Emmerich Kelih

(Graz)

Grapheme frequencies and word length in Slovene and Serbian: Interrelations and empirical findings

- Institut für Slawistik, Universität Graz
- http://www-gewi.uni-graz.at/quanta/ [Graz-project on quantitative text analysis]
- http://www.uni-graz.at/emmerich.kelih/ [emmerich.kelih@uni-graz.at]

Aims and objectives of the study:

Is there a systematic relation between grapheme frequencies and word length?

1. modelling grapheme frequencies: State of the art
2. importance of Menzerath’s law (word length vs. syllable length)
3. impact on relative frequencies of consonants/vowels
4. parameter interpretation of grapheme models
5. conclusion

Modelling grapheme rank frequencies of Slavic languages

I. Discrete probability models ?

negative Hypergeometric distribution

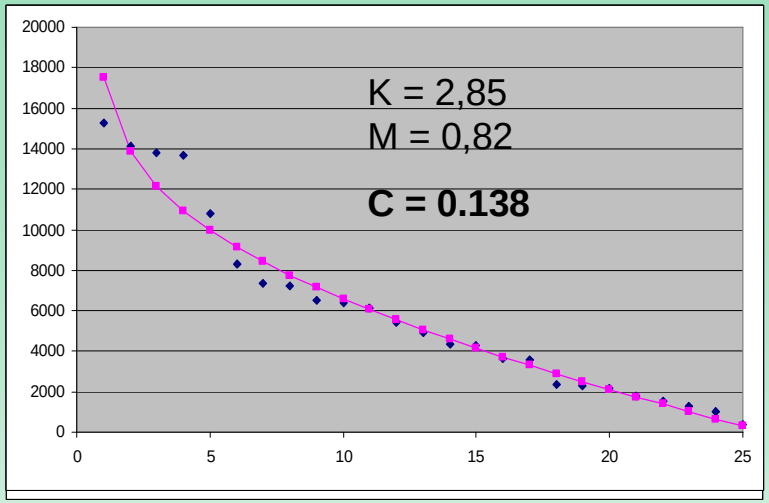
$$P_x = \frac{\binom{M+x-1}{x}\binom{K-M+n-x-1}{n-x}}{\binom{K+n-1}{n}}$$

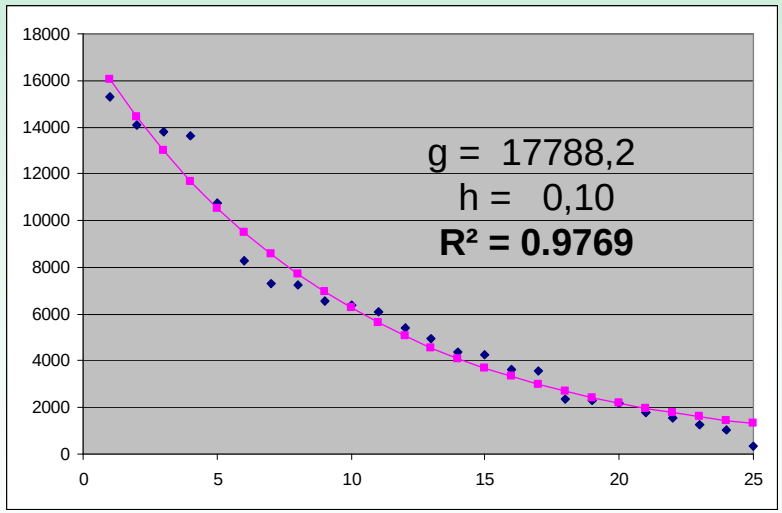
II. Continuous models ?

Re-Formulation of rank frequency models by Popescu/Altmann/Köhler 2009

„strata-model“

$$y = ge^{-hx} + me^{-nx} \ldots \qquad y = ge^{-hx}$$





Influence factors

1. Inventory size (under examination)
2. Repeat-Rate and Entropy
3. Parameter interpretation?

Starting point: Importance of Menzerath's law (ML):

"The relative number of sounds in the syllable decreases as the number of syllables in the word increases, or said in a different way: the more syllables in a word the shorter (relatively) they are."

"The longer the word, the shorter the syllables"

1. word length (WOL): number of syllables

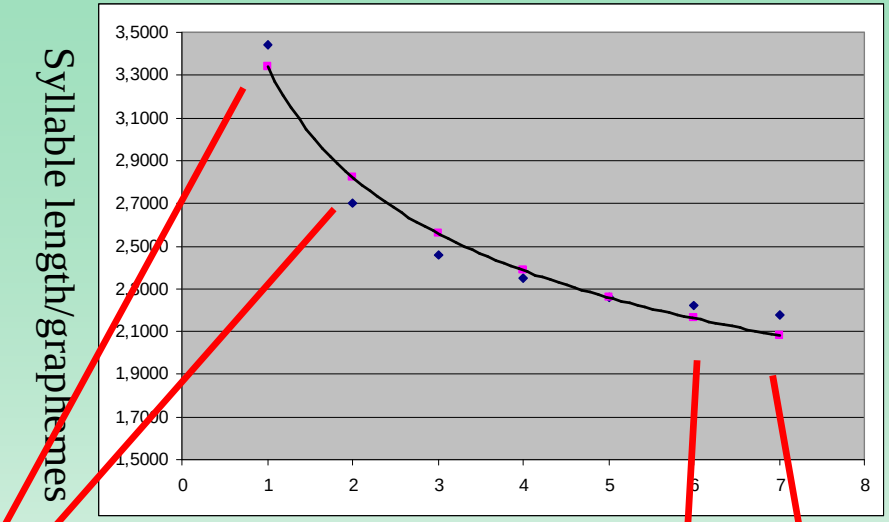
2. syllable length (SYL): number graphemes/phonemes

3. $SYL = a*WOL^{-b}$

- ML commonly accepted and empirically verified in many languages

- further implications of ML on grapheme frequencies, in particular on relative consonant/vowel frequencies?

The greater the word length, the „simpler“ the syllable structure of a word.



CV	CCCVC
CVC	CVCC
CCV	CVCCC
CCVC	CCVCC
V	VCC
CCCV	VC

CV	VC
CVC	CCVC
CCV	CCCV
V	CCVC

Hypotheses:

1. The „simpler“ the syllable structure, the higher the relative frequency of vowels.

2. The “simpler” the syllable structure, the lower the relative frequency of consonants.

Empirical analysis: Slovene Material

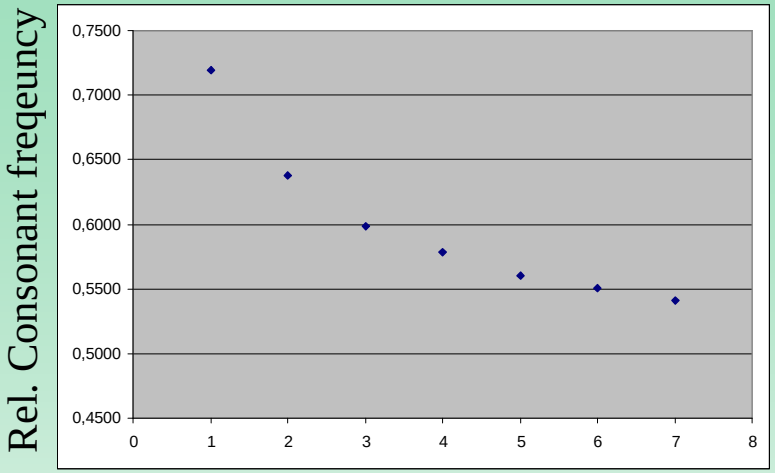
text sort	**number of texts**	**word form types**
dissertations	5 chapters	6144
private letters	30 (Ivan Cankar)	5182
sermons	32	7977
dramas	42 acts (Drago Jančar)	5616
complete corpus		24919

1. word length (WOL): number of syllables
2. word form types
3. <a, e, i, o, u> and syllabic <r> in some positions are counted as “vocalic graphemes”
4. syllable length (SYL): number of graphemes

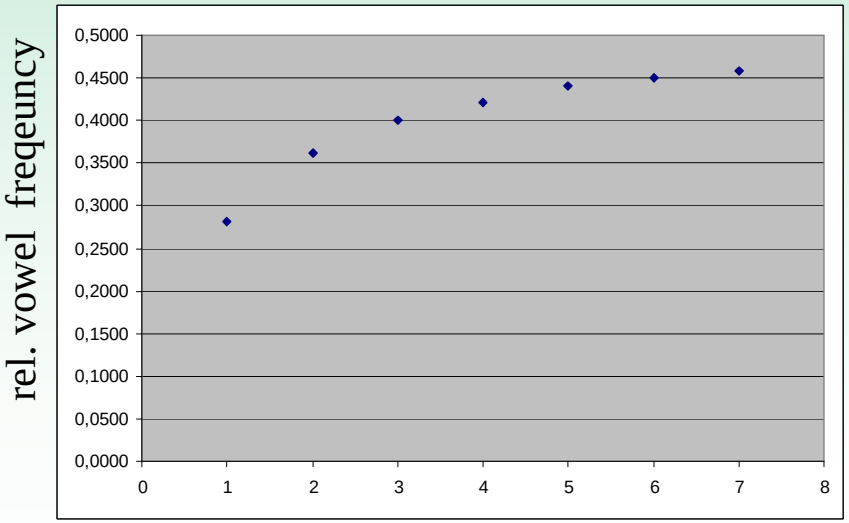
"The higher the word length (WOL), the smaller the relative consonant frequency (RCF)"

Complete Corpus

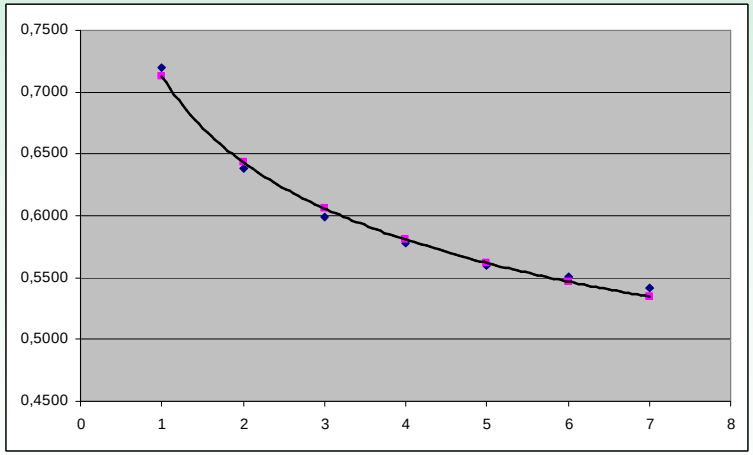
WOL/Syl	RCF
1	0,7193
2	0,6380
3	0,5989
4	0,5780
5	0,5602
6	0,5506
7	0,5412



Word length/syllables



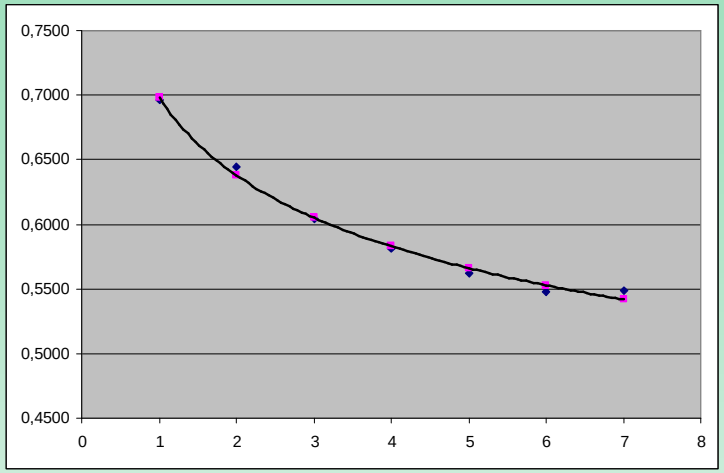
Word length/syllables



model: $RCF = 0.71*WOL^{-0.14}$

$R^2 = 0.99$

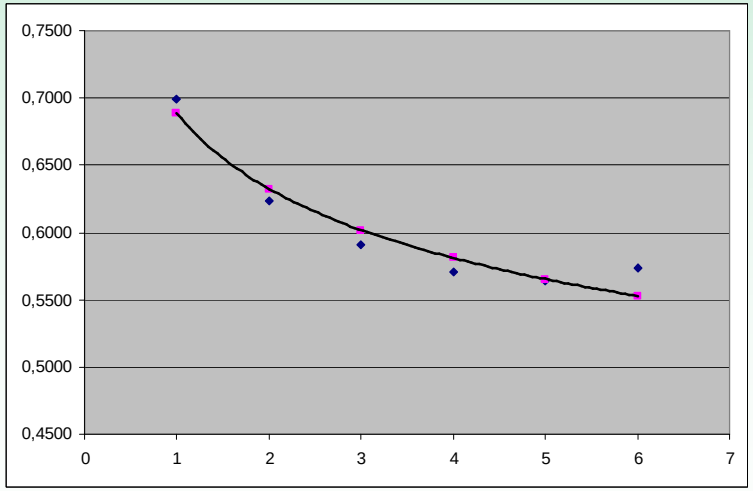
Modelling: Word length (WOL) and rel. cons. Frequencies (RCF)



Private letter

$$RCF = c*WOL^{-d}$$

	Paramter c	Paramter d	R^2
dissertations	0,6983	-0,1304	0,9903
dramas	0,6946	-0,1416	0,9776
sermons	0,7039	-0,1446	0,9981
private letters	0,6885	-0,1224	0,9349
complete corpus	0,7129	-0,1480	0,9920



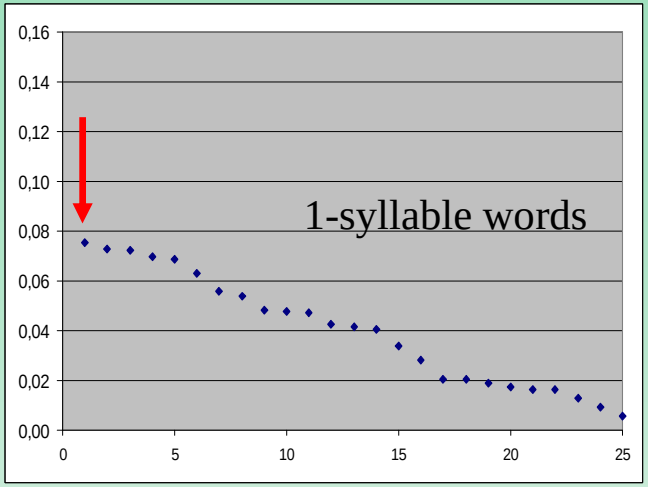
Dissertations

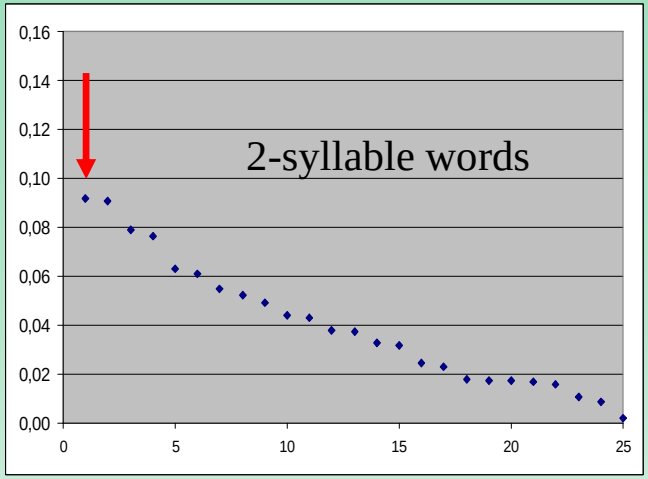
→ good fitting for all text sorts and the corpus

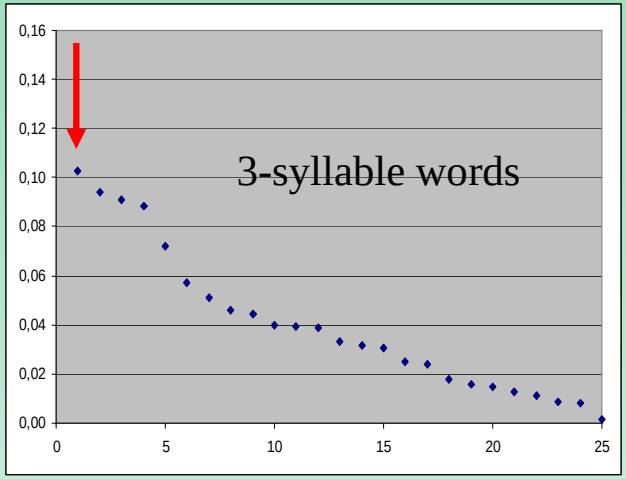
→ parameter *c* and *d* are text sort specific

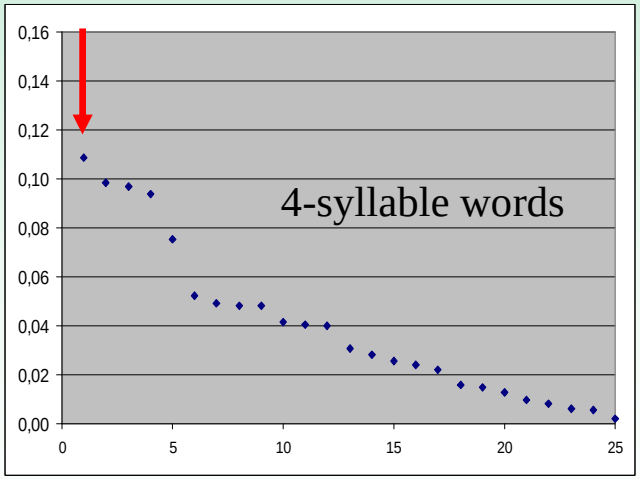
→ **consequences for grapheme frequency distributions?**

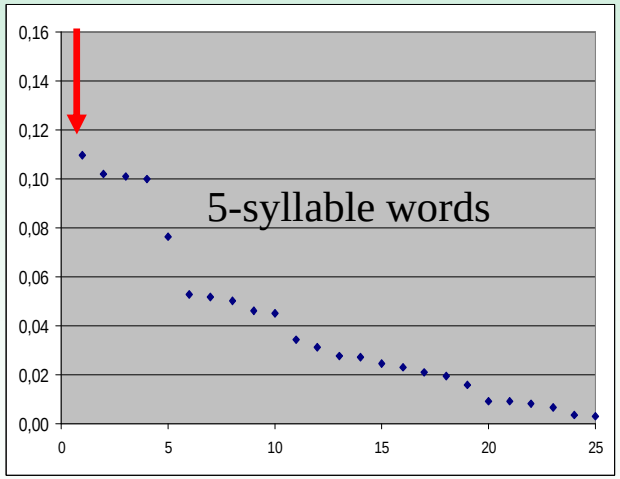
Relative rank grapheme frequencies of 1, 2, 3, 4 … *x* syllable words (corpus)

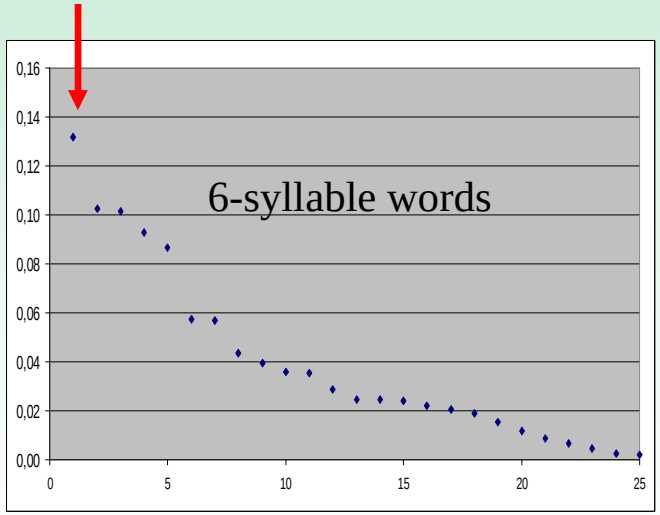












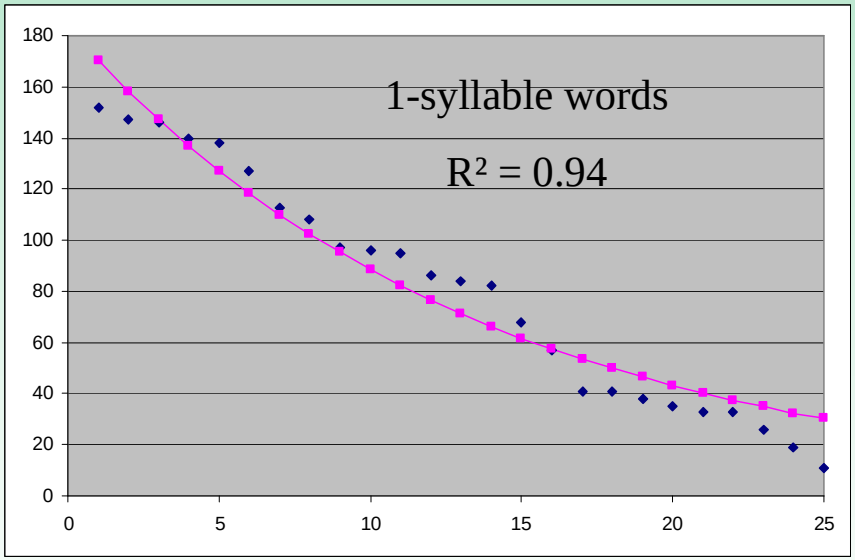
1. increase of p1 (first rank frequency) alongside with word length: (p1: 0.08 → 0.14)

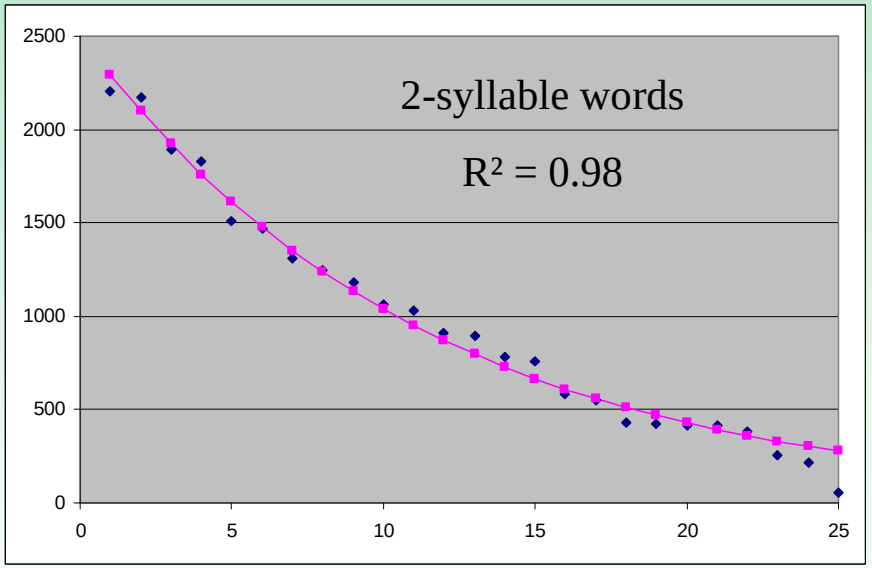
2. overall picture of the distribution curve changes !

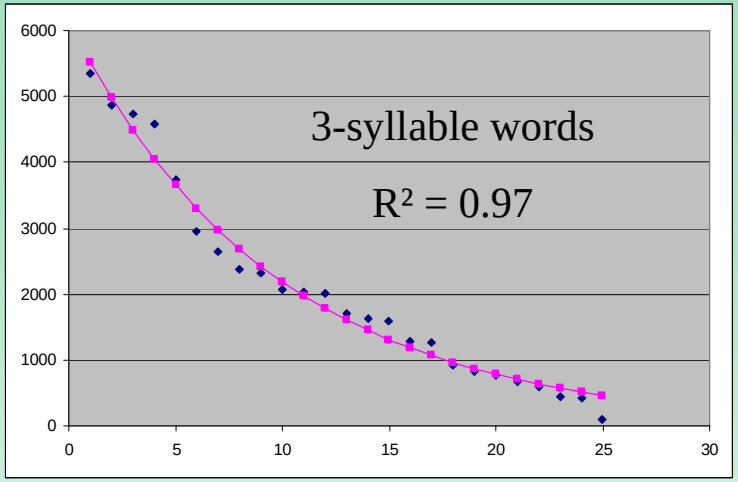
Parameter interpretation: Modelling grapheme frequencies

Popescu/Altmann/Köhler (2009): Zipf's law—another view

$$y = 1 + ge^{-hx}$$



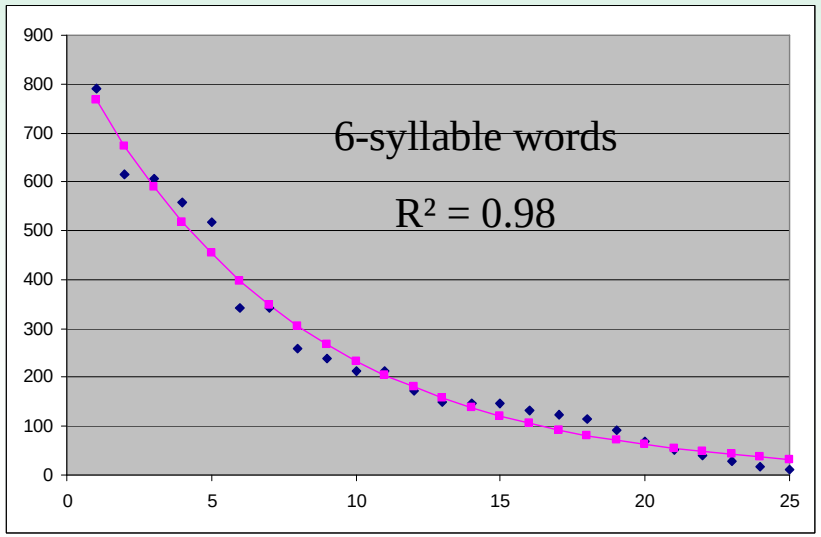




results for all word lengths

Wol/Syl.	g	| h |	R²
1	181,9781	**0,0731**	**0,9489**
2	2503,2874	**0,0883**	**0,9851**
3	6108,8006	**0,1028**	**0,9796**
4	5540,2261	**0,1115**	**0,9733**
5	2455,6990	**0,1175**	**0,972**
6	876,6859	**0,1325**	**0,9805**
7	400,1190	**0,1641**	**0,9825**

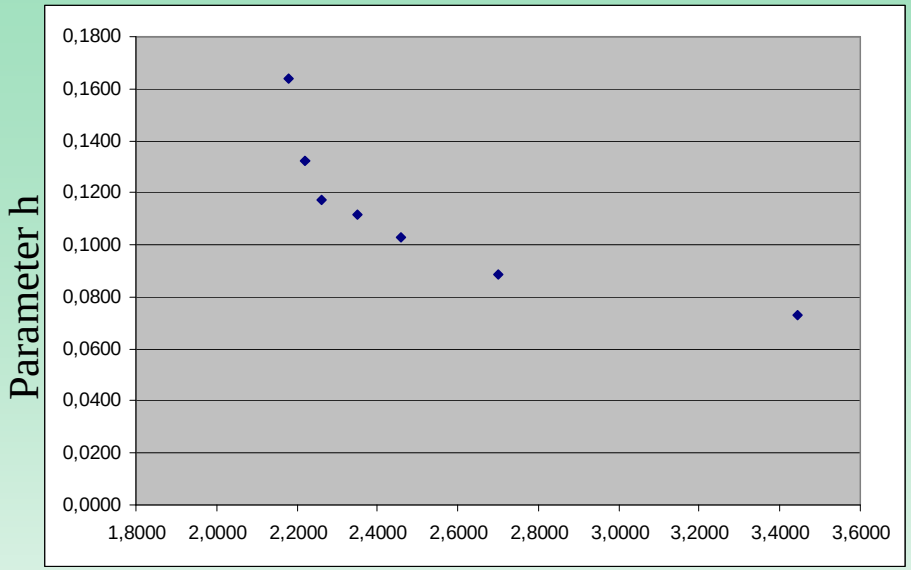
. . .



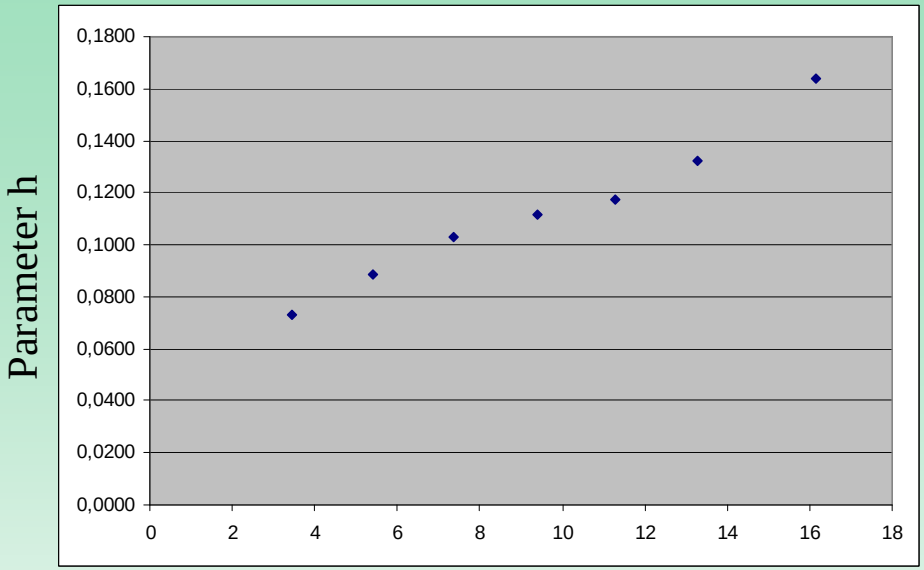
1. R^2 in all cases > 0.94 !
2. parameter *h* represents the steepness of the curve !

→ *h* depends on relative consonant frequency
→ *h* depends on syllable length
→ *h* depends on word length

Systematic behaviour of parameter *h* (corpus)



Mean syllable length/graphemes

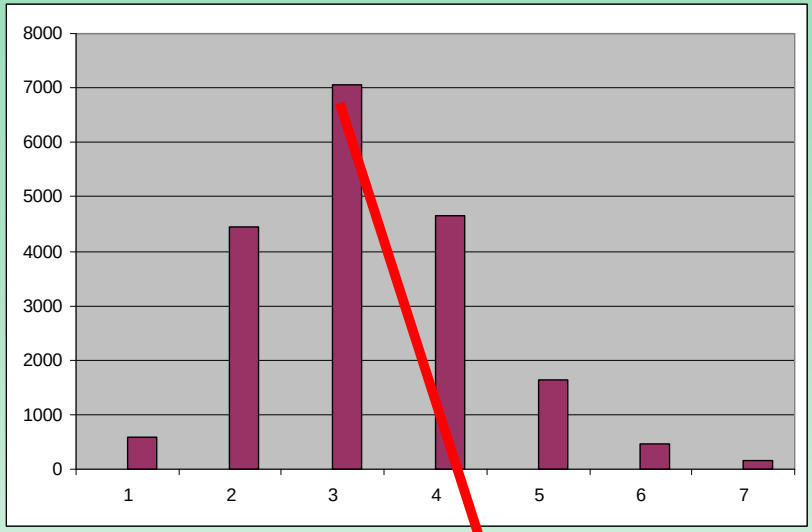


Mean word length/graphemes

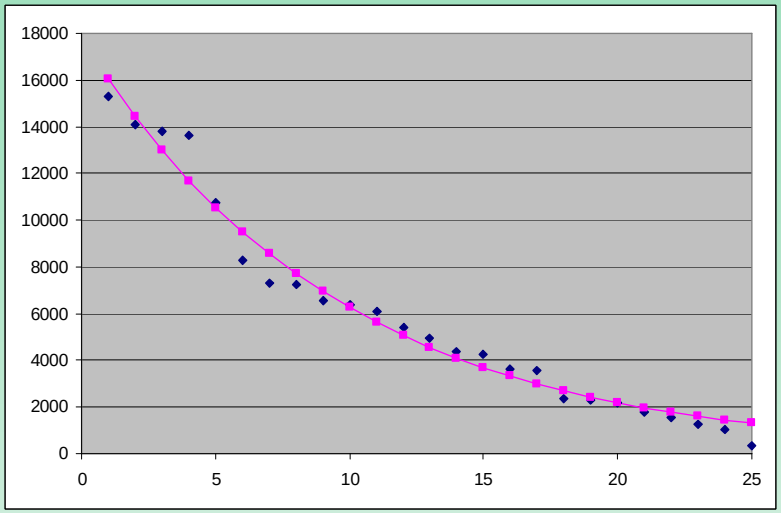
→ The longer the syllable, the lower parameter h !

→ The longer the word, the higher parameter h !

What about the grapheme frequency of the whole „corpus“?



Word length frequency



$y = 17787{,}78 * e^{-0.1047*x}$

$R^2 = 0.9769$

Wol/Syl.	g	| h|	R^2	f (Wol)
1	181,9781	**0,0731**	**0,9489**	**584**
2	2503,2874	**0,0883**	**0,9851**	**4440**
3	6108,8006	**0,1028**	**0,9796**	**7038**
4	5540,2261	**0,1115**	**0,9733**	**4658**
5	2455,6990	**0,1175**	**0,972**	**1640**
6	876,6859	**0,1325**	**0,9805**	**450**
7	400,1190	**0,1641**	**0,9825**	**141**

$$\frac{\sum h \cdot f(Wol)}{N} = 0.1030$$

***h* can be replaced**

→ $R^2 = 0.9767$ (!)

Major results

- same relation was for all other analysed Slovene text types too!
- same relation was obtained for Serbian too !
- ML controls the relative consonant/vowel frequency of words
- ML controls the shape of grapheme frequencies
- parameters of grapheme frequency models are in a systematic relation to

  >> mean syllable length

  >> mean word length

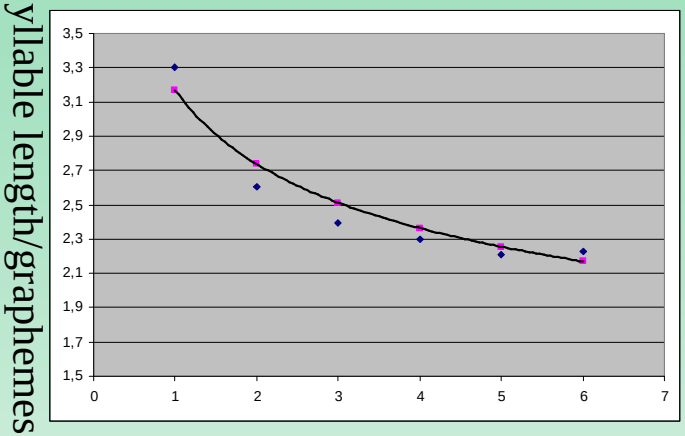
Precondition: Proof of Word length – Syllable length relation in Slovene

1. word length (WOL): number of syllables
2. word form types (!)
3. syllable length (SYL): graphemes/phonemes

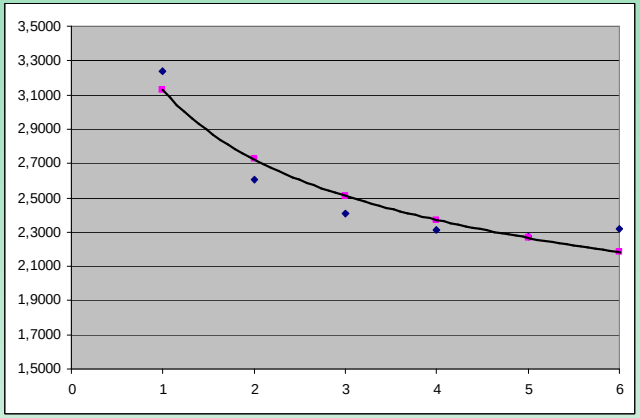
4. Slovene material:

text sort	**number of texts**	**word form types**
dissertations	5 chapters	6144
private letters	30 (Ivan Cankar)	5182
sermons	32	7977
dramas	42 acts (Drago Jančar)	5616
complete corpus		24919

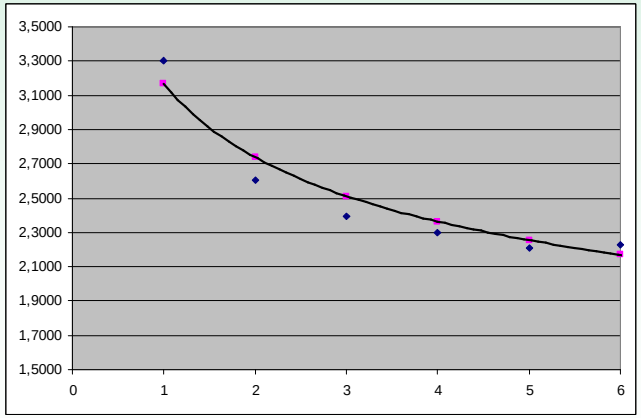
$$SYL = a*WOL^{-b}$$



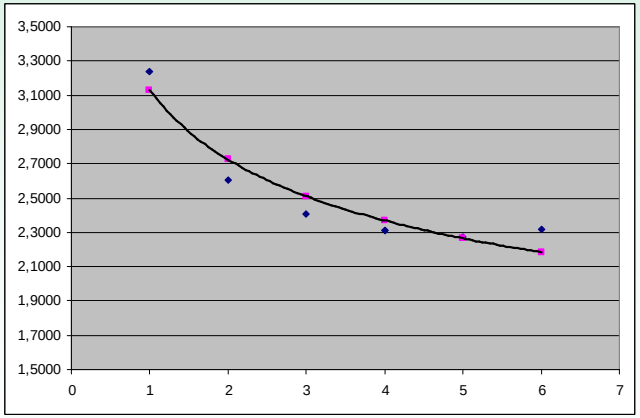
dissertations $SYL = 3.22*WOL^{-0.20}$ **$R^2 = 0.98$**



sermons $SYL = 3.23*WOL^{-0.22}$ **$R^2 = 0.97$**



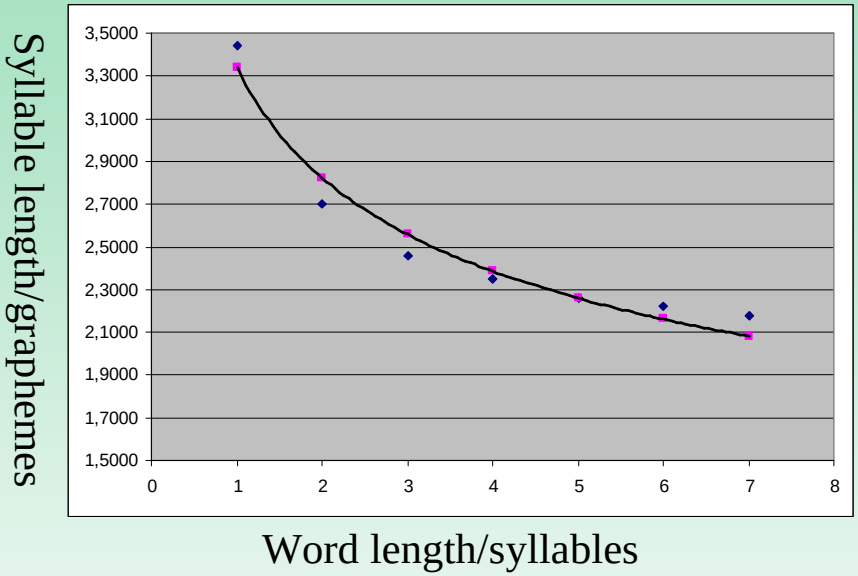
dramas $SYL = 3.17*WOL^{-0.21}$ **$R^2 = 0.98$**



private letters $SYL = 3.13*WOL^{-0.20}$ **$R^2 = 0.92$**

Modelling Word length – Syllable length

Replacing parameter *a* with the mean syllable length of 1-syllable words



Complete corpus SYL = $3.34*WOL^{-0.24}$ **R^2 = 0.95**

Summary of results

	Paramter a	Paramter b	R^2
dissertations	3,2254	-0,2099	**0,9832**
dramas	3,1700	-0,2115	**0,9000**
sermons	3,2371	-0,2271	**0,9742**
private letters	3,1315	-0,2012	**0,9259**
complete corpus	3,3405	-0,2429	**0,9584**

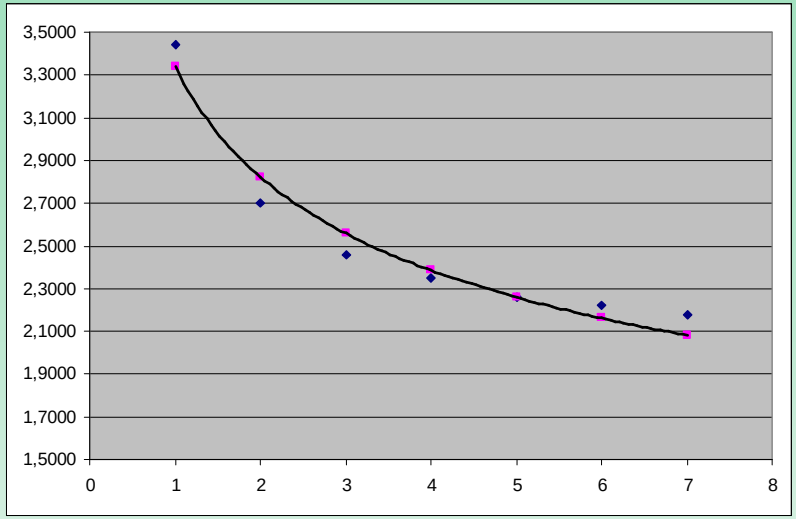
mean syllable length		**R^2**
dissertations	**3,2727**	**0,982**
dramas	**3,3003**	**0,8761**
sermons	**3,3134**	**0,9673**
private letters	**3,2386**	**0,8998**
complete corpus	**3,4435**	**0,9478**

→ parameter *a* of ML can be replaced by mean syllable length

→ only a marginal loss of information

→ low decrease of R^2 !

And finally: complete corpus



Complete corpus SYL = $3.34*WOL^{-0.24}$

R^2 = 0.95

Summary of results

	Parameter a	Parameter b	R^2
dissertations	3,2254	-0,2099	**0,9832**
dramas	3,1700	-0,2115	**0,9000**
sermons	3,2371	-0,2271	**0,9742**
private letters	3,1315	-0,2012	**0,9259**
complete corpus	3,3405	-0,2429	**0,9584**

→ ML abides WOL – SYL Relation in all text sorts and in the corpus !

→ parameter *a* as “starting point” for constituent lengths

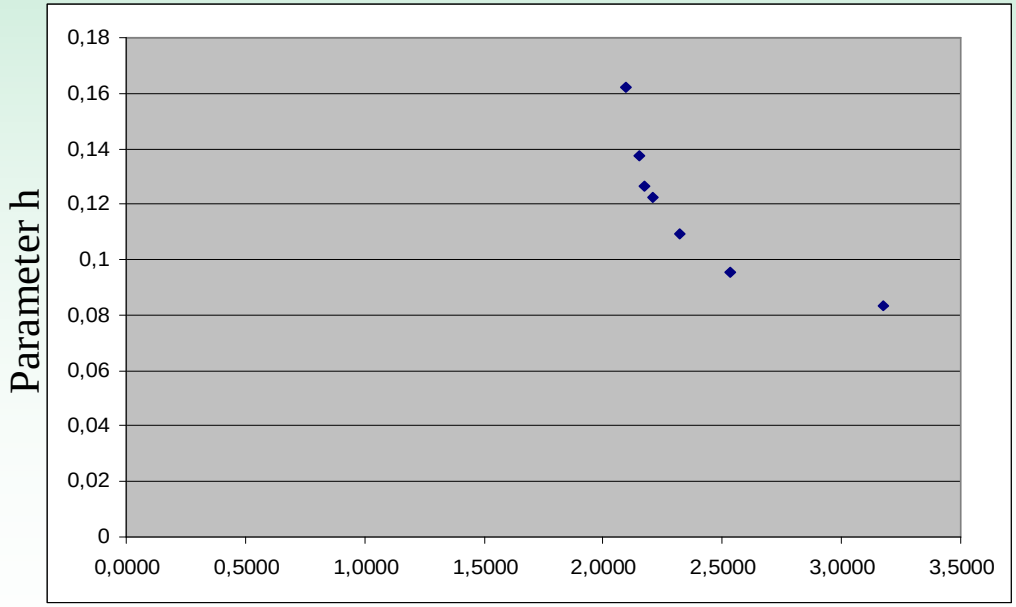
→ parameter a and *b* are text sort specific

→ but parameter *b* shows no significant differences

→ Hypothesis on a relation between grapheme frequencies and word length is justified !

Serbian Corpus

text sort	number of texts	word form types
scientific texts	10	4948
lit. prose	30	5216
journalistic texts	30	5436
sermons	32	4365
complete corpus		19965



Mean syllable length/graphemes