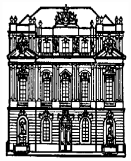
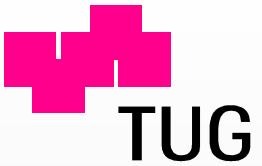




Peter Grzybek Emmerich Kelih Ernst Stadlober

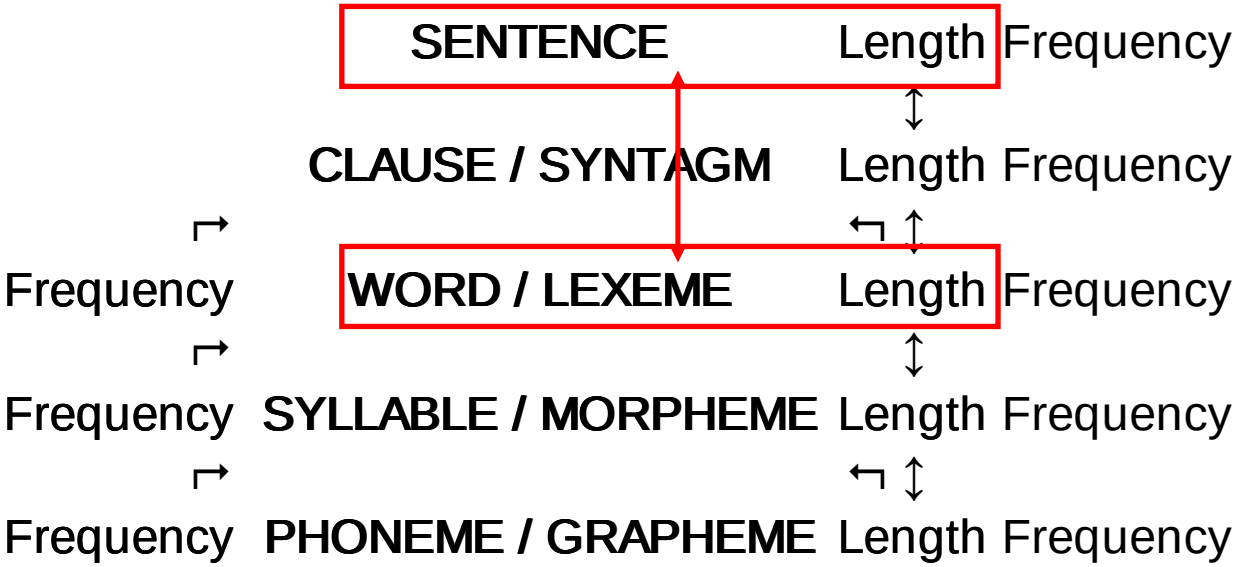
Long sentences, long words – short sentences, long words?

A Linguistic Contribution to the Study of Relationships Between Units of Different Levels

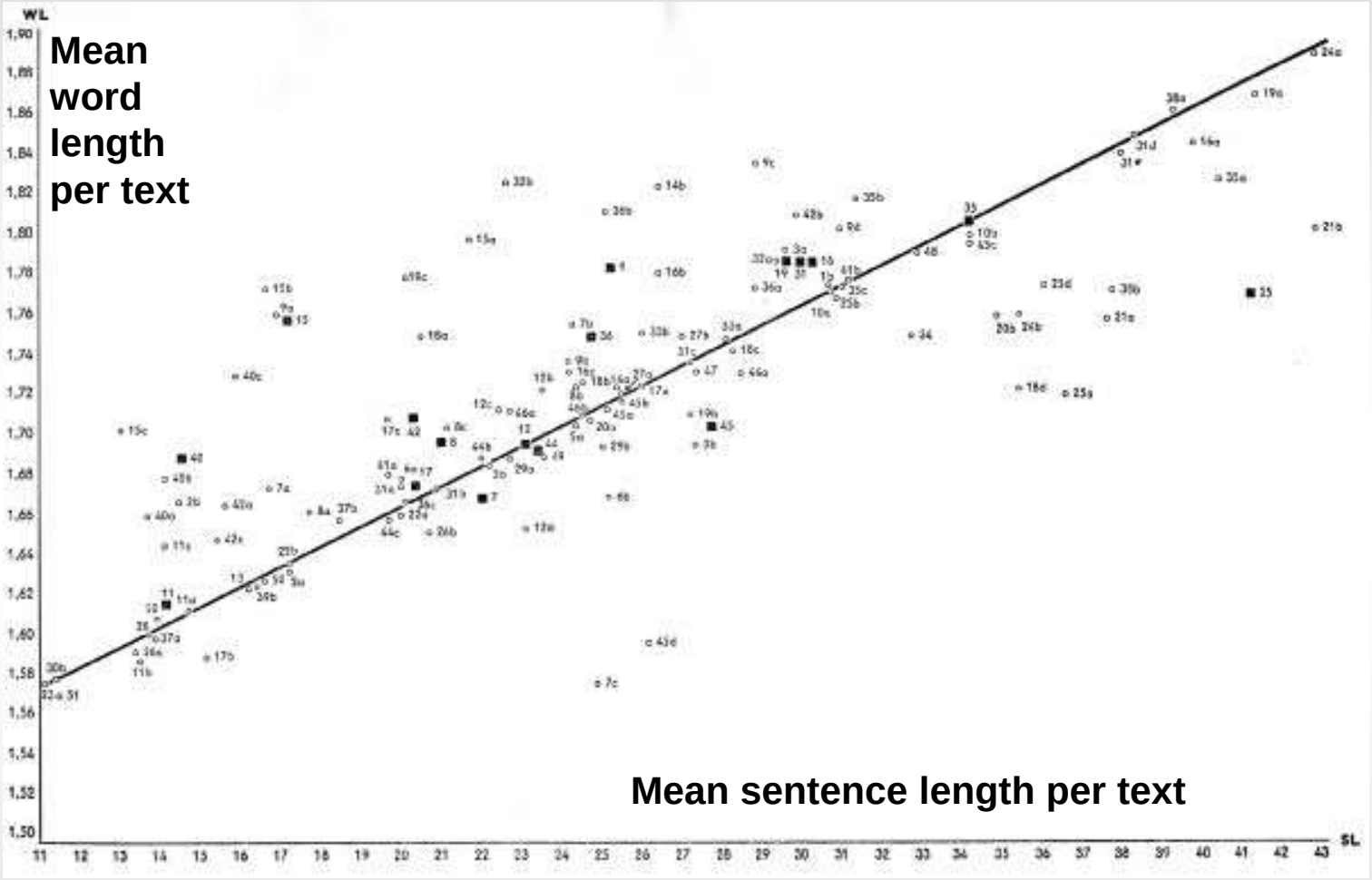
http://www-gewi.uni-graz.at/quanta

Synergetics In a Nutshell –

Frequencies and Dependencies



Sentence Length ⇔ Word Length



Data: 117 German Literary Prose Texts (Arens 1965)

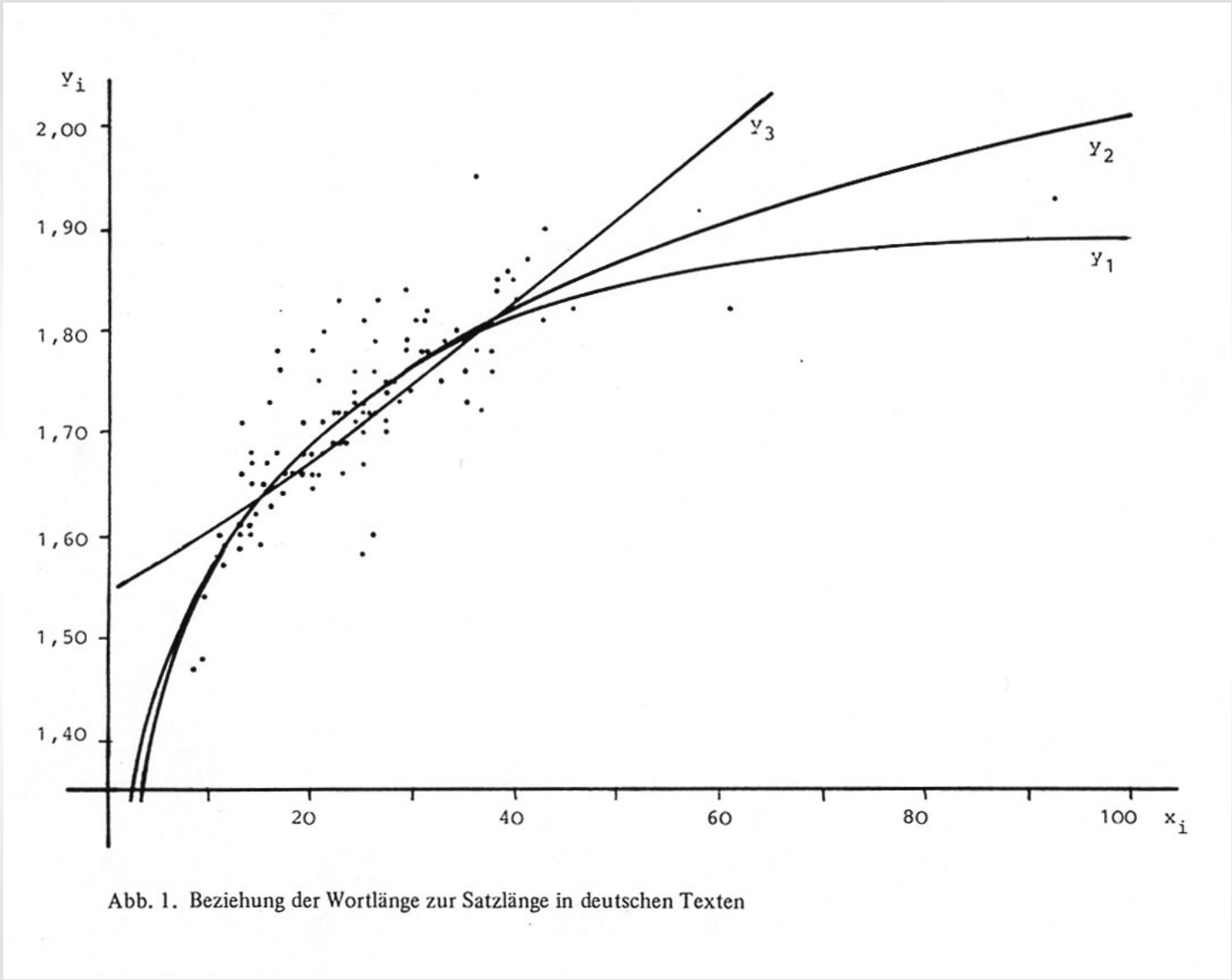


Abb. 1. Beziehung der Wortlänge zur Satzlänge in deutschen Texten

Re-Analysis: Altmann (1983) * Data: Arens (1965)

Menzerath's Law

Altmann's (1980) Interpretation and the Wimmer-Altmann Extension (2005)

The larger (longer) a construct, the smaller (shorter) its constituents

I	$y = ax^{-b}$	2
II	$y = ae^{cx}$	2
III	$y = ax^{-b}e^{cx}$	3
IV	$y = ae^{(d/x)}$	2
V	$y = ax^{-b}e^{(d/x)}$	3
VI	$y = ax^{-b}e^{cx}e^{(d/x)}$	4

Relationship Between Units of Different Levels

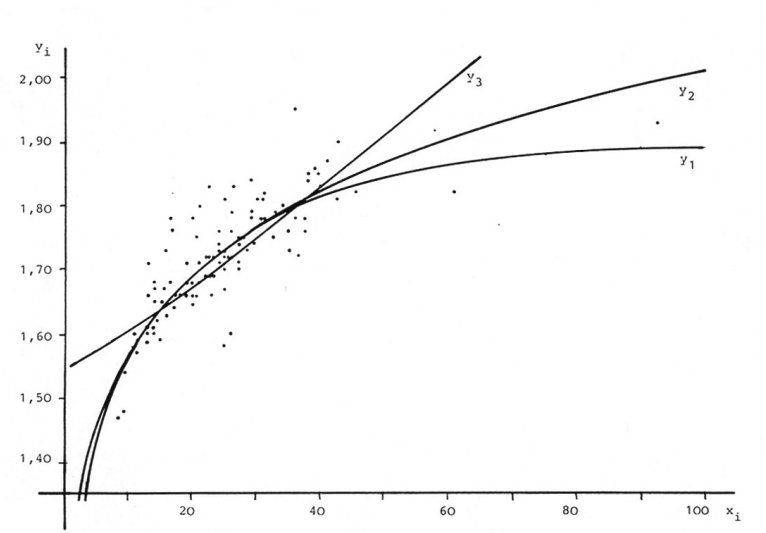


Abb. 1. Beziehung der Wortlänge zur Satzlänge in deutschen Texten

„Direct“ relations

(sentence --> clause / clause --> word)

$y_c = ax^{-b}$

Indirect relations

(sentence --> [clause] --> word)

Altmann-Arens Law

⇓

$y_w = cx^d$

$$y_c = a_1 x^{-b_1}$$

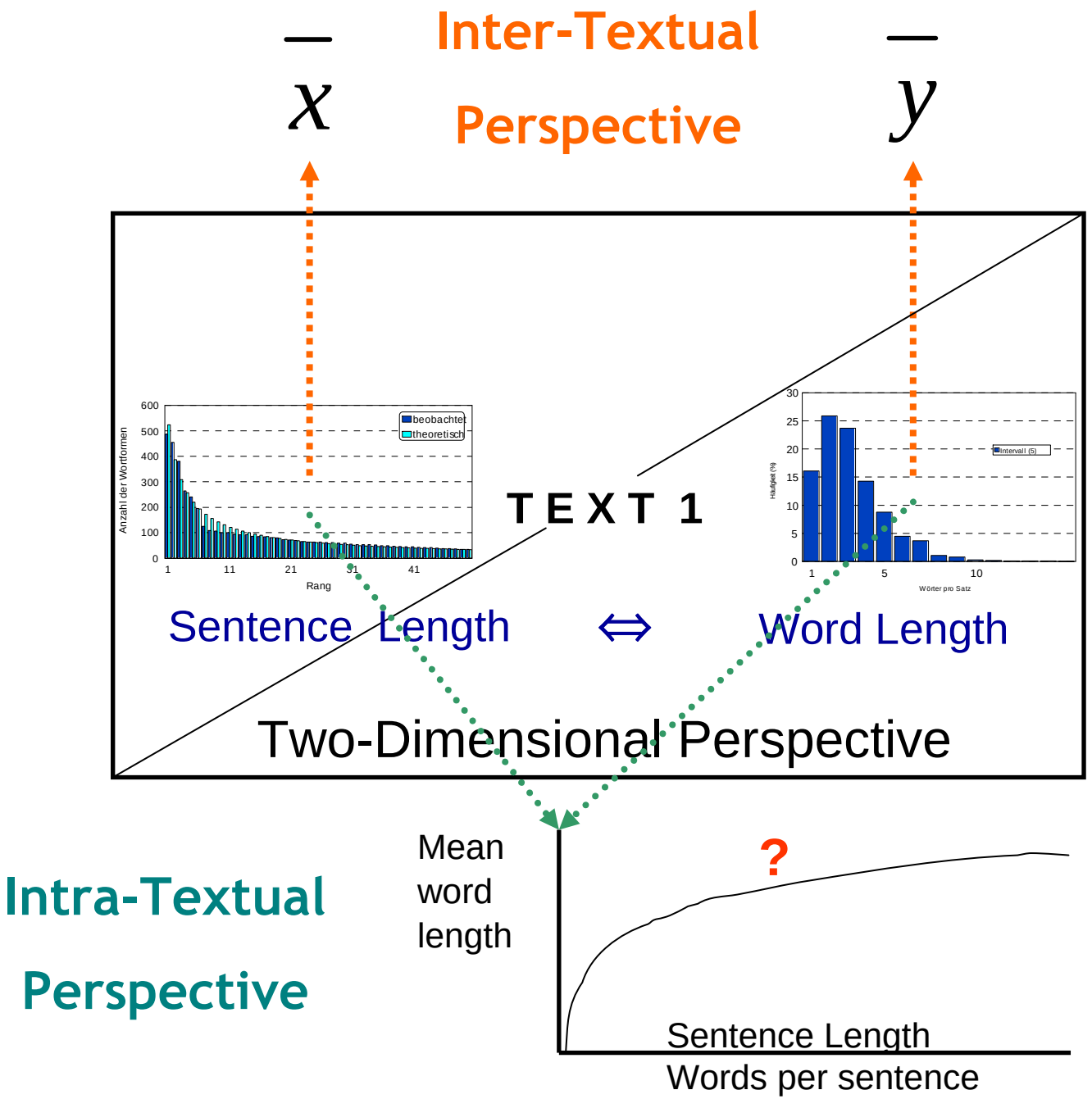
$$y_w = a_2 x^{-b_2} = a_2 \left(a_1 x^{-b_1} \right)^{-b_2} = a_2 a_1^{-b_2} x^{b_1 b_2} = c x^d$$

Testing the SL / WL Relation with a Larger Text Corpus (*N* = 404) and Different Text Sorts*

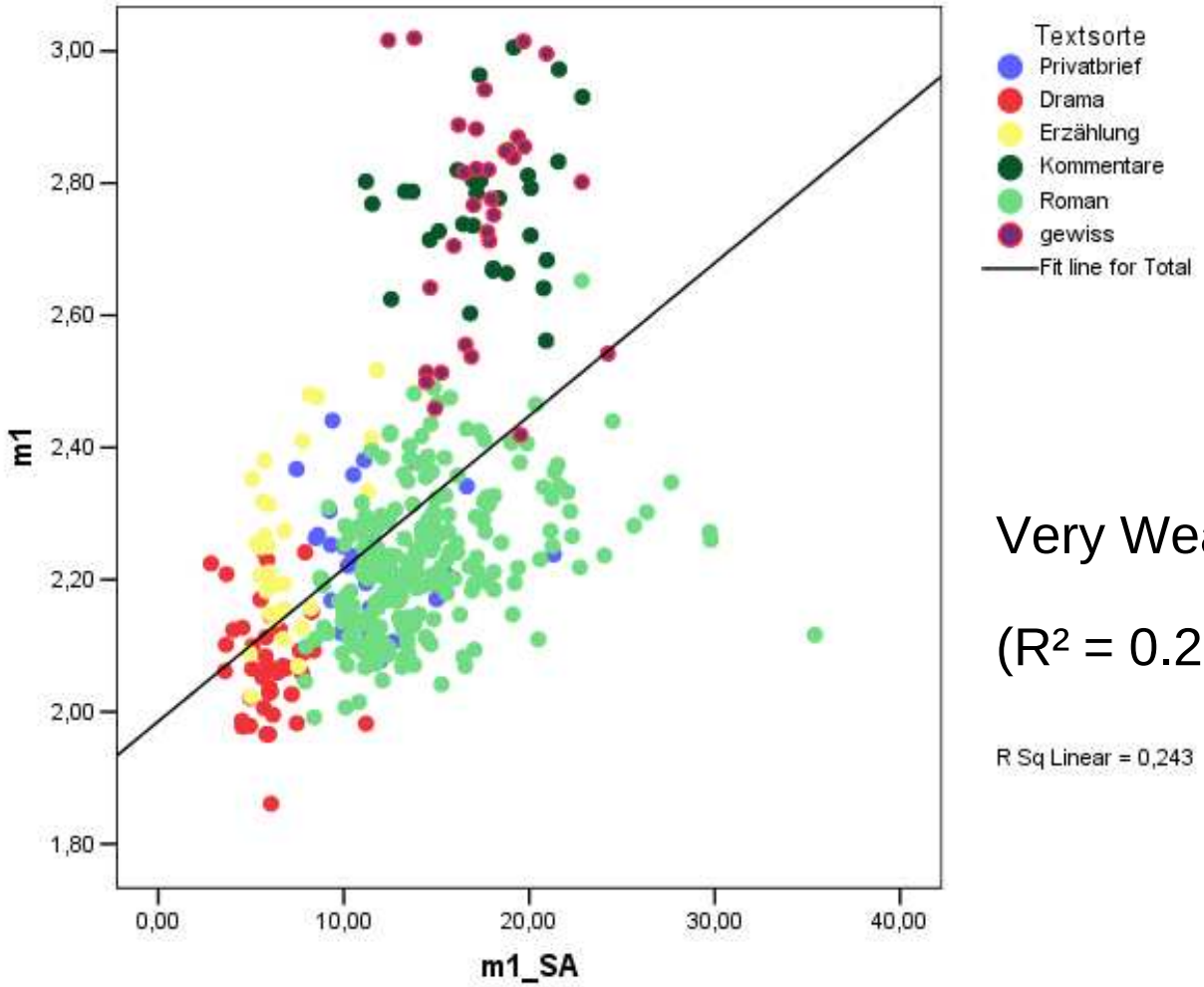
Text Sort	Author	N
Private Letter	A.P. Čechov	30
Drama	A.P. Čechov	44
Stories	A.P. Čechov	31
Comments	div.	30
Literary Prose	L.N. Tolstoj	239
Scholarly Prose	div.	30
		404

* Graz Text Data Base

http://quanta-textdata.uni-graz.at/


$\bar{x}$
Inter-Textual
Perspective
$\bar{y}$
Anzahl der Wortformen
beobachtet
theoretisch
Rang
TEXT 1
Häufigkeit (%)
Intervall (5)
Wörter pro Satz
Sentence Length ⇔ Word Length
Two-Dimensional Perspective
Intra-Textual
Perspective
Mean
word
length
?
Sentence Length
Words per sentence

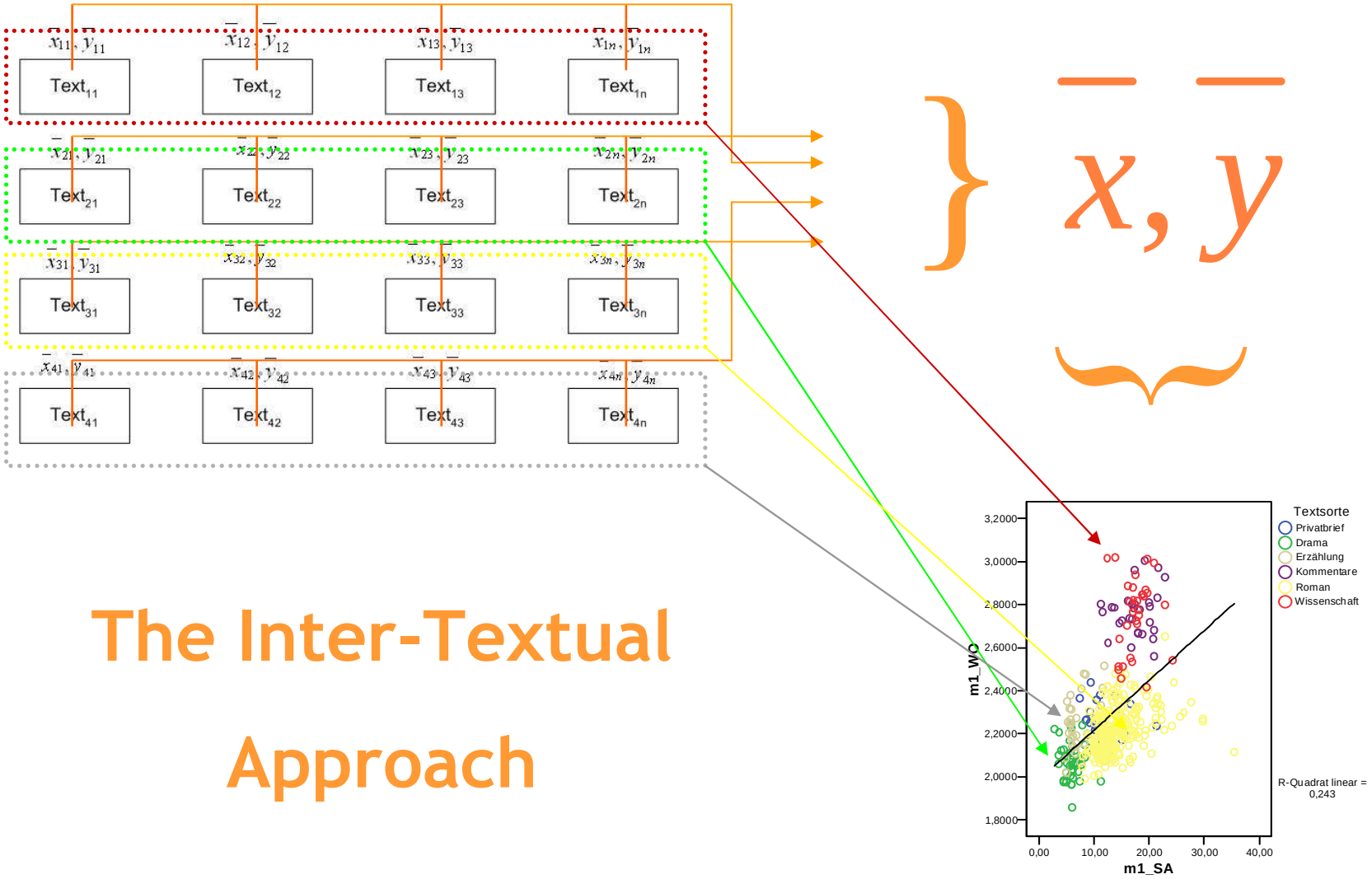
Combined Text Corpus



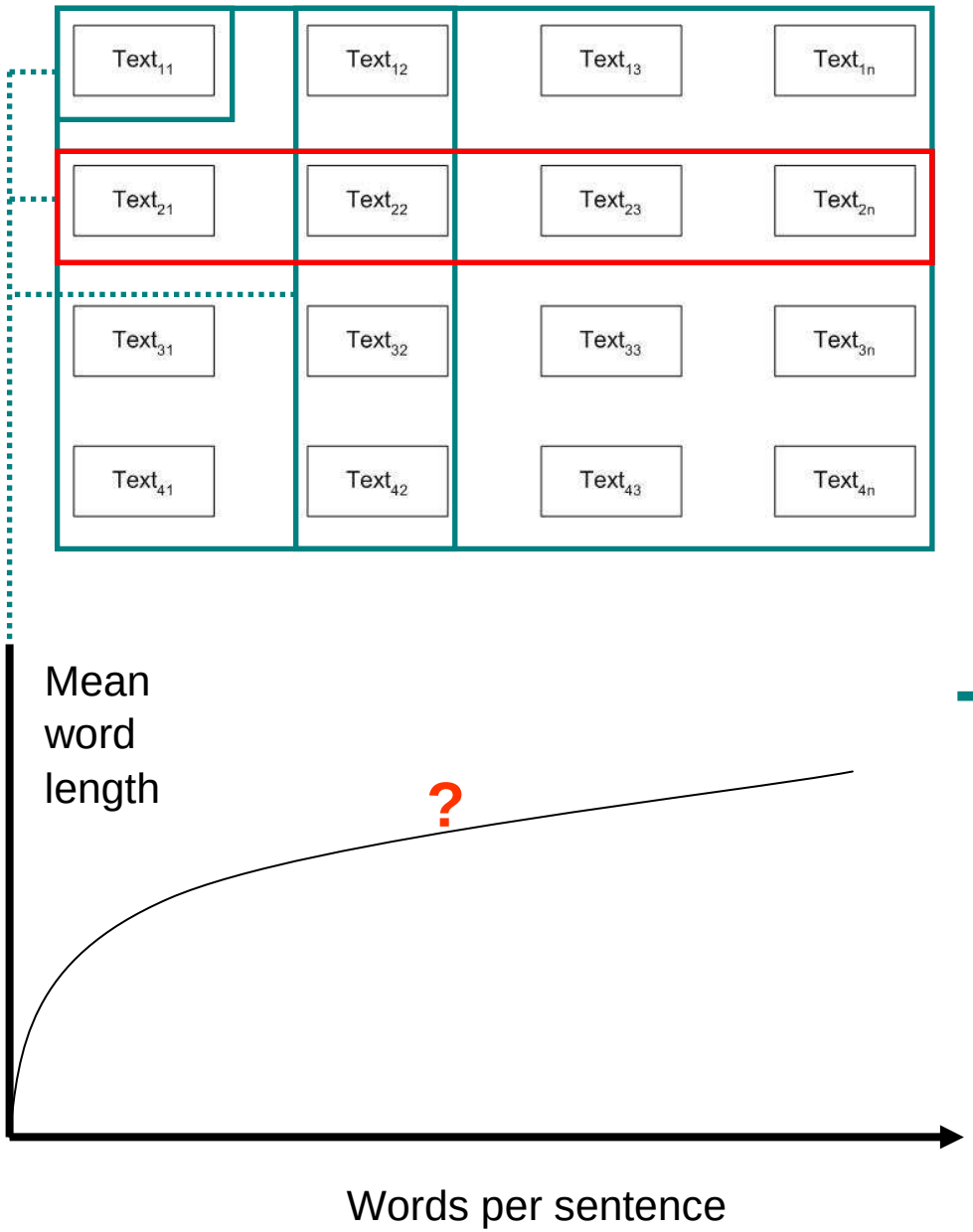
Very Weak WL / SL Relation

($R^2 = 0.24$) !

Need for a Systematization of SL / WL Studies



The Inter-Textual Approach



The Intra-Textual Approach

Text Basis For Intra-Textual Approach*

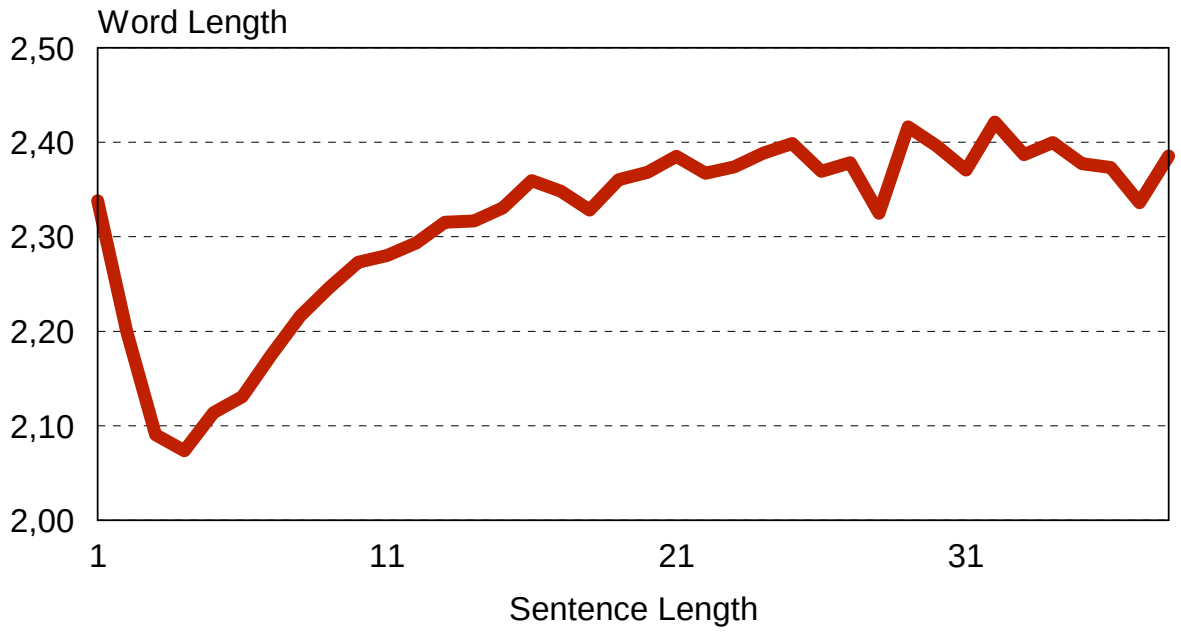
Text Sort	Author	N	Words		Sentences		m_1 WL	m_1 SL
			abs.	rel.	abs.	rel.		
Drama	A.P. Čechov	44	67.430	0,28	11.125	0,47	2,04	6,06
Private Letters	A.P. Čechov	30	56.751	0,23	4.178	0,18	2,19	13,58
	A.A. Achmatova	30						
	D. Charms	30						
	L.N. Tolstoj	30						
Literary Prose	L.N. Tolstoj	69	74.708	0,31	5.680	0,24	2,20	13,15
Comments	div.	60	43.263	0,18	2.556	0,11	2,67	16,93
Corpus		**293**	**242.152**	1	**23.539**	1	**2,25**	**10,29**

* Graz Text Data Base

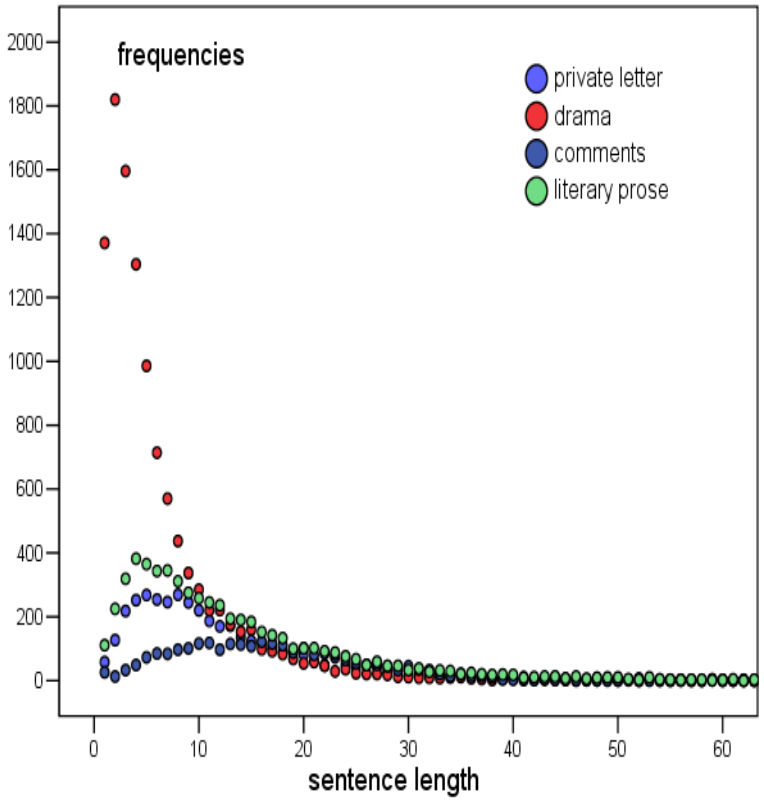
http://quanta-textdata.uni-graz.at/

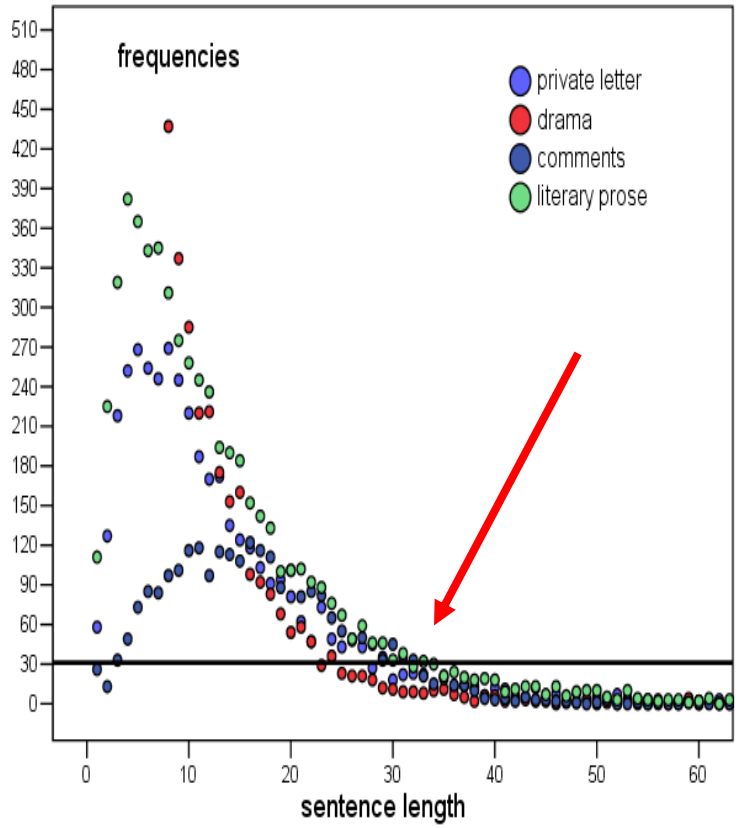
Mean word length, for individual sentence length classes

(Total corpus, sentence length: *f* > 100)

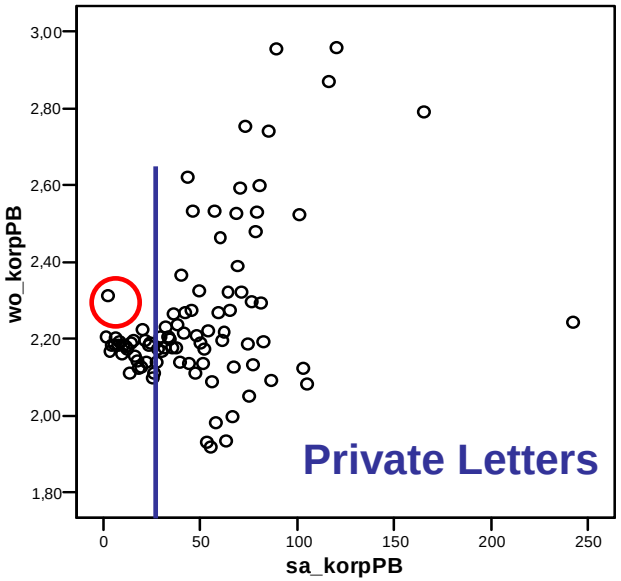


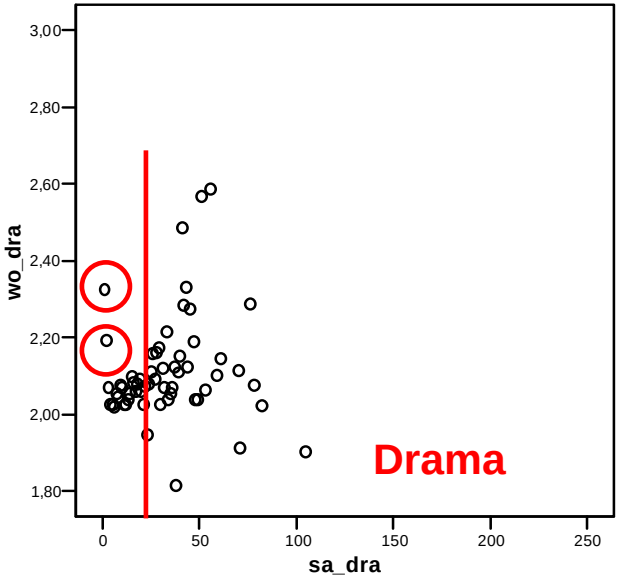
Frequency Distributions of Sentence Lengths

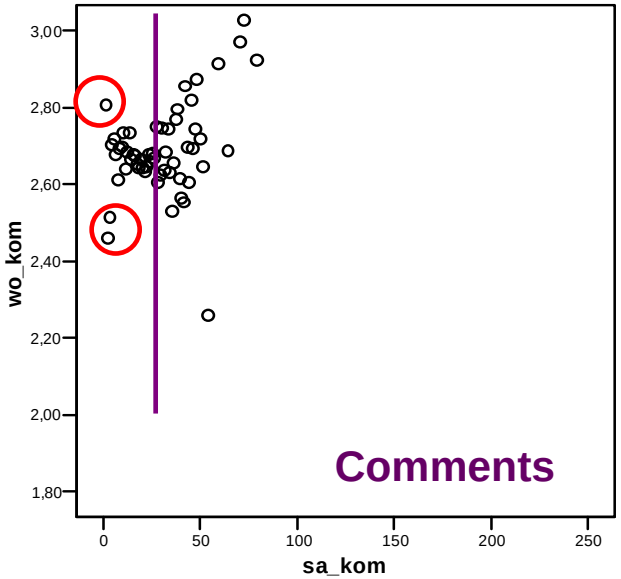


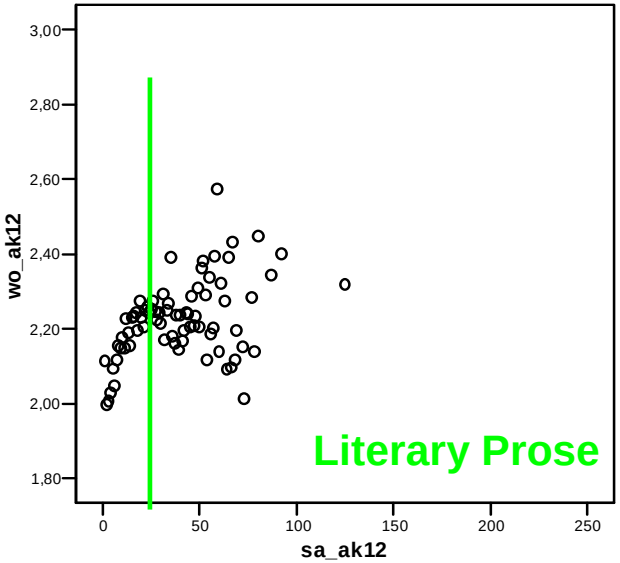


Empirical restriction: Frequency of sentence length > 30


wo_korpPB
sa_korpPB
Private Letters

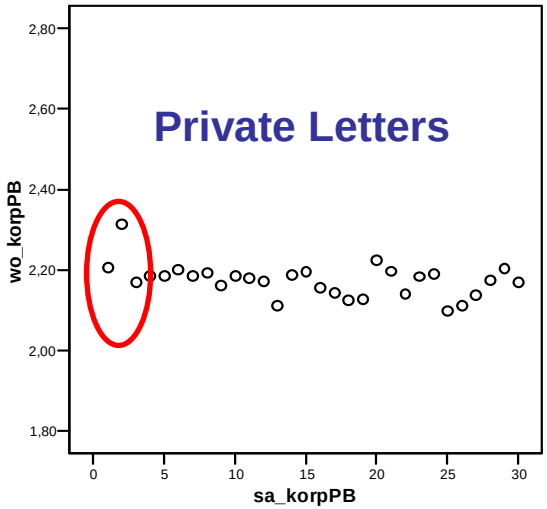

wo_dra
sa_dra
Drama

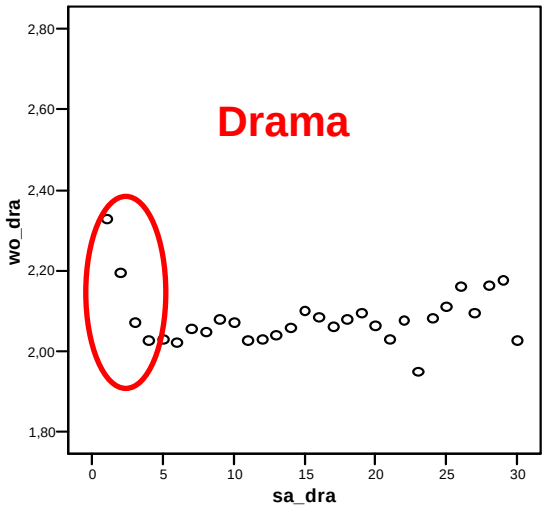

wo_kom
sa_kom
Comments

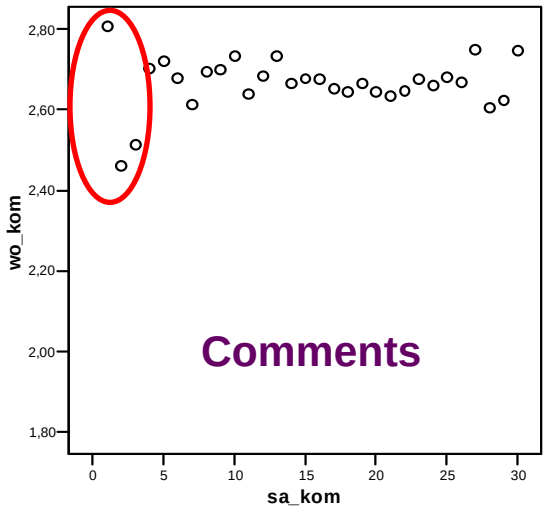

wo_ak12
sa_ak12
Literary Prose

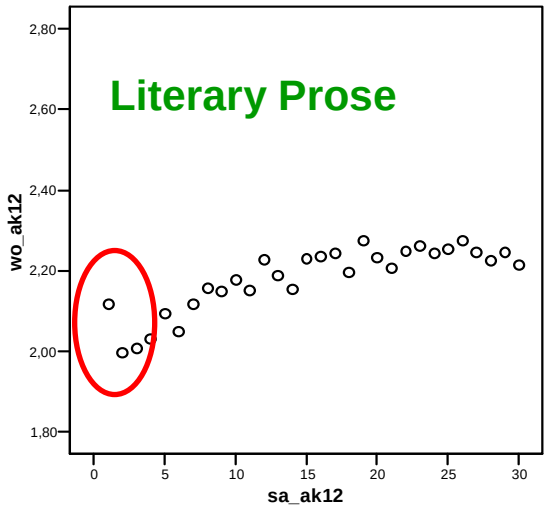
Word Length (WL) / Sentence Length (WL) Relation

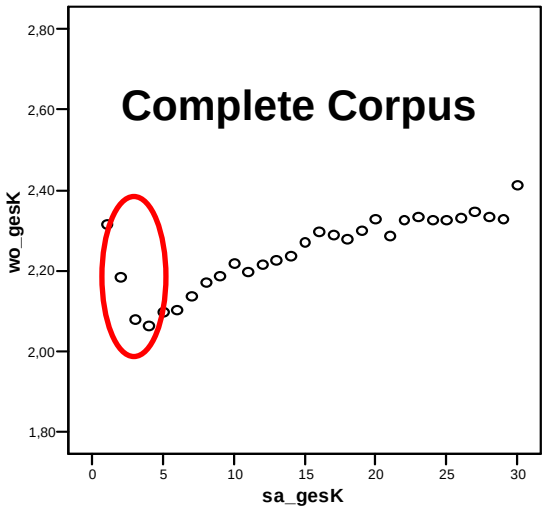
Restricted Conditions: SL < 30



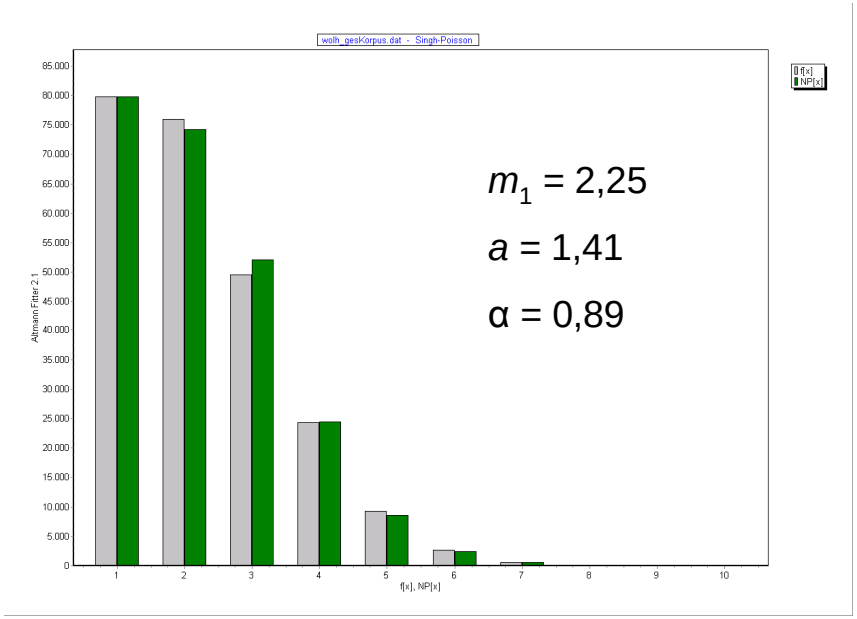








Word Length Frequencies in the Corpus



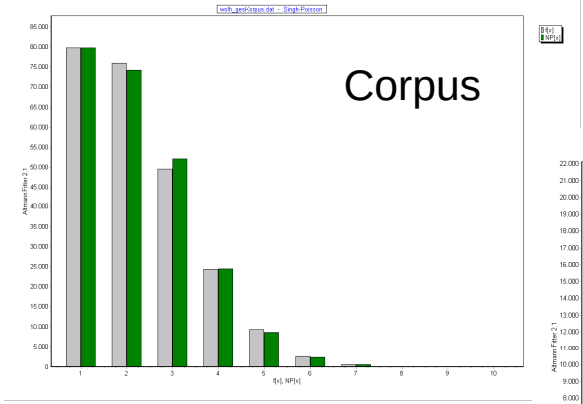
Singh-Poisson (a, α)

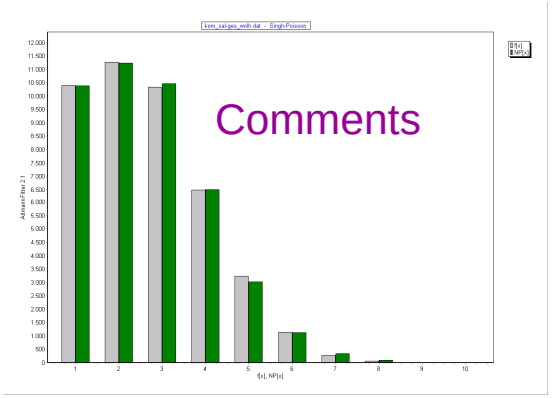
$$P_x = \begin{cases} 1-\alpha+\alpha e^{-a} & x=1 \\ \dfrac{\alpha a^{(x-1)} e^{-a}}{(x-1)!} & x=2,3,\ldots \end{cases}$$

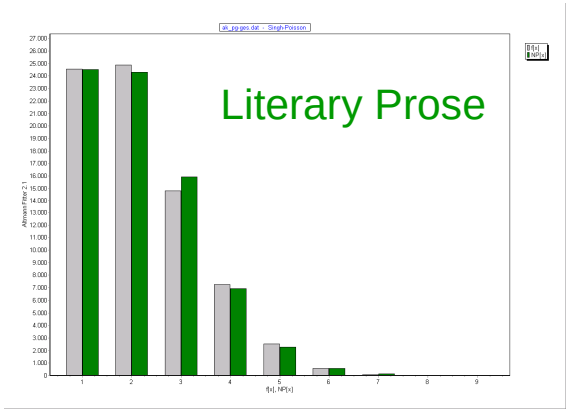
Poisson (a) ≡ Singh-Poisson (a,1)

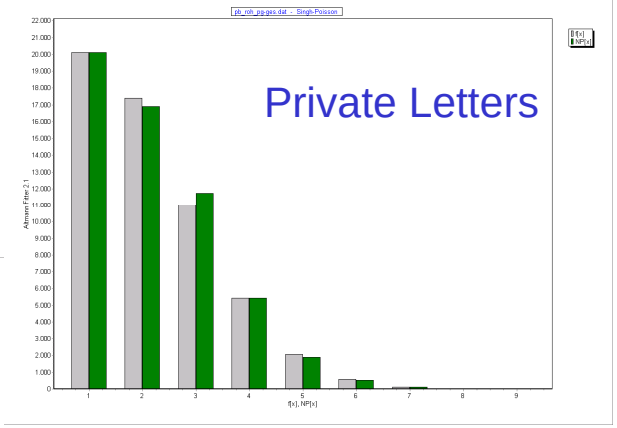
$$P_x = \frac{a^{(x-1)} e^{-a}}{(x-1)!} \qquad x=1,2,3,\ldots$$

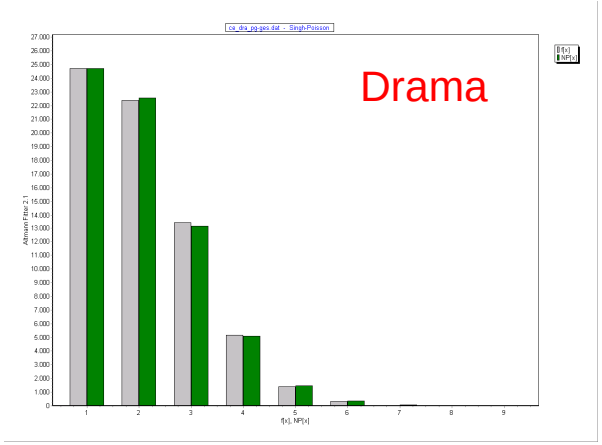
Overall Word Length Frequencies in Different Text Types











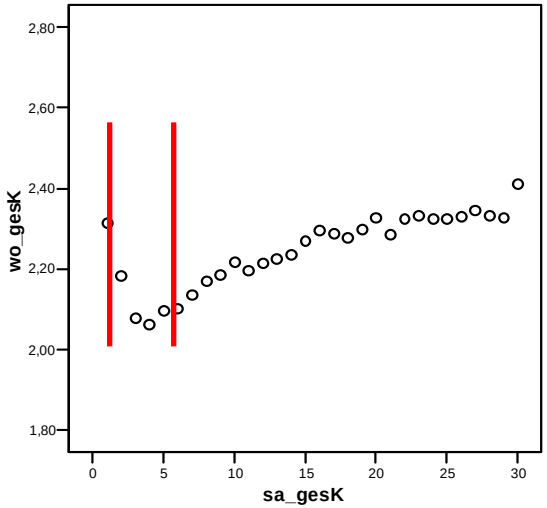
Text type	m_1	s	a	α
Corpus	2,25	1,50	1,41	0,89
Comments	2,67	1,64	1,86	0,90
Lit. prose	2,20	1,48	1,31	0,92
Private letters	2,19	1,48	1,39	0,86
Drama	2,07	1,44	1,16	0,92

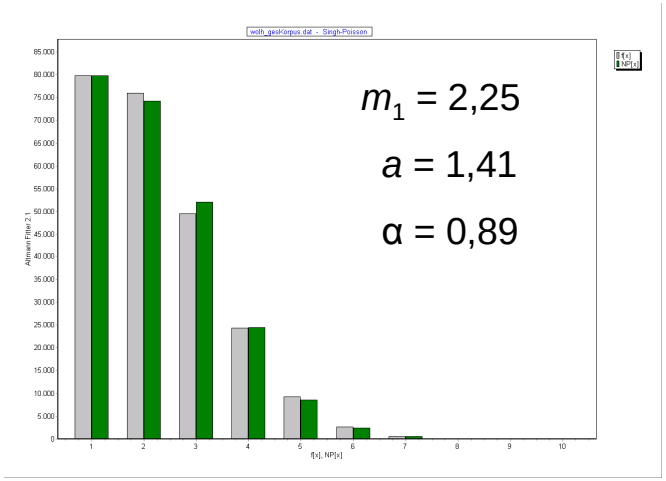
→Significant deviation from homogeneity for all text types ($p < 0.0001$)

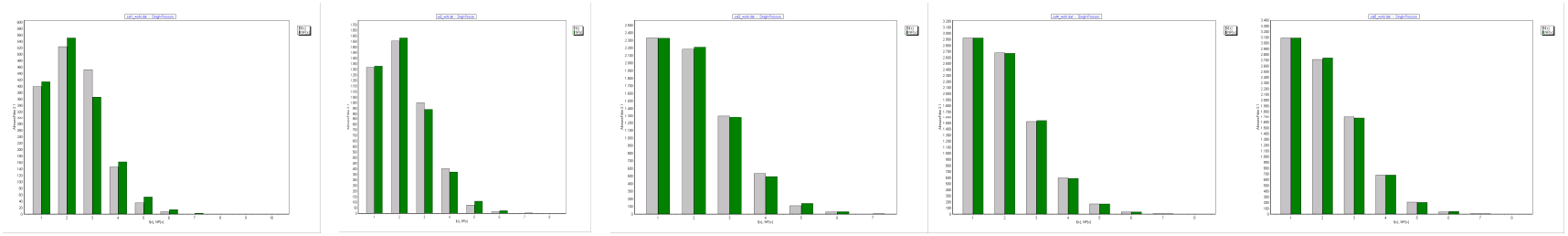
→ One type of frequency model

SL 1-5: Distribution Singh-Poisson

Total Corpus (Length 1-5 = 12% of Corpus)



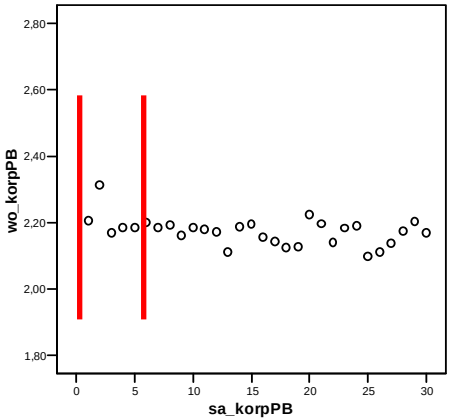


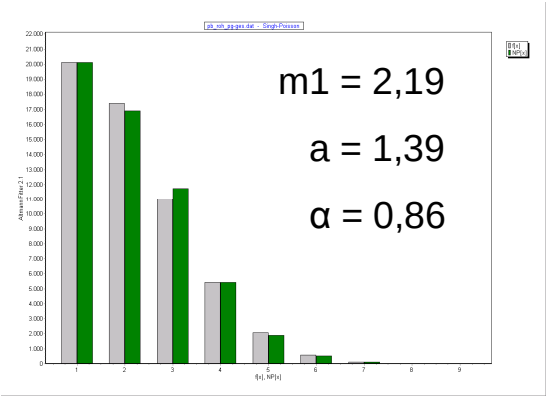


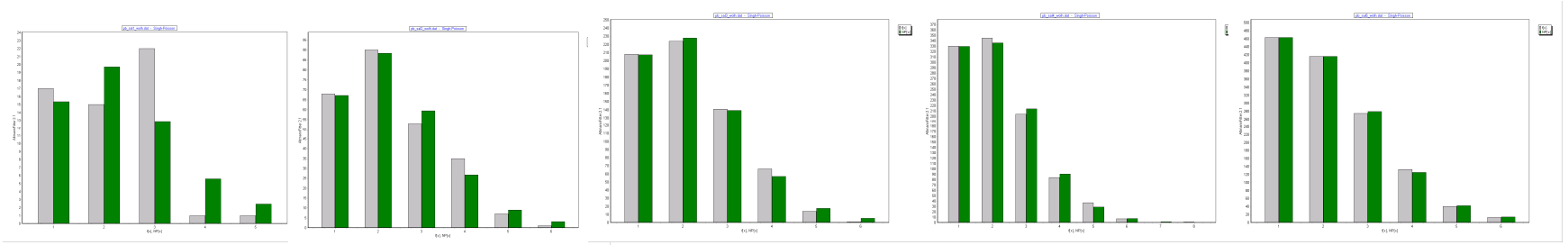
	1	2	3	4	5
m_1	2,32	2,18	2,08	2,06	2,10
a	1,33	1,19	1,16	1,15	1,22
α	0,99	0,99	0,94	0,92	0,90
fit	–	+	+	+	+

SL 1-5: Distribution Singh-Poisson

Private Letters (Length 1-5 = 5.6% of sub-corpus)



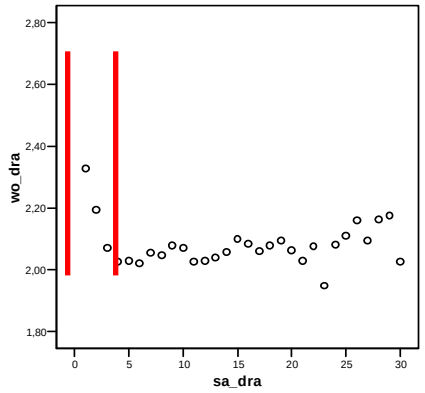


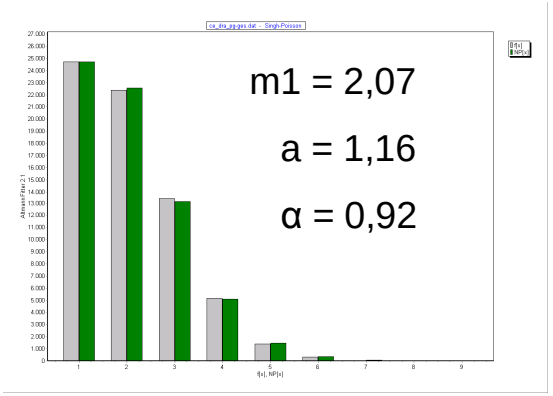


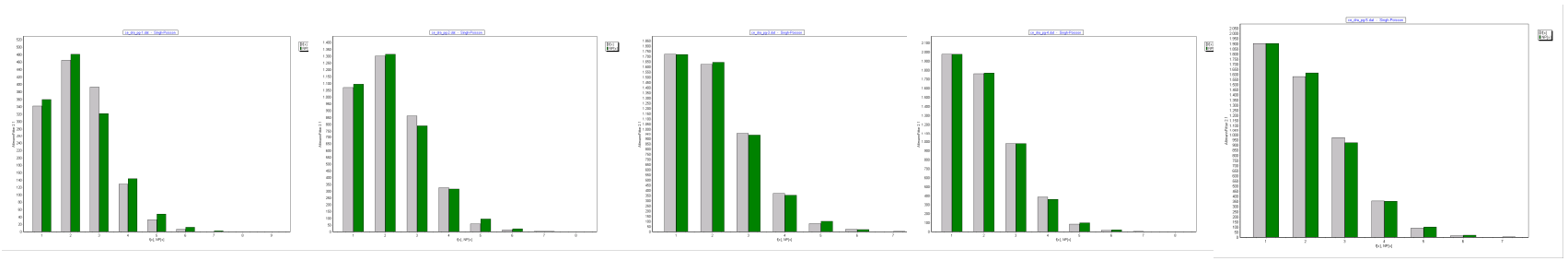
	1	2	3	4	5
m_1	2,18	2,32	2,17	2,19	2,19
a	1,31	1,35	1,22	1,27	1,34
α	0,99	0,99	0,97	0,94	0,89
fit	–	+	+	+	+

SL 1-5: Distribution Singh-Poisson

Drama (Length 1-5 = 29.58% of sub-corpus)



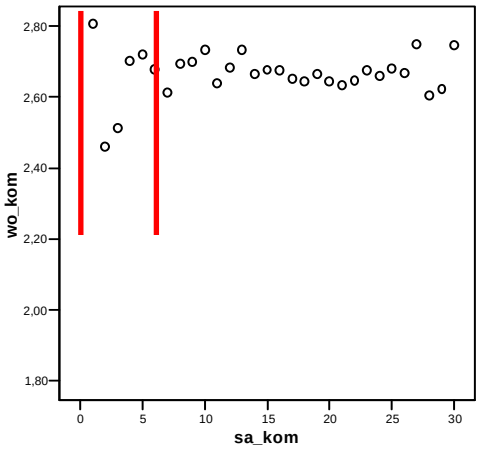


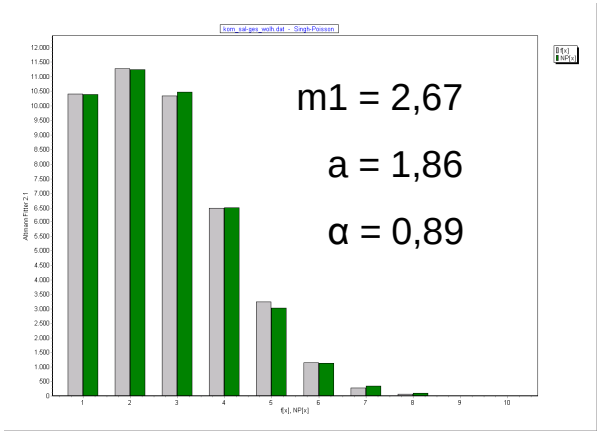


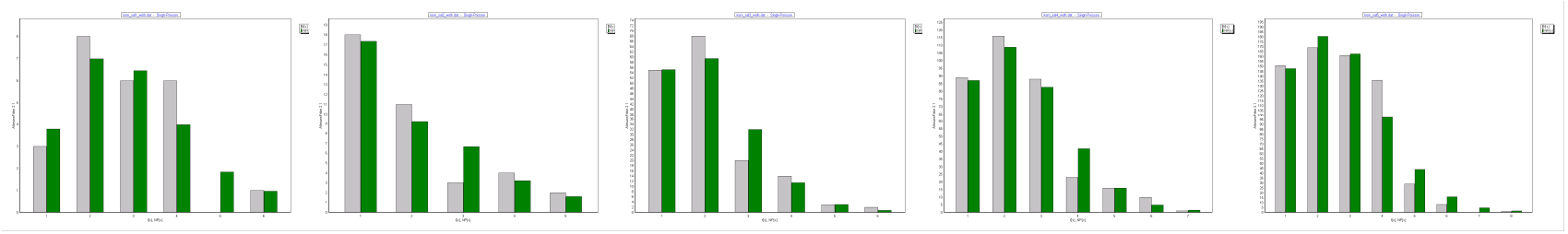
	1	2	3	4	5
m_1	2,33	2,19	2,07	2,03	2,03
a	1,34	1,20	1,14	1,11	1,14
α	0,99	0,99	0,94	0,93	0,90
fit	+	+	+	+	+

SL 1-5: Distribution Singh-Poisson

Comments (Length 1-5 = 2.8% of sub-corpus)



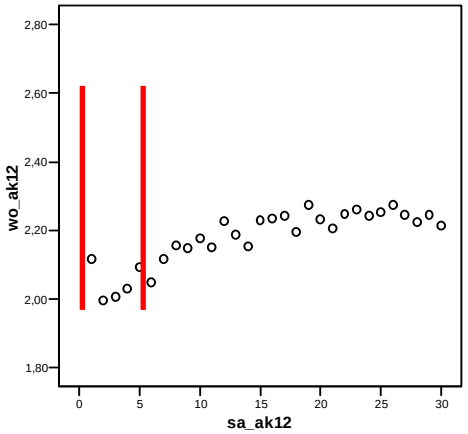


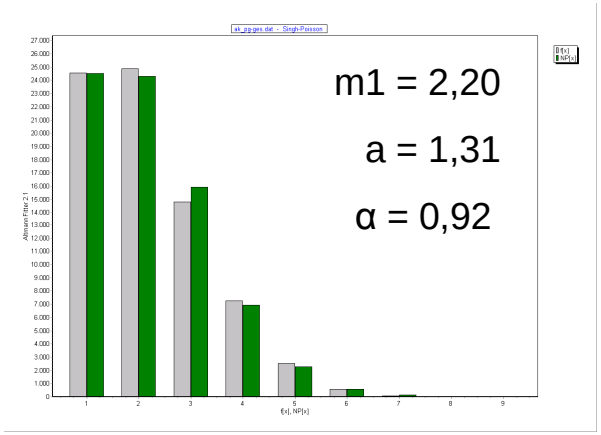


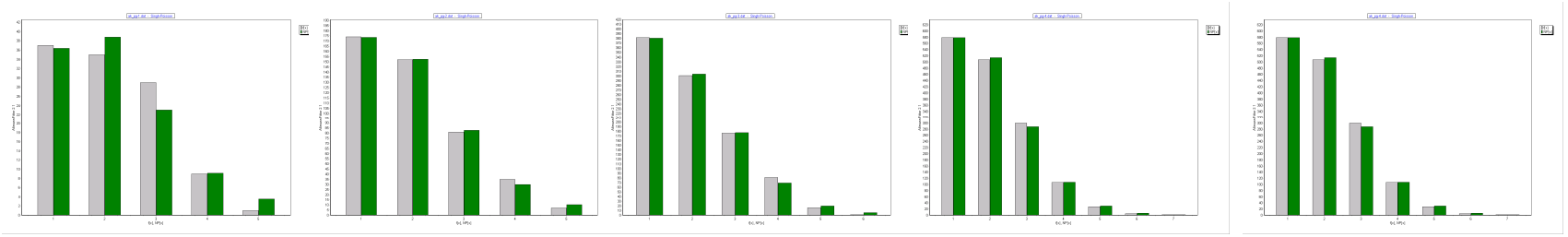
	1	2	3	4	5
m_1	2,79	1,97	2,06	2,40	2,62
a	1,85	1,45	1,08	1,52	1,80
α	0,99	0,71	0,99	0,95	0,93
fit	+	+	–	–	–

SL 1-5: Distribution Singh-Poisson

Literary Prose (Length 1-5 = 6.5% of sub-corpus)

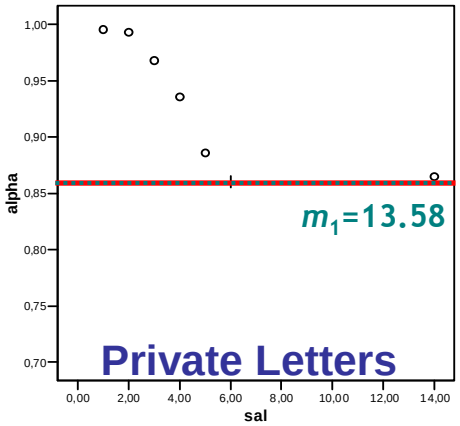


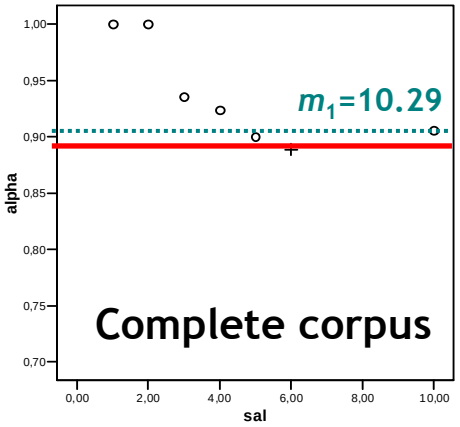


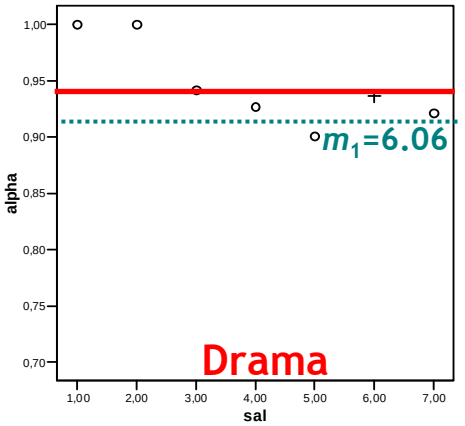


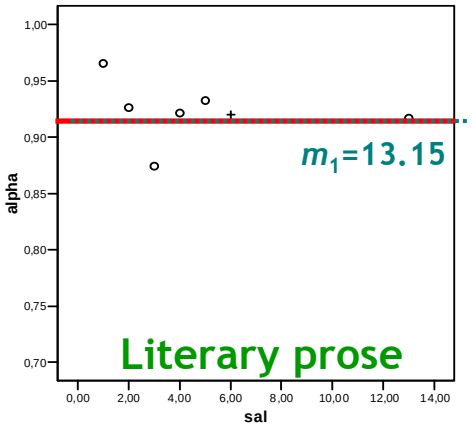
	1	2	3	4	5
m_1	2,12	2,00	2,01	2,03	2,10
a	1,19	1,08	1,17	1,12	1,18
α	0,97	0,93	0,87	0,92	0,93
fit	+	+	+	+	+

Parameter α (word length), for specific sentence lengths (comparison with mean and sub-corpus)



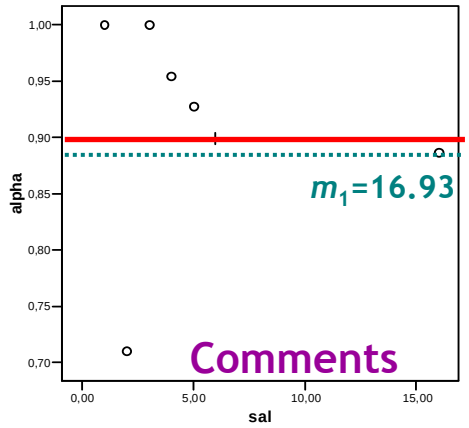






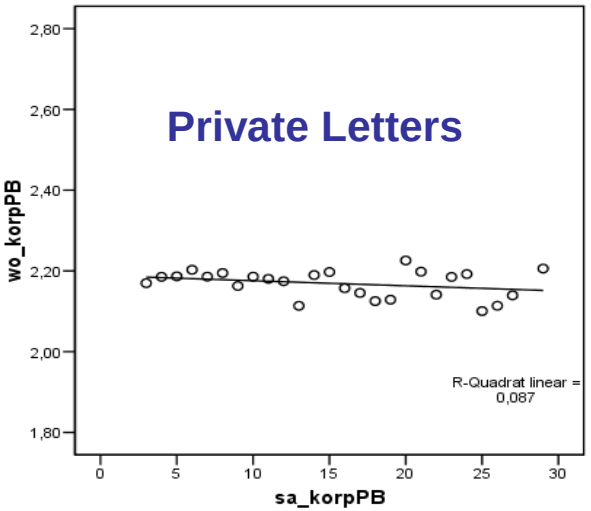
(sub)corpus mean

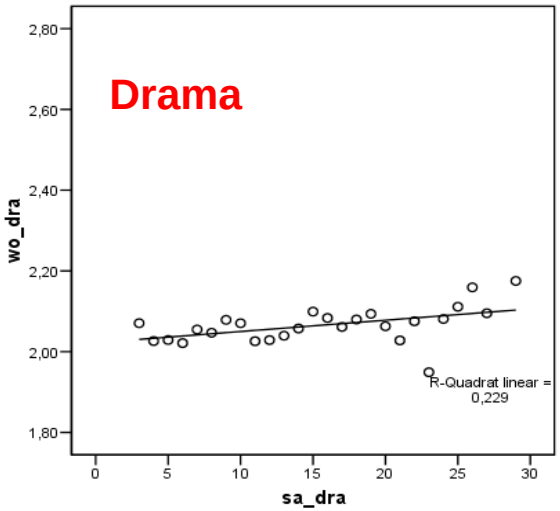
Sentence length mean

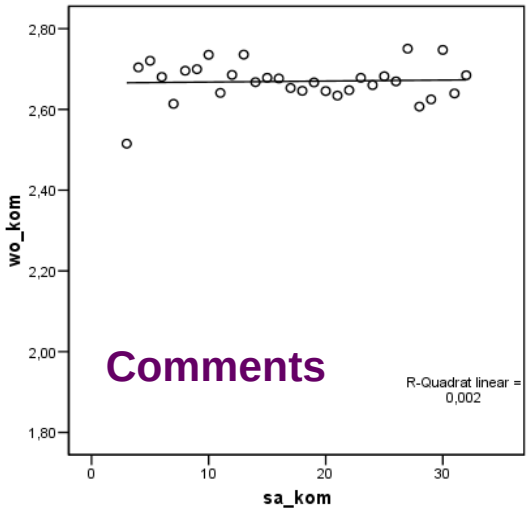


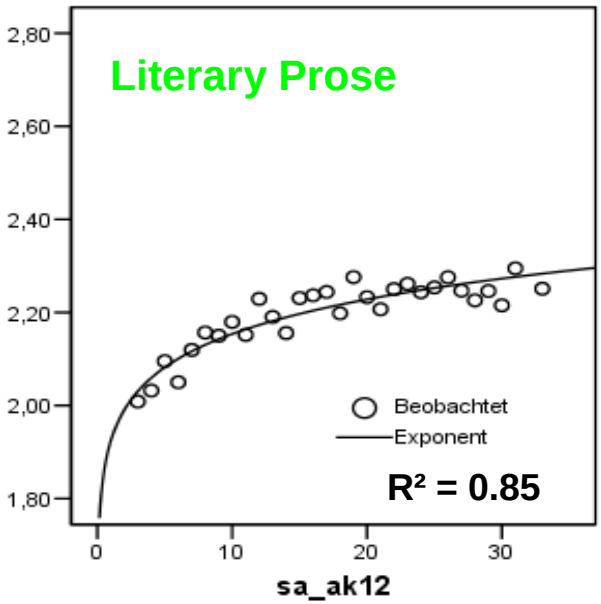
WL / SL Relation in Different Text Types

(Restricted Conditions)







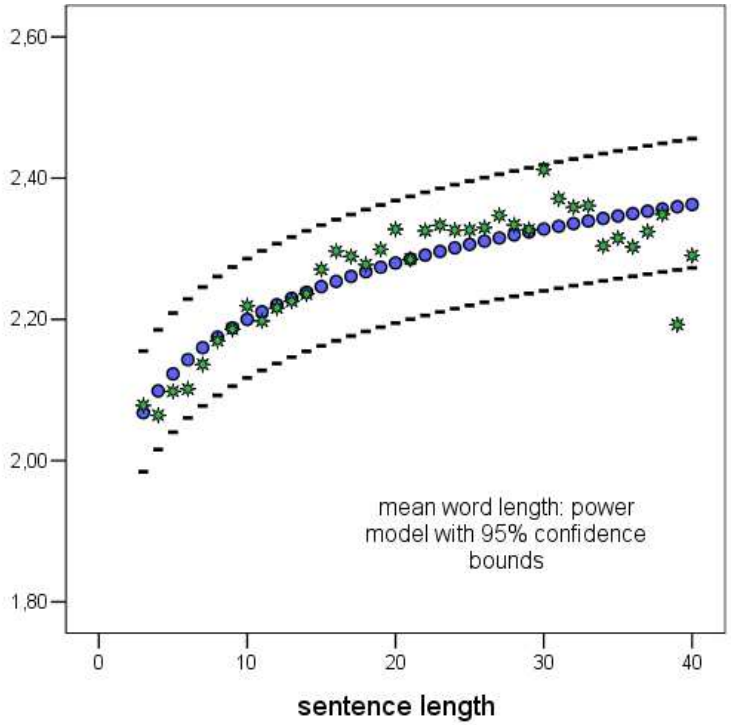


The SL / WL Relation in the Complete Corpus

(Restricted Conditions)

$y = 1.95x^{0.05}$

$R^2 = 0.79$

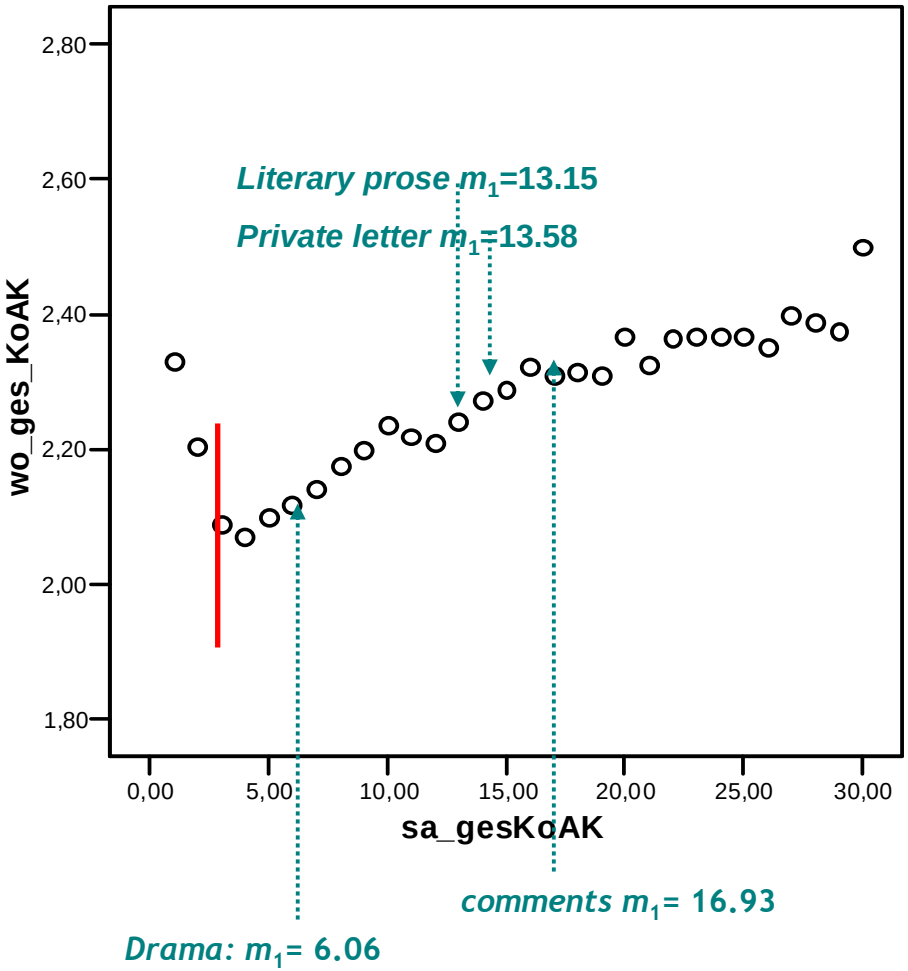


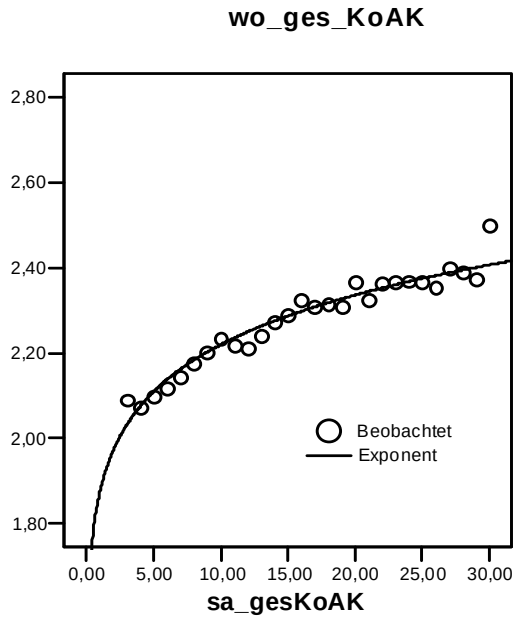
Is the non-linear curve type an artificial result of the dominance of literary prose texts ?

A Heterogeneous Sub-Corpus without literary texts

Drama + Letters + Comments

(17859 sentences, 167444 words)





$y = 1.87x^{0.07}$

$R^2 = 0.94$

Major Results

Inter-Textual vs. Intra-Textual Analysis

Inter-Textual Perspective

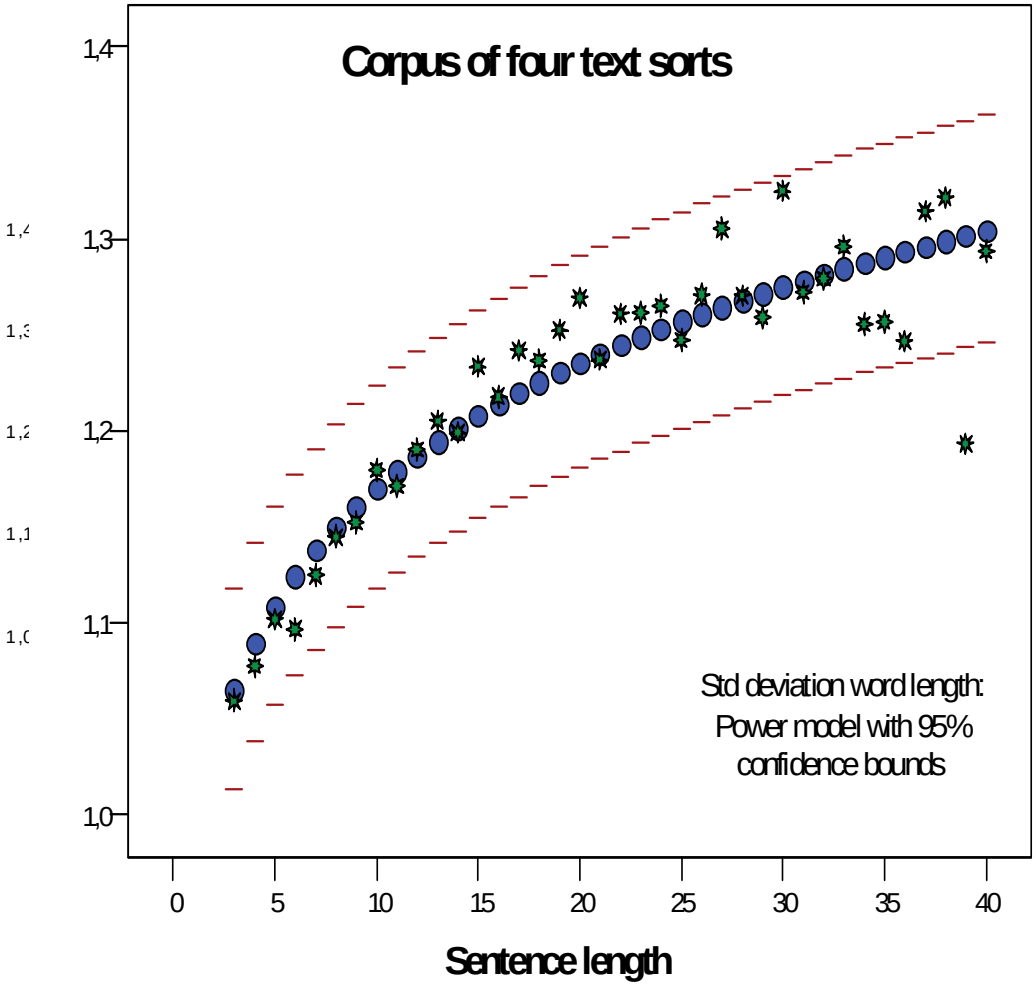
- Possibly slight (linear or non-linear) dependence of WL on SL

Intra-Textual Perspective (Corpus and Homogeneous Sub-Corpora)

- High variability of sparse data (points) (class frequency ≤ 30)
- Specific structure of very short sentences

- **Impact of Text Types:**
  - Linear or non-linear WL / SL relation (second-order Menzerath)
  - Non-linear WL / SL for heterogeneous text material

Standard Deviation of Word Length vs Sentence Length



$y = 0.98x^{0.08}$

$R^2 = 0.86$

Standard Deviation of Word Length vs Sentence Length