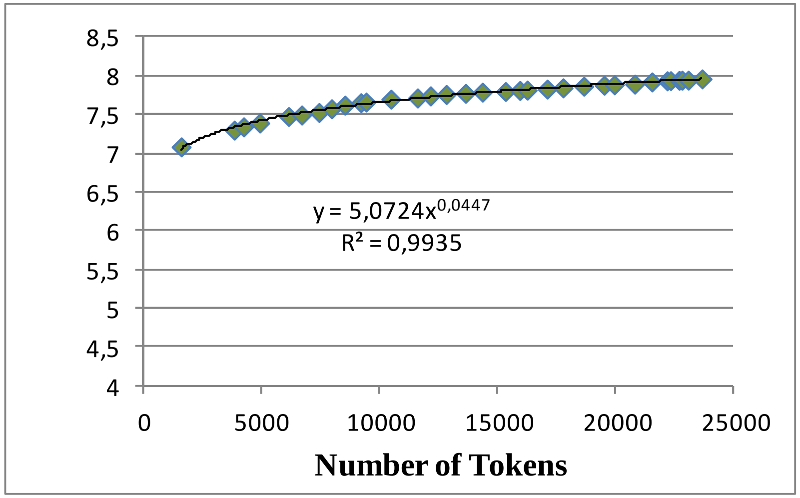
    1.1. number of syllables

    1.2. number of graphemes

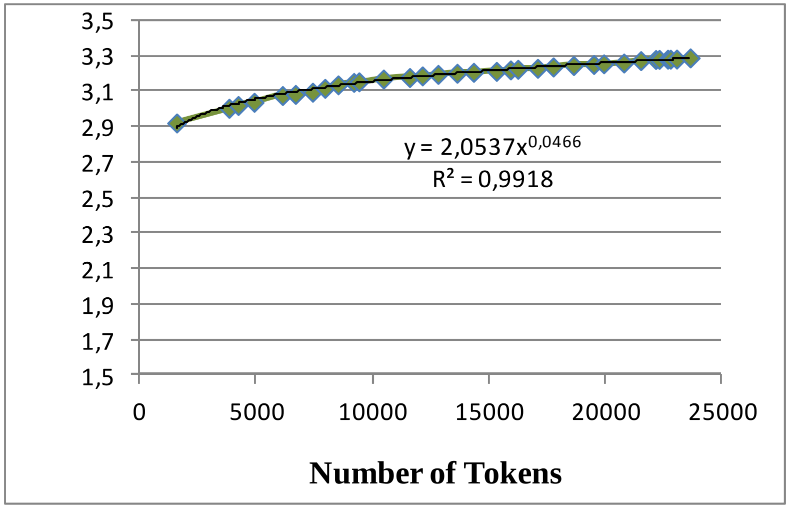
2. What about the interrelation between number of types (V) and word length?

Russian

word length (graphemes)



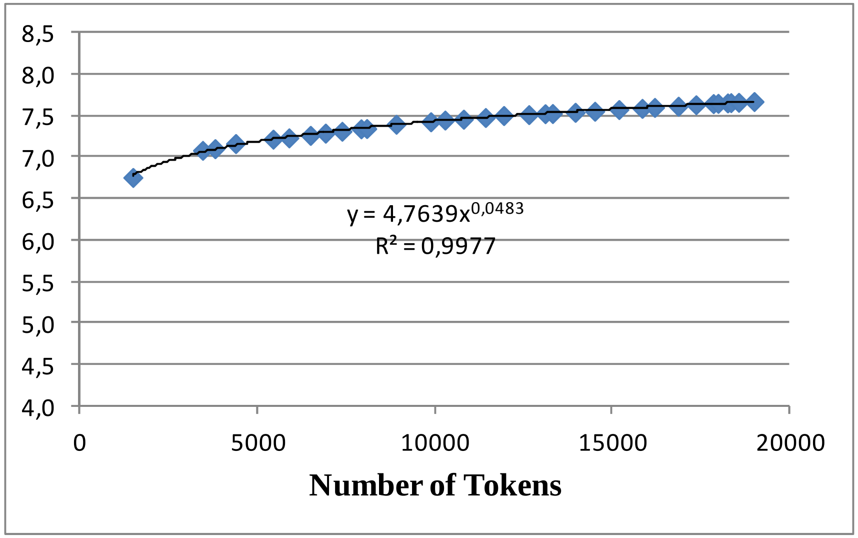
word length (syllables)



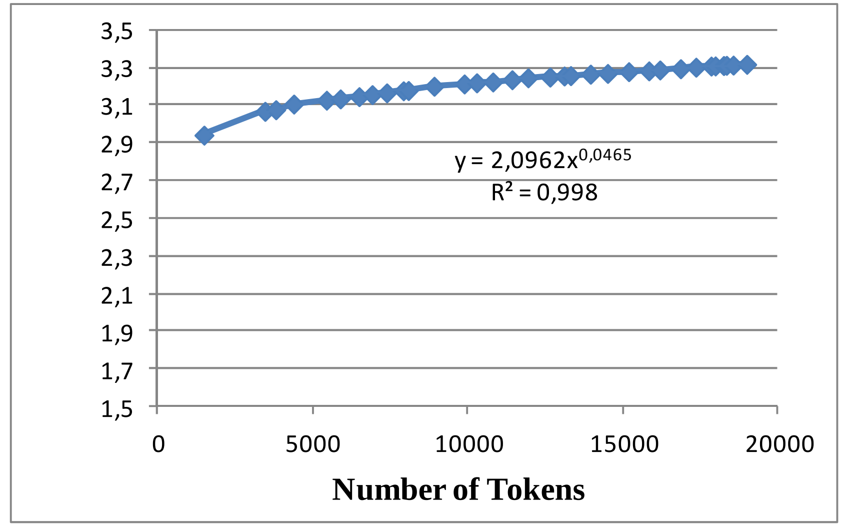
→ **both cases:** systematic increase of word length with increasing text length (V)!

Bulgarian

word length (graphemes)

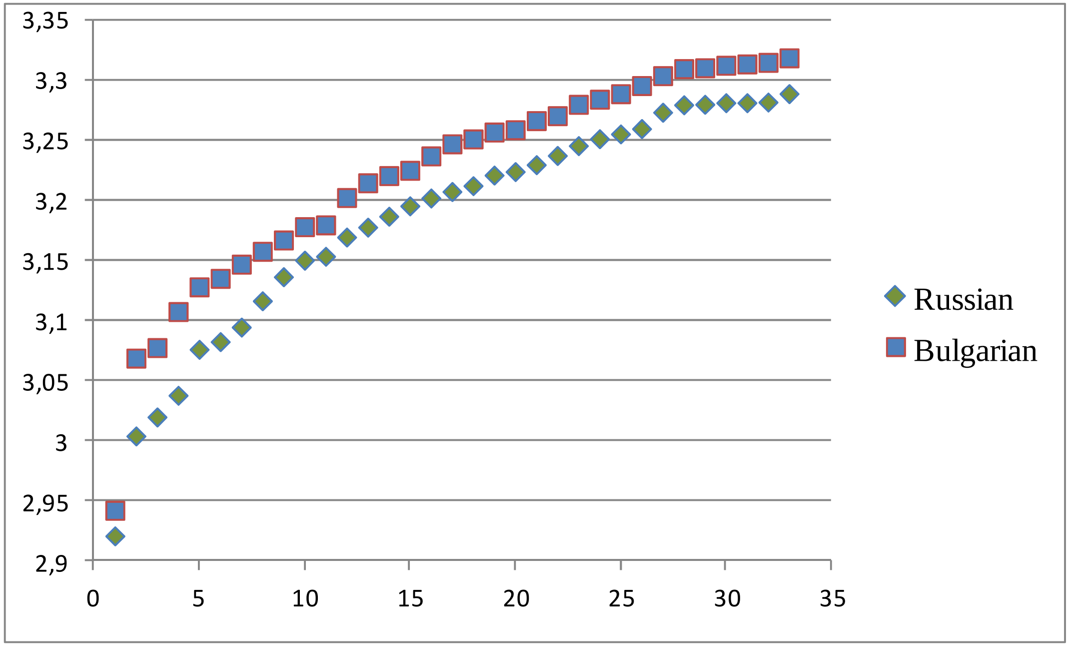


word length (syllables)



→ **again:** systematic increase of word length with increasing text length (V)!

Comparision: Russian and Bulgarian



Language	**Wol (Syllables)**	**R^2**
Russian	Wol (syl) = $2.054*V^{0.047}$	0.99
Bulgarian	Wol (syl) = $2.097*V^{0.047}$	0.99

- same model for the interrelation between *V* and *Wol (syl.)* !
- **no** significant differences between Russian and Bulgarian

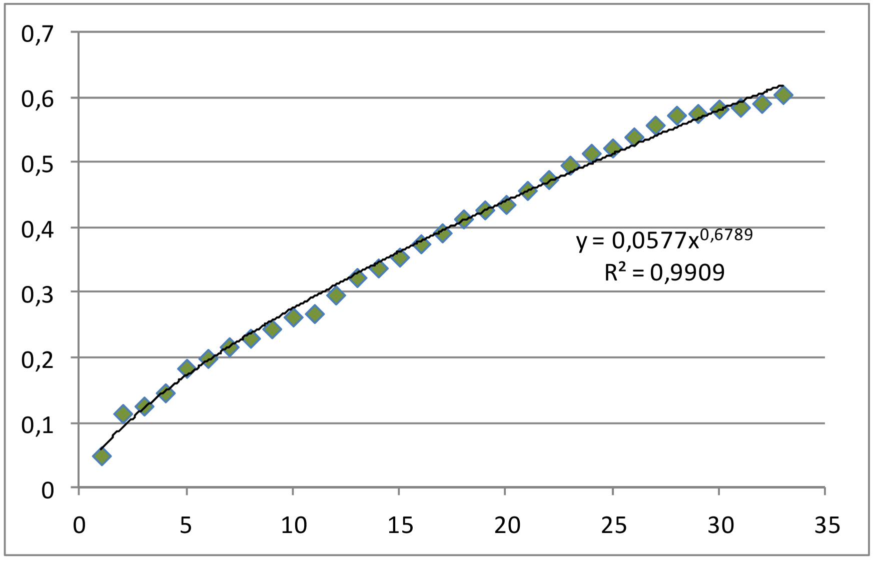
explanation of the interrelation between *V* and *Wol*

→ successive introduction of new word form types.

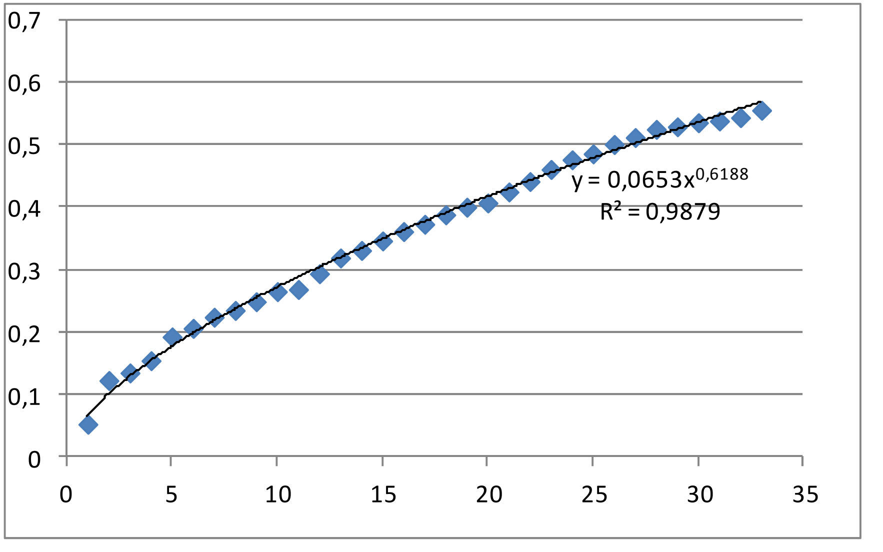
→ systematic increase of synsemantic and autosemantic word forms.

→ frequency behaviour of **Hapax Legomena (HL)**

increase of *HL* in cumulated chapters:



Russian

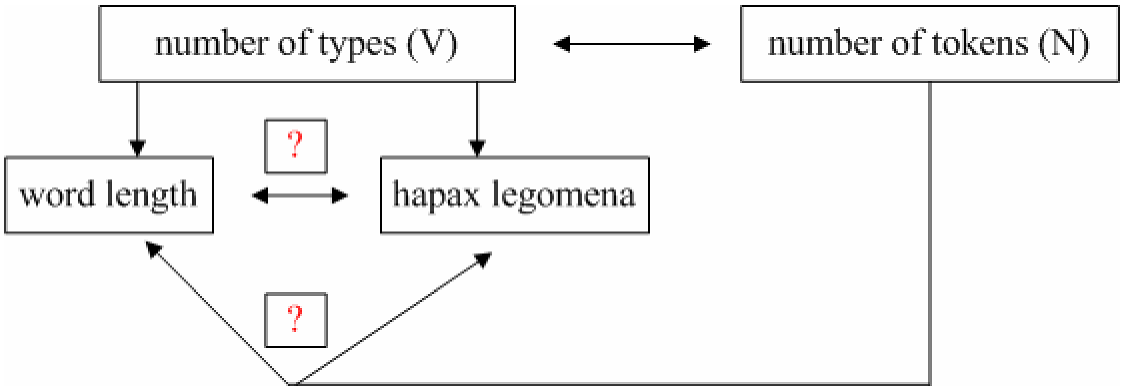


Bulgarian

preliminary results:

➢systematic behaviour of different text/language characteristics

- number of types (*V* vocabulary)
- number of tokens (*N* text length)
- word length
- increase of hapax legomena



major results:

- ➢vocabulary size and text length are systematically interrelated
- ➢hapax legomena increase with increasing vocabulary size
- ➢word length is influenced by text length (V) in general

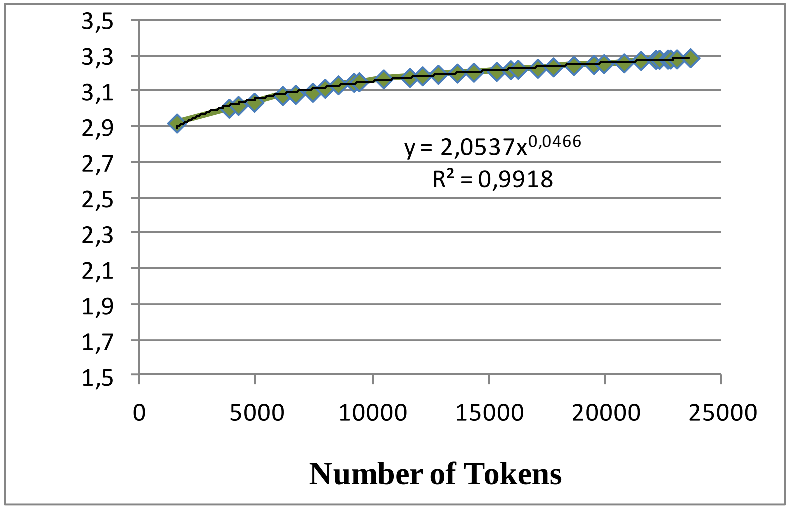
consequences:

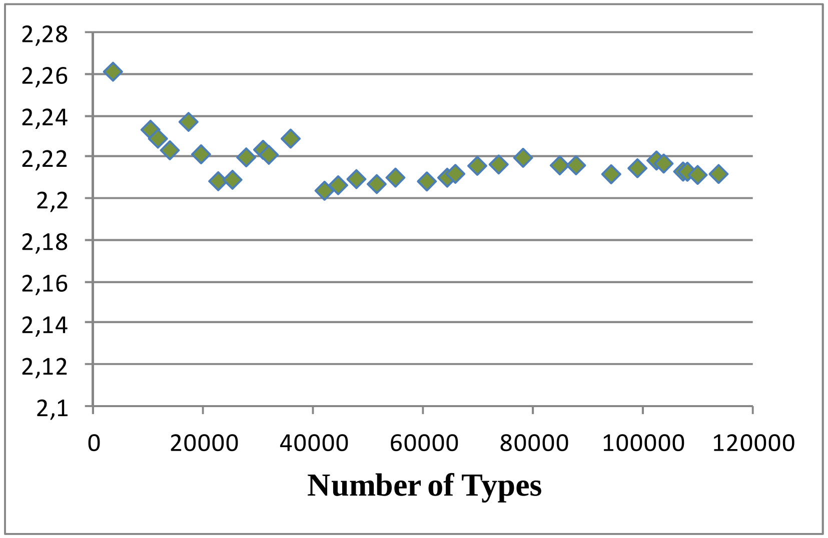
- ➢ importance of the factor “text length” for word length studies
- ➢ comparison of texts/languages?

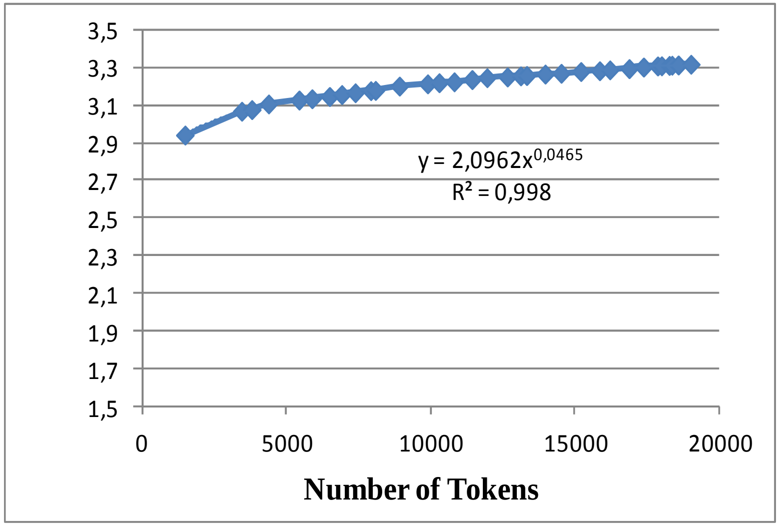
→only in texts of the “same” length

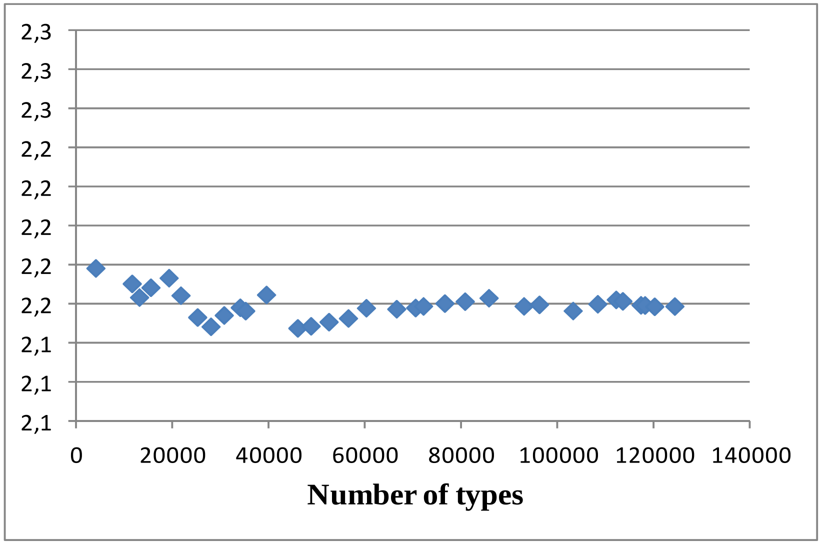
→parallel texts as reliable source of TTR and WOL studies

what about word length and number of tokens ?









→ Decrease of word length with increasing number of tokens!