

Words and Numbers

**In Memory of Peter Grzybek
(1957-2019)**

Editors

Emmerich Kelih

Reinhard Köhler

RAM-Verlag

2020

Words and Numbers

In Memory of Peter Grzybek (1957-2019)

Editors

Emmerich Kelih

Reinhard Köhler

© Copyright 2020 by RAM-Verlag

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
Germany
RAM-Verlag@t-online.de
<http://ram-verlag.eu>

ISBN: 978-3-942303-89-7



Peter Grzybek
(1957-2019)

Contents

Foreword	
<i>Emmerich Kelih, Reinhard Köhler</i>	1
Sentence Length and Word Length: Testing Arens' Law and Menzerath-Altmann's Law	
<i>Lu Wang, Chengcheng Ren, Yahui Guo</i>	2
Sentence Lengths in Ukrainian: From Words to Finite Verbs via Autosemantics	
<i>Solomija Buk, Andrij Rovenchak</i>	16
Evolution of Chinese Word Length Motifs	
<i>Heng Chen, Haitao Liu, Gabriel Altmann</i>	27
Language Identification by Simple Character Profiles	
<i>Eric S. Wheeler, with Sheila Embleton, Dorin Uritescu</i>	44
A Classification of the Celtic Languages Based on Grapheme Frequencies	
<i>Andrew Wilson, Ján Mačutek</i>	53
Finding the Author of a Translation. An Experiment in Authorship Attribution Using Machine Learning Methods in Original Texts and Translations of the Same Author	
<i>George Mikros</i>	69
Some Aspects of a Sign Language Quantitative Analysis	
<i>Jiří Langer, Jan Andres, Martina Benešová, Dan Faltýnek</i>	81
Adnominal Valency in Modern Russian	
<i>Sergey Andreev</i>	104
A Model of Clause Properties and the Zipf-Alekseev Function	
<i>Haruko Sanada, Gabriel Altmann</i>	120
The Flexibility of Parts-of-Speech Systems and Their Grammar Efficiency	
<i>Relja Vulanović and Tayebah Mosavi Miangah</i>	129
Intrinsic Intentionality and Linguistic Meaning: An Historical Outline	
<i>Hermann Moisl</i>	148

Theoretical Thoughts and Practical Advice on the Length of Shots by Early Soviet Film Directors <i>Veronika Schmidt</i>	167
Fitting the Menzerath-Altmann Law: How Much Data Do You Need? <i>Andrei Beliankou, Reinhard Köhler</i>	178
Sentence and Paragraph in the Light of Menzerath-Altmann's Law <i>Volker Gröller</i>	194
The Impressive Story of a Unique Collaboration <i>Ernst Stadlober</i>	200
The Peter Grzybek Memorial Archive of Slavic Studies Publications <i>Emmerich Kelih, Hermann Moisl</i>	216
Commented Bibliography of Peter Grzybek (Or a Short Contribution to His Impact on Quantitative Text Analysis) <i>Emmerich Kelih, Veronika Schmidt</i>	219

Foreword

Writing a foreword for a memorial volume seems to be much harder than writing one for a Festschrift. No Festschrift had been dedicated to Peter Grzybek during his lifetime, since he had only passed his 60th year of life. Moreover, Peter wasn't a friend of more or less reflected academic tradition and he viewed the fashion for Festschriften with a certain degree of scepticism.

However, we believe: honour where honour is due and memory where memory is due. We therefore decided to publish a memorial volume for Peter.

The idea for this was born quite quickly after his passing away and we decided to ask his close academic colleagues and friends to honour Peter in this way. Many of the published articles are very close to Peter's own interests, whereas while some of them are about quantitative text analysis and quantitative linguistics generally. Both areas were without any doubt the most focussed research areas in the last decades of his life. Karl-Heinz Best apologized for not contributing to this *Gedenkschrift* due to health reasons, but he expressed his deepest sympathy for Peter and his close relation to his works in quantitative linguistics.

The publication of this memorial volume was financially supported by the Faculty of Philological and Cultural Studies of the University of Vienna, for which we would like to express our gratitude. We also have to acknowledge the professional and timely submission and resubmission of the articles and proofs by the authors. Many thanks also go to Jutta Altmann (RAM-Verlag) from the publishing house for the uncomplicated publication of the volume and to Daniel Ross for his thorough improvement of the English of articles written by non-native speakers.

Vienna, Trier in autumn 2020

Emmerich Kelih & Reinhard Köhler

Sentence Length and Word Length: Testing Arens' Law and Menzerath-Altmann's Law

Lu Wang^{1,2}, Chengcheng Ren², Yahui Guo³

Abstract

The present study focuses on the relationship between sentence length (SL) and word length (WL) and tests two hypotheses. Firstly, according to Arens' law, we assume that longer mean SL per text associates with longer mean WL per text. Due to the small size of our data, the results show weak correlation. Secondly, according to Menzerath-Altmann's law, the length of the construct performs a shortening effect on the length of its direct components. Since clause length (CL) is in between, the formula of the indirect relationship SL-WL is derived from the direct SL-CL (monotonical shortening effect) and CL-WL (shortening effect overlaps increasing effect) relationships. We assume that SL-WL abides by the derived formula which implies the relationship includes both decrease and increase effects, and fitting results support the assumption.

Keywords: Arens' law; Menzerath-Altmann's law; sentence length; word length

1. Introduction

The study concerning linguistic units and the components originated from Menzerath's (1928, 1954) phonetic research, which discovered a negative effect from syllable to sounds: the longer a syllable, the shorter its sounds. Later, Altmann (1980, 1983) extended Menzerath's findings, which is now known as Menzerath's law or Menzerath-Altmann's law, i.e. "The longer a language construct the shorter its components (constituents)." (Altmann, 1980) Further, he formalized the linguistic relation into the mathematic functions:

$$y = ax^b e^{-cx} \quad (1)$$

$$y = ax^b \text{ (when } c = 0) \quad (2)$$

$$y = ae^{-cx} \text{ (when } b = 0) \quad (3)$$

where x stands for the length of the construct, y the mean length of the components, and a , b , c are parameters. Formula (1) is the general form while Formula (2) and (3) are special forms when $c = 0$ and $b = 0$ respectively.

Altmann (1983) emphasized, Menzerath-Altmann's law is only applicable to a given construct and its direct components. Therefore, the relationship will not necessarily be negative when a level in a multi-level system is skipped. Take sentence length (SL) and word length (WL) as an example, sentence directly consists of clauses, therefore the longer the sentences, the shorter the clauses; while clause directly consists of words, thus the longer the clauses, the

¹ Computational Linguistics and Digital Humanities, University of Trier; Germany and School of Foreign Languages, Dalian Maritime University, Dalian, China, wanglu-chn@hotmail.com.

² School of Foreign Languages, Dalian Maritime University, Dalian, China.

³ Handan Vocational Center, Handan, China.

shorter the words or the shorter the clauses, the longer words. Thus, we conclude: longer sentences result in longer words. Does it hold true? It was in Arens' (1965) investigation, 117 German literary prose were analyzed to study the relationship between SL and WL. The results show that, as SL increases WL also increases, i.e. longer sentences, longer words, which is called Arens' law.

In Arens' work SL was measured by the number of words, however, it was not taken as independent variable. Arens took the mean sentence length of each entire text as the independent variable and the mean word length of the entire text as the dependent variable. Therefore, Arens' finding is actually the relationship between mean sentence length and mean word length. Grzybek & Stadlober (2007), Grzybek et al (2007, 2008) and Grzybek (2010) mention Arens' method as inter-textual perspective, while Menzerath-Altmann's law which takes individual sentence length as independent variable is mentioned as intra-textual perspective.

The present study attempts to explore the relationship between sentence length and word length following both inter-textual and intra-textual perspectives by testing the following two hypotheses.

(1) According to Arens' law, there is a positive effect from the mean sentence length per text (\overline{SL}) to the mean word length per text (\overline{WL}).

Hypothesis 1: The longer the mean sentence length per text, the longer the mean word length per text, i.e.

$$\overline{WL} = a\overline{SL}^b . \quad (4)$$

(2) According to Menzerath-Altmann's Law, given SL measured in terms of clause number (= number of clauses in the sentence) and clause length (CL) measured in terms of word number,

$$CL = a_1SL^{b_1}e^{-c_1SL} \quad (5)$$

and WL measured in terms of syllable number,

$$WL = a_2CL^{b_2}e^{-c_2CL} \quad (6)$$

Substituting CL in formula (6) with formula (5),

$$WL = a_2a_1^{b_1}SL^{b_1b_2}e^{-c_1b_2SL-c_2a_1SL^{b_1}e^{-c_1SL}} \quad (7)$$

Since in direct relationships $b = 0$ is rarely seen, here we only assume $c_1 = 0$ and/or $c_2 = 0$. When $c_1 = 0$,

$$WL = a_2a_1^{b_1}SL^{b_1b_2}e^{-c_2a_1SL^{b_1}} \Rightarrow WL = aSL^b e^{-cSL^d} \quad (8)$$

When $c_2 = 0$, the formula is as same as Formula (1)

$$WL = a_2a_1^{b_1}SL^{b_1b_2}e^{-c_1b_2SL} \Rightarrow WL = aSL^b e^{-cSL} \quad (9)$$

When $c_1 = 0$ and $c_2 = 0$, the formula is as same as Formula (2)

$$WL = a_2a_1^{b_2}SL^{b_1b_2} \Rightarrow WL = aSL^b \quad (10)$$

Formula (5) and (6) are special forms of Formula (4), when $d = 0$ and when $d = 0, c = 0$ re-

spectively. Formula (10), as mentioned by Altmann (1983), is taken as a theoretical evidence for “longer sentences, longer words”.

Hypothesis 2: The word length measured by syllable number depends on the sentence length measured by word number, i.e.

$$WL = aSL^b e^{-cSL^d}. \quad (8)$$

2. Data

Ten academic articles (Table 1) are prepared to test the hypotheses. Those academic articles contain figures, tables, titles of figures and tables, references, which are not sentences and therefore excluded.

Table 1
Data description

	Title	Number of sentences	Number of tokens
Text 1	Text difficulty and the Arens-Altman law (Grzybek, 2010).	112	3768
Text 2	Do we have problems with Arens' law? A new look at the sentence-word relation (Grzybek & Stadlober, 2007).	84	2788
Text 3	The relationship of word length and sentence length: the inter-textual perspective (Grzybek, Stadlober & Kelih, 2007).	84	2368
Text 4	The relation between word length and sentence length: an intra-systemic perspective in the core data structure (Grzybek, Kelih & Stadlober, 2008).	107	3126
Text 5	Close and distant relatives of the sentence: some results from russian (Grzybek, 2013)	103	3619
Text 6	Introductory remarks: on the science of language in light of the language of science (Grzybek, 2006)	151	4466
Text 7	On the systematic and system-based study of grapheme frequencies: a re-analysis of German letter frequencies (Grzybek, 2007)	87	2545
Text 8	Quantitative text typology: the impact of word length (Grzybek, Stadlober, Kelih, Antić, 2005)	119	3004
Text 9	Slavic letter frequencies: a common discrete model and regular parameter behavior? (Grzybek, Kelih & Stadlober, 2009)	141	4387
Text 10	Regularities of Estonian proverb word length: frequencies, sequences, dependencies (Grzybek, 2014)	132	4172
Total		1120	34243

SL, measured by clause number, depends on the sum of main clauses and subordinate clauses (Example 1). CL/SL, measured by word number, depends on the sum of words in a clause/sentence, where numbers, symbols and formula (involving arithmetic and exponentiation) are transformed into words (Example 2 & 3) while in-text citation and contents in brack-

ets (Example 3 & 4) are excluded. If a formula involves more than arithmetic and exponentiation, the whole sentence is excluded (Example 5). WL, measured by syllable number, depends on the sum of syllables of the word (Example 6).

Example 1:

- Sentence: This view contains, of course, no theoretical foundation as to the question which specific factors influence text difficulty in what way or to what degree; yet it offers a theoretically based post-hoc answer to the question why the reduction to only a couple of seemingly elementary factors has made this concept to have such a success story.
- Clause 1: This view contains, of course, no theoretical foundation as to the question
- Clause 2: which specific factors influence text difficulty in what way or to what degree;
- Clause 3: yet it offers a theoretically based post-hoc answer to the question
- Clause 4: why the reduction to only a couple of seemingly elementary factors has made this concept to have such a success story.

Example 2:

- Sentence: With condition I = 32 – used for our re-analysis of the parameters below – 21 texts showed $C > 0.001$ and for the remaining rest $C \leq 0.002$ was obtained.
- Clause 1: With condition I equal to thirty-two – used for our re-analysis of the parameters below – twenty-one texts showed C greater than zero point zero zero one and for the remaining rest C less than zero point zero zero two was obtained.

Example 3:

- Sentence: The regression lines follow the equation $y = b + ax$ (that is, in our case, $M = b + aK$); here, b is a constant determining the regression intercept, and a is the regression coefficient which determines the steepness for the rise or decline of the line.
- Clause 1: The regression lines follow the equation y equal to b plus a times x ;
- Clause 2: here, b is a constant determining the regression intercept,
- Clause 3: and a is the regression coefficient
- Clause 4: which determines the steepness for the rise or decline of the line.

Example 4:

Sentence: It seems reasonable to start from this inter-textual end, tentatively maintaining Altmann's (1983: 32) assumption as to less variance across samples than for individual texts, consequently predicting even worse results for individual texts (i.e., for the intra-textual situation).

Clause 1: It seems reasonable to start from this inter-textual end, tentatively maintaining Altmann's assumption as to less variance across samples than for individual texts, consequently predicting even worse results for individual texts.

Example 5:

Sentence: For linear relations, this can be done by reference to a t-test statistic

$$t = \frac{|b_1 - b_2|}{\sqrt{\frac{s_{y1.x1}^2 \cdot (n_1 - 2) + s_{y2.x2}^2 \cdot (n_2 - 2)}{n_1 + n_2 - 4} \cdot \left(\frac{1}{Q_{x1}} + \frac{1}{Q_{x2}}\right)}}$$

with $DF = n_1 + n_2 - 4$ degrees of freedom and $Q_x = \sum(x - \bar{x})^2$.

Example 6:

Words:	Th	regres-	line	fol-	th	equa-	y	equa	t	b	plu	a	time	x
	e	sion	s	low	e	tion		l	o	s	s			
Sylla- bles:	1	3	1	2	1	3	1	2	1	1	1	1	1	1

3. Results and discussion

3.1. Testing Hypothesis 1: \overline{SL} vs. \overline{WL}

The fitting result of \overline{SL} - \overline{WL} relationship is shown in Table 2 and Figure 1. Both the linear and non-linear models show weak correlation with $R^2 = 0.2934$ and $R^2 = 0.3021$ respectively. The fact that we used a small corpus, only ten data points, might be a disadvantage. However, large corpus also reports unsatisfied results: Grzybek, Stadlober & Kelih (2007) analyzed Russian literary prose *Anna Karenina*, converted its 239 chapters to 239 data points and finally obtained $R^2 = 0.15$ for linear model and $R^2 = 0.18$ for power model. The models predict a negative relation between \overline{SL} and \overline{WL} , which according to Arens' law should be a positive one. But given the unsatisfied goodness-of-fit and the scattered points, we do not consider the data reflects any tendency.

Table 2
Mean sentence length per text and mean word length per text

	\bar{SL}	\bar{WL}	$\bar{WL} = a\bar{SL} + b$	$\bar{WL} = a\bar{SL}^b$
Text 8	25.24	1.8492	a = 2.0398	a = 2.9742
Text 3	28.19	1.7082	b = -0.0088	b = -0.1518
Text 4	29.21	1.7668	R ² = 0.2934	R ² = 0.3021
Text 7	29.25	1.8255		
Text 6	29.57	1.7886		
Text 9	31.11	1.7835		
Text 10	31.6	1.762		
Text 2	33.19	1.7052		
Text 1	33.64	1.7893		
Text 5	35.13	1.7251		

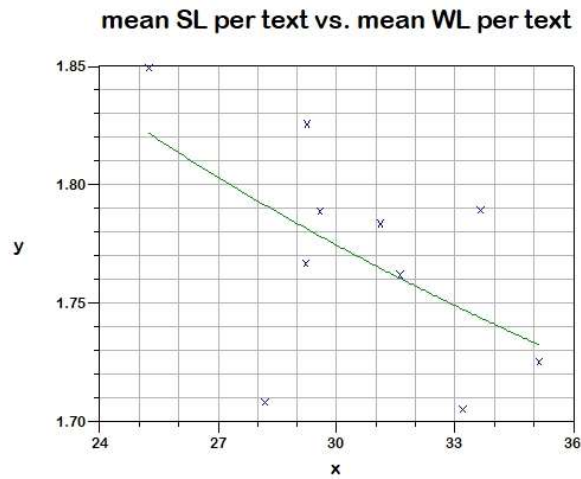


Figure 1 Fitting the data to Arens' law (power function)

3.2. Testing Hypothesis 2: SL vs. WL

The SL-CL data illustrates a clearly decreasing tendency, implying a shortening effect from SL on CL. Formula (1) ~ (3) are all applicable. Formula (1) and (2) obtain similar goodness-of-fit values, both better than Formula (3). Considering less parameter is preferred, we choose Formula (2) with $c = 0$, as shown in Table 3 and Figure 2. Considering Formula (7), we confirm $b_1 \neq 0$ and $c_1 = 0$.

Table 3
Fitting Menzerath-Altmann's law to SL-CL data

SL	CL
1	20.6
2	14.63
3	13.17
4	11.92
5	10.82
6	10.54

a = 20.2161
b = -0.3874
R² = 0.9852

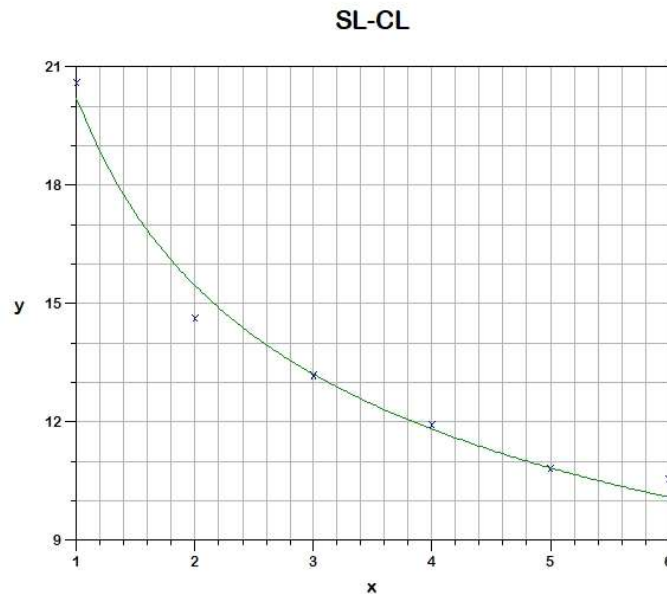


Figure 2 Fitting Menzerath-Altmann's law to SL-CL data

To our surprise, the CL-WL relationship is not negative (Figure 3). The data points form a convex curve, which means besides the shortening effect from CL, there exists another factor against, hence stronger than, the shortening. The general form of Menzerath-Altmann's law, Formula (1), is exactly appropriate to capture it, as in this formula e^{-cx} expresses the shortening effect from the construct and x^b stands for the opposite effect (Altmann 1983; Cramer, 2005). From the viewpoint of synergetic linguistics, this opposite effect may be a regulation from the language system, since WL can not be shortened infinitely (Köhler, 1986). The fitting results are shown in Table 4 and Figure 3. Despite $R^2 = 0.7624 < 0.8$, the tendency of the data confirms that both the shortening effect and an opposite effect exist, thus $b_2 \neq 0$ $c_2 \neq 0$.

Table 4
Fitting Menzerath-Altmann's law to CL-WL data

CL	WL	CL	WL	Formula (1)	Formula (2)
3	1.52	19	1.76	a = 1.3167	a = 1.4775
4	1.44	20	1.8	b = 0.1453	b = 0.0633
5	1.65	21	1.79	c = 0.006	R ² = 0.6648
6	1.68	22	1.74	R ² = 0.7624	
7	1.74	23	1.83		
8	1.71	24	1.74		
9	1.76	25	1.81		
10	1.75	26	1.82		
11	1.78	27	1.82		
12	1.77	28	1.76		
13	1.77	29	1.78		
14	1.77	30	1.9		
15	1.76	31	1.83		
16	1.81	32	1.83		
17	1.78	33	1.79		
18	1.78				

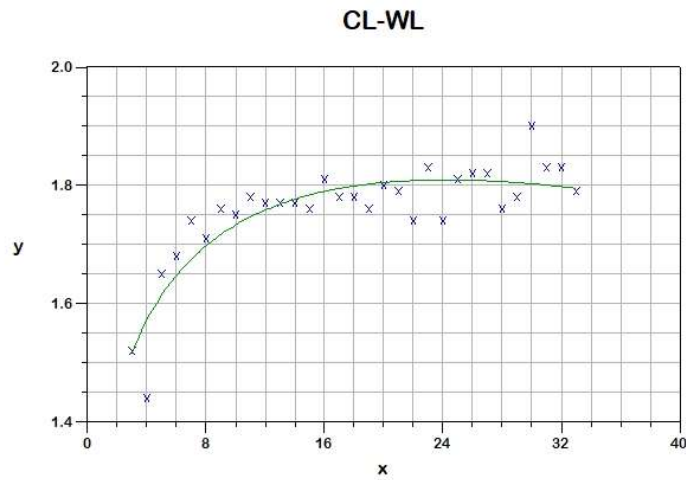


Figure 3 Fitting Formula (1) to CL-WL data

The indirect relationship of SL-WL is derived from Menzerath-Altmann's law. As shown by Formula (8), the general form includes six parameters, if the direct relationships both involve all the three parameters a , b and c as Formula (1). Given $c_1 = 0$, $b_2 \neq 0$ and $c_2 \neq 0$, Formula (8) with four parameters a , b , c , and d should be the best model for SL-WL. Formula (9) & (10) are also fitted to see the difference. From the results in Table 5 and Figure 4, we can see that each formula obtains unsatisfied goodness-of-fit. However, given such fluctuation of data points, it is difficult to find a model.

Table 5
Fitting Menzerath-Altmann's law to SL-WL data

SL	WL	SL	WL	SL	WL	Formula (10) ($c_1 = 0, c_2 = 0$)	Formula (9) ($c_1 \neq 0, c_2 = 0$)	Formula (8) ($c_1 = 0, c_2 \neq 0$)
10	1.82	24	1.75	38	1.78	a = 1.9859	a = 2.097	a = 2.3021
11	1.95	25	1.78	39	1.77	b = -0.0318	b = -0.0573	b = 0.0318
12	1.76	26	1.83	40	1.76	R ² = 0.2857	c = -0.0009	c = 0.1477
13	1.83	27	1.78	41	1.72		R ² = 0.2937	d = 0.000007
14	1.8	28	1.81	42	1.69			R ² = 0.2857
15	1.82	29	1.74	43	1.76			
16	1.83	30	1.82	44	1.84			
17	1.83	31	1.79	45	1.72			
18	1.81	32	1.81	46	1.83			
19	1.81	33	1.74	47	1.79			
20	1.81	34	1.72	48	1.81			
21	1.83	35	1.72	49	1.73			
22	1.75	36	1.77	50	1.75			
23	1.86	37	1.76					

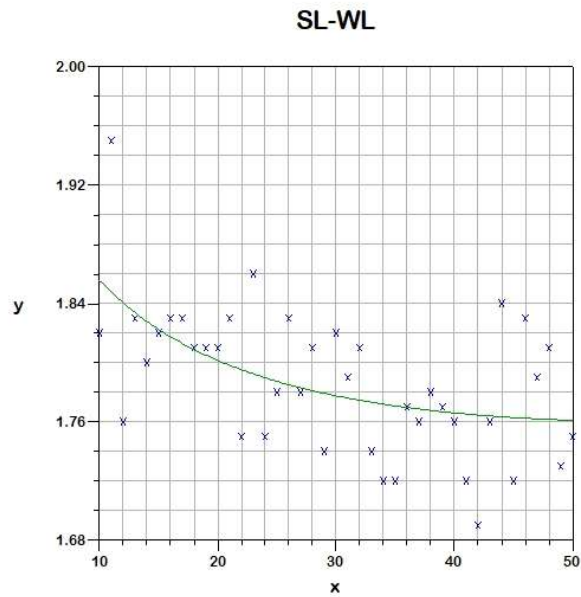


Figure 4 Fitting Formula (1) to SL-WL data

To reveal the SL-WL tendency, we adopt moving average method. The smoothed data abides by Menzerath-Altmann's law with $R^2 = 0.8847$ (Figure 5 and Table 6). The curve shows a clearly decreasing tendency, different from the prediction of the law.

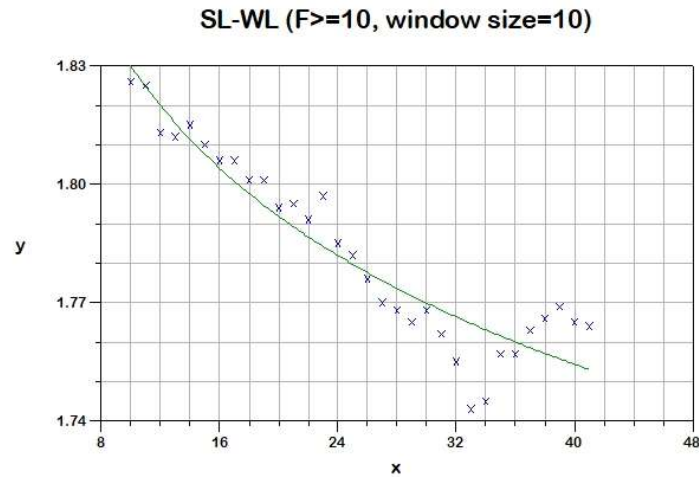


Figure 5 Moving average (window size = 10)

Generally, in order to avoid unreliability, only data points with more than 10 observations are taken into consideration. But this may cause another “unreliability”, especially when we face with sparse linguistic data. It is not likely to obtain large amount of extremely long sentences as well as extremely short sentences. Inevitably, we will lose the head and the tail part of the complete curve, i.e. we will see merely part of the truth rather than the whole. As Figure 5 shows, SL-WL forms a decreasing curve, however, if we scrutinize the whole curve with all the data points (Figure 6) and its moving average curve (Figure 7), we find concave shapes. This fact, despite surprising, is in accordance with the explanation for the direct relations of Menzerath-Altman’s law that the shortening and the opposite effects co-exist (Altmann, 1983). From Figure 7, the shortening effect dominates in initial stage causing WL decreases, then around SL=60 the opposite effect shows stronger force and WL grows.

Why SL-WL shows neither positive relation as the law predicts nor monotonically negative relation as in its partial curve in Figure 5? Considering the limitation of SL and WL, it is possible to create extremely long or short sentences, but not possible to make that long words or zero-syllable words. Consequently, WL will not increase or decrease infinitely; the amplitude of the change of WL must be restricted within some limits. Otherwise, the transmission security is not guaranteed, the coding and encoding effort (Köhler, 2005) increases too much, etc. and thus the stable state of the language system break. As a result, WL will fluctuate rather than monotonically decreasing. The fitting results of three different processed data are shown in Table 6 and Figure 8.

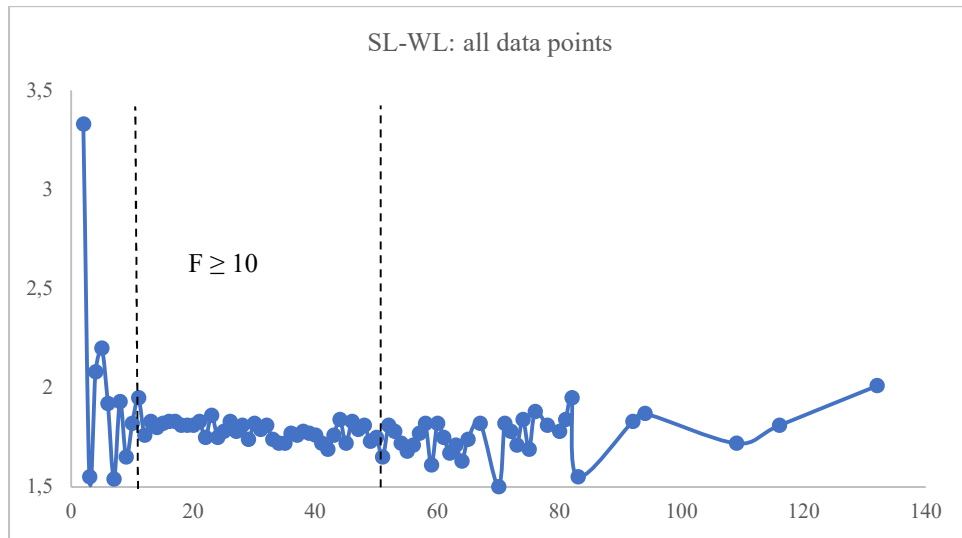


Figure 6 SL-WL data

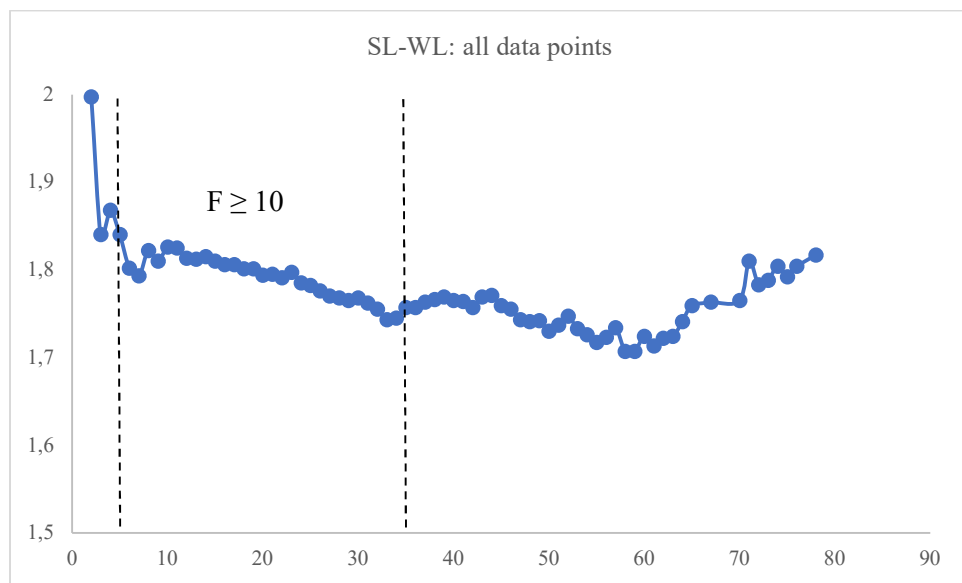


Figure 7 Moving average with window size = 10 (all data points)

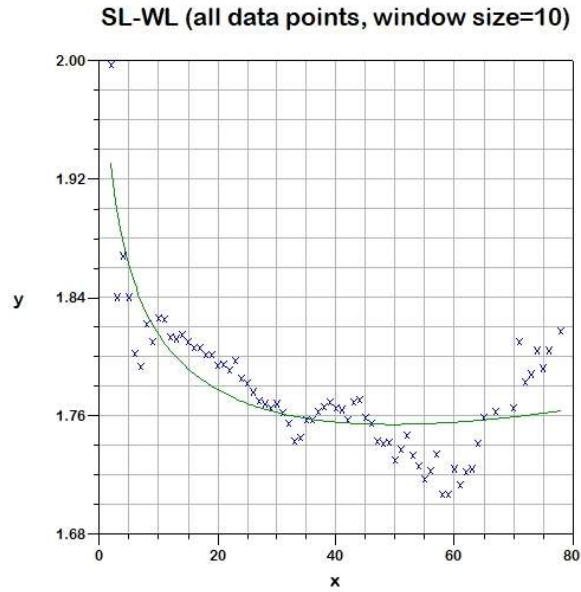


Figure 8 Fitting Menzerath-Altman's law to smoothed SL-WL data

Table 6
Fitting Menzerath-Altman's law to SL-WL data

	Formula (1)	Formula (2)	Formula (3)
All data points	a = 2.6332 b = -0.153 c = -0.0035 R ² = 0.3484	a = 2.1872 b = -0.0557 R ² = 0.185	a = 1.8669 c = 0.0008 R ² = 0.0405
Moving average of all data points	a = 1.985 b = 0.0425 c = -0.0008 R ² = 0.6677	a = 1.9153 b = -0.0223 R ² = 0.5765	a = 1.821 c = 0.0006 R ² = 0.3323
Moving average with F ≥ 10	a = 1.9773 b = -0.034 c = -0.0001 R ² = 0.8847	a = 1.9633 b = -0.0305 R ² = 0.8843	a = 1.8433 c = 0.0013 R ² = 0.8482

Purely from a mathematical point of view, if a text or a corpus contains n words (tokens), the mean word length of the text (\overline{WL}) is

$$\overline{WL} = \frac{\sum_{i=1}^n WL(i)}{n}.$$

Given a sentence, consisting of j words ($1 \leq j \leq n$), the mean word length of this sentence (\overline{wl}) is

$$\overline{wl} = \frac{\sum_{i=1}^j WL(i)}{j}.$$

When the sentence is long enough to approximate the whole text, i.e. $j \approx n$,

$$\lim_{j \rightarrow n} \overline{wl} = \lim_{j \rightarrow n} \frac{\sum_{i=1}^j WL(i)}{j} = \frac{\sum_{i=1}^n WL(i)}{n} = \overline{WL}.$$

A long sentence, with large amount and various words, would statistically more possible to reach a $\overline{wl} \approx \overline{WL}$, or even $\overline{wl} = \overline{WL}$. As for a short sentence, with few words or even one word, its \overline{wl} is more random and unlikely to approach the whole text's \overline{WL} . Regardless of a positive or negative difference, \overline{wl} of short sentences would locate farther from \overline{WL} in axis than the \overline{wl} of long sentences.

Yet a question: from the linguistic point of view, why should \overline{wl} approach \overline{WL} ? A possible reason is that, \overline{WL} can be seen as a compromise resulting from the influencing factors of the language system. The longer the sentence, the more its words affected by the factors, such as structural restriction, frequency, transmission security requirements etc., then finally becomes closer to \overline{WL} .

Our data did not reach the stable WL value, since the long sentences are neither long enough nor frequent enough. The longest contains 132 words with only 1 observation, which is too small compared with the whole corpus 34243 tokens. The authors are not native speakers might also be a reason for the lack of long sentences.

4. Conclusion

The present study tests two hypothesis, inter-textual and intra-textual, concerning sentence length and word length.

1. Arens' law assumes that, the longer the mean sentence length per text, the longer the mean word length per text. Due to the small size of the corpus which contains ten texts, the results show weak correlation on both linear and non-linear models.
2. The relationship of SL, measured by word number, and WL, measured by syllable number, as an indirect relationship, is derived from SL-CL and CL-WL according to Menzerath-Altmann's law. The original data shows fluctuation; while the smoothed data verifies that, the derived function $WL = aSL^b e^{-cSL}$, which is also the general form of the law, captures the SL-WL relationship. As Altmann (1983) explained, this formula implies the shortening effect from the structure (SL), and the opposite effect, which may come from the regulation of the language system, collaboratively affect WL. We only observed a decreasing-increasing tendency, however, the relation would not increase infinitely. The regulation effect of the language system will be triggered if WL grows over a limitation, vice versa. Thus, as SL grows, WL will fluctuate and gradually approach a stable value, i.e. the whole corpus' mean word length. This assumption needs to be tested by large amount of data in future investigations.

References

- Altmann, G.** (1980). Prolegomena to Menzerath's law. In: Grotjahn, R. (eds.), *Glottometrika 2: 1-10*. Bochum: Brockmeyer.
- Altmann, G.** (1983). H. Arens' "Verborgene Ordnung" und das Menzerathsche Gesetz. In: Faust, M. (ed), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik: 31-39*. Tübingen: Narr.
- Arens, H.** (1965). *Verborgene Ordnung: die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute*. Düsseldorf: Pädagogischer Verlag Schwann.
- Cramer, I.** (2005). Das Menzerathsche Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 659-688*. Berlin: de Gruyter.
- Grzybek, P.** (2006). Introductory Remarks: On the Science of Language in Light of the Language of Science. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues: 1-14*. Dordrecht, NL: Springer.
- Grzybek, P.** (2007). On the Systematic and System-based Study of Grapheme Frequencies. A Re-analysis of German Letter Frequencies. *Glottometrics 15*, 82-91.
- Grzybek, P.** (2010). Text difficulty and the Arens-Altman law. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language: Structures, Functions, Interrelations, Quantitative Perspectives: 57-70*. Wien: Praesens.
- Grzybek, P.** (2013). Close and Distant Relatives of the Sentence: Some Results from Russian. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics: 44-58*. Belgrade: Academic Mind.
- Grzybek, P.** (2014). Regularities of Estonian Proverb Word Length: Frequencies, Sequences, Dependencies. In: Baran, A., Laineste, L., Voolaid, P. (eds.), *Scala Naturae. Festschrift in Honour of Arvo Krikman: 121-148*. Tartu: ELM Scholarly Press.
- Grzybek, P., Kelih, E. & Stadlober, E.** (2008). The Relation between Word Length and Sentence Length: an Intra-systemic Perspective in the Core Data Structure. *Glottometrics 16*, 111-121.
- Grzybek, P., Kelih, E. & Stadlober, E.** (2009). Slavic Letter Frequencies: A Common Discrete Model and Regular Parameter Behavior? In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 17-33*. Lüdenscheid: RAM.
- Grzybek, P. & Stadlober, E.** (2007): Do We Have Problems with Arens' law? A new Look at the Sentence-Word Relation. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: Dedicated to Professor Gabriel Altmann on the Occasion of His 75th Birthday: 205-218*. Berlin: De Gruyter.
- Grzybek, P., Stadlober, E. & Kelih, E.** (2007). The Relation of Word Length and Sentence Length: The Inter-Textual Perspective. In: Decker, R., Lenz, H.-J. (eds.), *Advances in Data Analysis: 611-618*. Berlin: Springer.
- Grzybek, P., Stadlober, E., Kelih, E. & Antić, G.** (2009). Quantitative Text Typology: The Impact of Word Length In: Weihs, C., Gaul, W. (eds.), *Classification. The Ubiquitous Challenge: 53-64*. Heidelberg, New York: Springer.
- Köhler, R.** (1984). Zur Interpretation des Menzerathschen Gesetzes. *Glottometrika 6*, 177-183.
- Köhler, R.** (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.
- Köhler, R.** (2005). Synergetic linguistics. In: R. Köhler, G. Altmann & G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin: de Gruyter.
- Menzerath, P.** (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.

Sentence Lengths in Ukrainian: From Words to Finite Verbs via Autosemantics

Solomija Buk¹, Andrij Rovenchak²

Abstract

We report the results of a sentence length analysis in Ukrainian long prose texts by Ivan Franko. The distributions of lengths measured in several units are fitted using continuous models. The simplest unit considered is an orthographic word, next we analyze the number of autosemantic (meaningful) parts of speech per sentence, and finally count the number of finite (non-infinitive) verbal forms. Alongside the log-normal distribution, a proper fitting is achieved with function $x^{-a-1} e^{-bx-c/x}$, especially in the domain of small length values.

Keywords: sentence length; Ukrainian language; text corpus; Ivan Franko

1. Introduction

We first established contact with Peter Grzybek in the mid-2000s and met in person during the conference “Modern Methods in Linguistics” held in Budmerice (Slovakia) in October 2006. His helpful comments and suggestions allowed in particular the development of an approach to model the word–clause–sentence level of the Menzerath–Altmann law for Ukrainian texts (Buk & Rovenchak 2008).

One of the problems related to the Menzerath–Altmann law is the analysis of sentence lengths (cf. Sanada 2016). It is also linked to authorship/genre attribution (Yule 1939, Sichel 1974, Kelih et al. 2006) and to studies on syntactic complexity (Khany & Kafshgar 2016) and syntactic hierarchy depth (Yang 2019). Its spoken-language analog, utterance length, is used in studies of language disorders (Yaruss 1999) and psycholinguistics and psychiatry (Scarborough et al. 1991, Özcan & Kuruoğlu 2018). Such lengths can be expressed in a variety of linguistic units, ranging from syllables and morphemes to clauses.

In the present work, we focus on measuring sentence lengths using three different units. First, all orthographic words, i.e., alphanumeric sequences between two spaces and/or punctuation marks, are counted. Second, only words being autosemantic parts of speech are counted. These are meaningful parts of speech, namely nouns, verbs, adjectives, adverbs, numerals, and pronouns only. And finally, the third approach consists in counting finite verbs, i.e., all verb forms except for infinitives. This is often considered equivalent to counting clauses (cf. Wimmer et al. 2003, p. 162, Maillart & Parisse 2019) but is not the case for Ukrainian and other Eastern Slavonic languages (cf. Roukk 2007). Finiteness is a vaguely defined concept itself (Kalinina & Sumbatova 2007), and at least in the case of Indo-European languages it is linked

¹ Department for General Linguistics, Ivan Franko National University of Lviv, Ukraine, solomija@gmail.com.

² Department for Theoretical Physics, Ivan Franko National University of Lviv, Ukraine, andrij.rovenchak@gmail.com.

to verbal forms that differentiate person and number (Fortson 2011, p. 107) and some other grammatical categories (Finch 2000, p. 92-93), that is, exclude infinitives, participles, gerunds, etc. Our approach (Buk & Rovenchak 2008) included a special treatment of participles applied to the clause definition in Ukrainian.

In the subsequent chapters, we report results of sentence length studies based on the analysis of long prose fiction by Ivan Franko, a Ukrainian writer, scholar, and public figure, using a tagged text corpus of his works (Buk 2013, Rovenchak & Buk 2018).

2. Sources

The complete set of Ukrainian long prose fiction works by Ivan Franko (Pastukh 1996; Denysiuk 2008) except the *Petriji j Dovbuščuky*, 1st edition (1875–76) was under consideration. The last-mentioned novel was not included in the analysis because of the big influence of Church Slavonic language on it. The author himself considered the language of that edition of the novel as “unprocessed and imperfect”.

The list of titles is:

- BC1 *Boa constrictor*, 1st edition (1878–84);
- BC2 *Boa constrictor*, 2nd edition (1905–07);
- BSm *Boryslav smijetsja* [*Boryslav Laughs*] (1880–81);
- DDO *Dlja domašnjoho ohnyšča* [*For the Hearth*] (1892);
- NSB *Ne spytavšy brodu* [*Without Asking a Wade*] (1885–86);
- OSu *Osnovy suspil'nosti* [*Pillars of Society*] (1894–95);
- PD2 *Petriji j Dovbuščuky*, 2nd edition (1909–12);
- PSt *Perekhresni stežky* [*The Cross-paths*] (1900);
- VSh *Velykyj šum* [*The Great Noise*] (1907);
- ZBe *Zakhar Berkut* (1883).

To make some comparisons, we also included the following two texts classified as short prose:

- chu *Čuma* [*Plague*] (1889);
- pan *Pantalakha* (1902).

Note that both these short prose texts were first published in Polish and later appeared as self-translations in Ukrainian, with some modifications and additions. They are a part of the respective parallel corpus of Ivan Franko’s self-translations (Buk 2012).

While usually ends of sentences are marked by a period, an exclamation or question mark, or an ellipsis (‘.’, ‘!’, ‘?’, and ‘...’), we also consider punctuation marks separating direct and author’s speech (semicolon ‘:’ and comma ‘,’) as sentences boundaries (Martin et al. 2003, Buk et al. 2019).

As observed previously (Buk et al. 2019), several discrete models yield satisfactory fits for sentence lengths (with the determination coefficient $R^2 > 0.96$), namely, Jackson–Nickols, extended positive negative binomial, hyper-Pascal, Consul–Jain–Poisson, as well as mixed negative binomial distributions (Wimmer & Altmann 1999). It is another confirmation of previous results on sentence length from several authors reporting Poissonian-type behavior (Sichel 1974; Ishida & Ishida 2007; Pande & Dhami 2015). In the following chapter we focus on continuous models instead (cf. Mačutek 2007, Mačutek & Altmann 2007). While there is no special reason to prefer either discrete or continuous approaches, we still believe that – in some cases – a

continuous dependence can catch intrinsic relations in a more straightforward fashion, especially when moving between disciplines.

3. Raw data

In Figures 1–4, dependencies of sentence lengths in various units described in the Introduction are presented: all words, autosemantic words, and finite verbs. P_j represents the fraction of sentences having length j (i.e., containing j respective units). Note that while the minimum length of a sentence measured in all words is one, zero lengths are possible when measuring in autosemantic words or finite verbs.

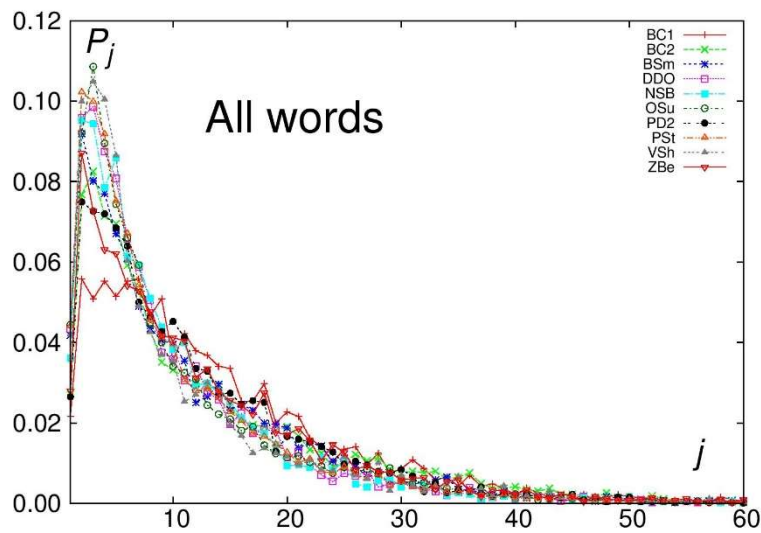


Figure 1 Sentence lengths measured in all words

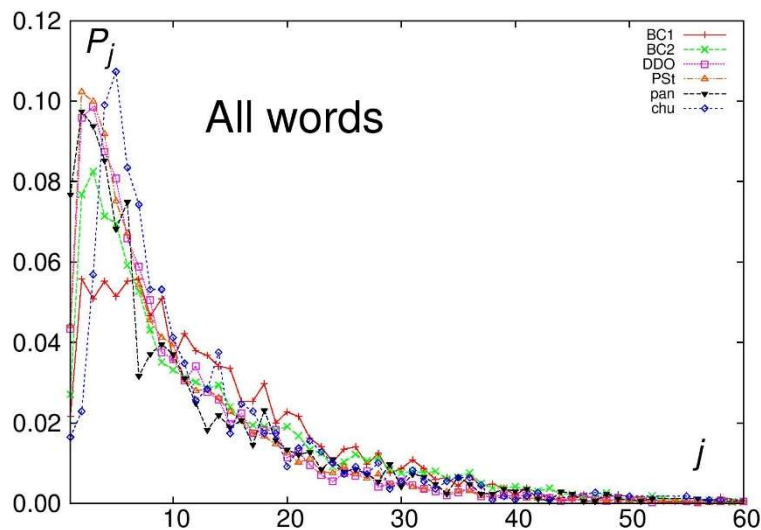


Figure 2 Sentence lengths measured in all words for some long prose and two short texts for comparison

From Figure 1 it might seem that the length distribution for the shortest text – BC1 – has a shape significantly different from other texts. To check whether this is the effect of text size we have also analyzed two short texts (see Figure 2). As one can see, no qualitative difference is observed between longer and shorter text, so the BC1 behavior is just specific for this text only.

As we have mentioned in the Introduction, autosemantic parts of speech are the following: nouns, verbs, adjectives, adverbs, numerals, and pronouns. There is also a specific class of words in Ukrainian called *присудкові слова* ‘predicative words’, which is not given a part-of-speech mark in dictionaries. By the syntactic role, such words are close to verbs, so we counted them as autosemantic parts of speech as well (cf. Buk et al. 2019). Examples include, in particular, *треба* ‘ \approx it is needed’, *школа* ‘ \approx it is a pity’, *нема* ‘ \approx there is no’, and some others.

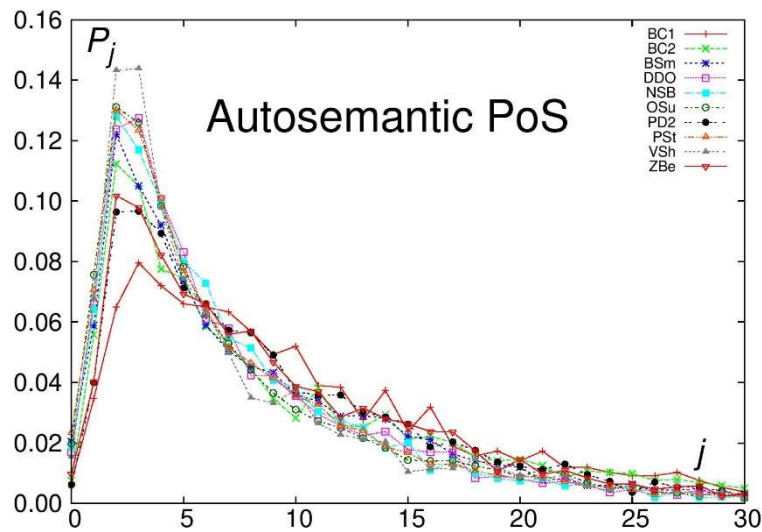


Figure 3 Sentence lengths measured in autosemantic words

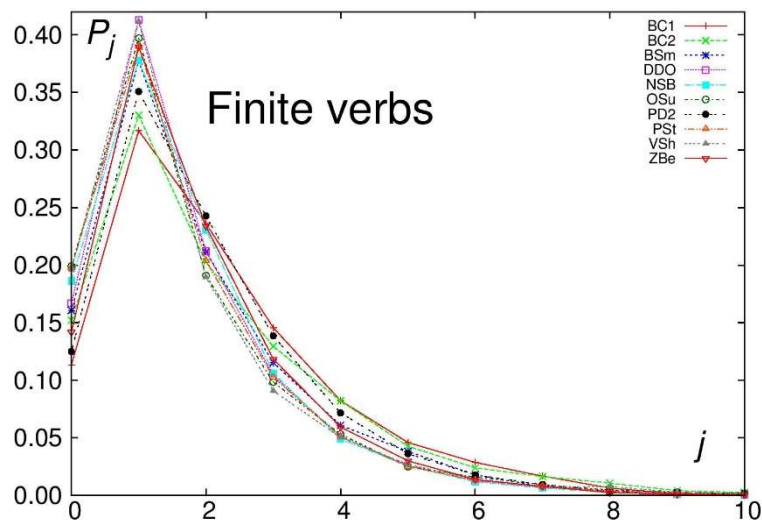


Figure 4 Sentence lengths measured in finite verbs

Counting finite (non-infinitive) verbal forms in sentences is linked to counting clauses. The latter is, however, a more complicated task as verbless clauses are not uncommon – not only

with incomplete sentences – in various languages across the globe (Miller 1999, Idiátov 2010, Landolfi et al. 2010, Aikhenvald 2018). As can be seen from Figure 4, the shapes of dependences describing sentence lengths measured in finite verbs are more uniform compared to those obtained using the previous two considered units, all words and autosemantic words. Therefore, one can expect that parameters of sentence length distributions measured in finite verbs can be utilized in author/genre attribution.

4. Models

The first continuous model we tested for sentence lengths was the log-normal distribution known to yield a good description, at least for the number of words per sentence (William 1940, Limpert et al. 2001). For the log-normal distribution, the probability density function is given by:

$$P_x = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}, \quad (1)$$

where μ and σ are, respectively, the mean value and standard deviation of the variable's natural logarithm $\ln x$ (Johnson et al. 1994). In the case of autosemantic PoS and finite verbs, the argument in (1) was shifted by unity to avoid singularities at $x = 0$.

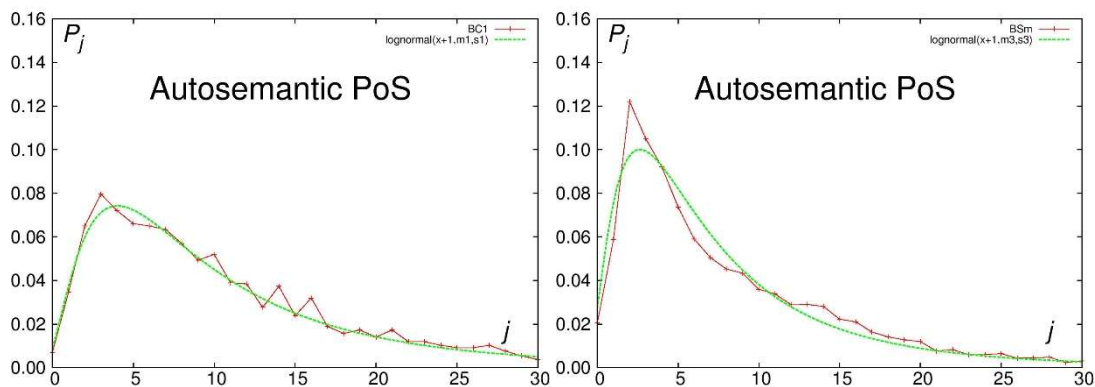


Figure 5 Log-normal fits for sentence length measured in autosemantic words for BC1 (left panel) and BSm (right panel).

Fitting parameters for the log-normal distribution are shown in Table 1. The values of the determination coefficient R^2 in all cases are quite close to unity. In particular, the best fits are achieved for sentence lengths measured in finite verbs yielding $R^2 > 0.99$. For lengths in all words, R^2 values are greater than 0.96. Finally, the fits for the number of autosemantic PoS per sentence are only slightly worse than for all words, with the lowest $R^2 = 0.945$ in the case of BC2.

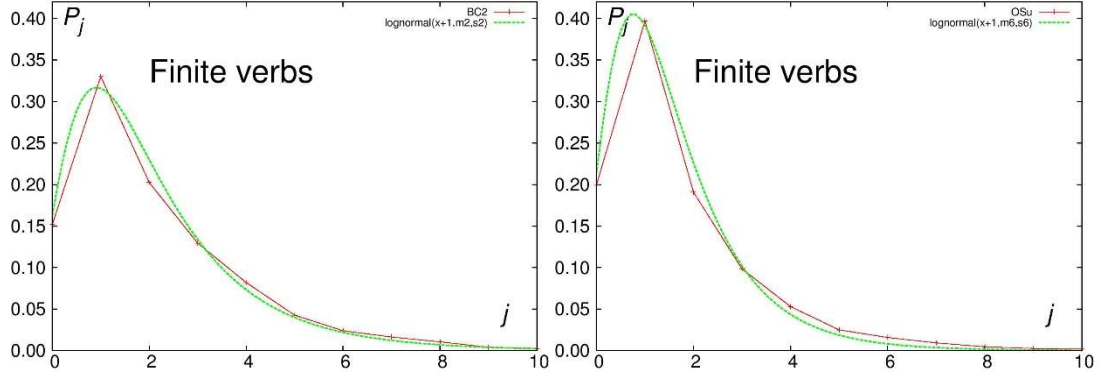


Figure 6 Log-normal fits for sentence length measured in finite verbs for BC2 (left panel) and OSu (right panel).

Table 1
Fitting parameters for the log-normal distribution

Text	All words		Autosemantic PoS		Finite verbs	
	μ	σ	μ	σ	μ	σ
BC1	2.415	0.972 $R^2 = 0.965$	2.231	0.786 $R^2 = 0.983$	1.021	0.532 $R^2 = 0.998$
BC2	2.176	1.016 $R^2 = 0.977$	1.998	0.820 $R^2 = 0.945$	0.964	0.562 $R^2 = 0.992$
BSm	2.090	1.035 $R^2 = 0.981$	1.926	0.798 $R^2 = 0.965$	0.881	0.505 $R^2 = 0.993$
DDO	1.936	0.942 $R^2 = 0.991$	1.826	0.728 $R^2 = 0.976$	0.829	0.469 $R^2 = 0.995$
NSB	1.977	0.935 $R^2 = 0.985$	1.839	0.732 $R^2 = 0.980$	0.850	0.501 $R^2 = 0.9997$
OSu	1.922	0.944 $R^2 = 0.989$	1.791	0.742 $R^2 = 0.977$	0.809	0.496 $R^2 = 0.992$
PD2	2.187	0.974 $R^2 = 0.983$	2.059	0.771 $R^2 = 0.974$	0.957	0.503 $R^2 = 0.999$
PSt	1.915	0.948 $R^2 = 0.989$	1.810	0.744 $R^2 = 0.978$	0.822	0.500 $R^2 = 0.996$
VSh	1.880	0.927 $R^2 = 0.977$	1.745	0.702 $R^2 = 0.951$	0.793	0.478 $R^2 = 0.991$
ZBe	2.205	1.028 $R^2 = 0.965$	2.059	0.788 $R^2 = 0.966$	0.887	0.479 $R^2 = 0.997$

Aiming to catch the high peaks at low values of lengths ($x = 2, 3$), we have also applied the following phenomenologically introduced function:

$$P_x = \frac{\beta b^{\alpha/\beta}}{\Gamma(\alpha/\beta)} x^{\alpha-1} \exp(-bx^\beta), \quad (2)$$

where $\Gamma(z)$ stands for Euler's gamma function. The constant factor is obtained from the normalization condition in the following form:

$$\int_0^{\infty} P_x dx = 1 . \quad (3)$$

It appeared, however, that the values of β yielding best fits are rather low ($\beta \sim 0.1$) and thus correspond in the limit of $\beta \rightarrow 0$ to the logarithmic dependence $\ln x$ in the exponential found in the log-normal distribution (1).

A proper description of the small length values was achieved by using the following function (cf. Grzybek 2012, 2015):

$$P_x = \frac{1}{2(b/c)^{a/2} K_a(\sqrt{bc})} x^{-a-1} \exp\left(-bx - \frac{c}{x}\right) \quad (4)$$

with $K_\nu(z)$ standing for the modified Bessel function of the second kind (Abramowitz & Stegun 1972: Chap. 9). The constant factor is obtained using the normalization condition (3).

In Figure 7, fitting results are compared for three above-mentioned models (1), (2), (4). One can easily notice that both (1) and (2) have almost identical behavior in the domain of small sentence lengths. On the other hand, model (4) yields a better description for $j = 2-3$ and even catches properly the number of shortest sentences ($j = 1$).

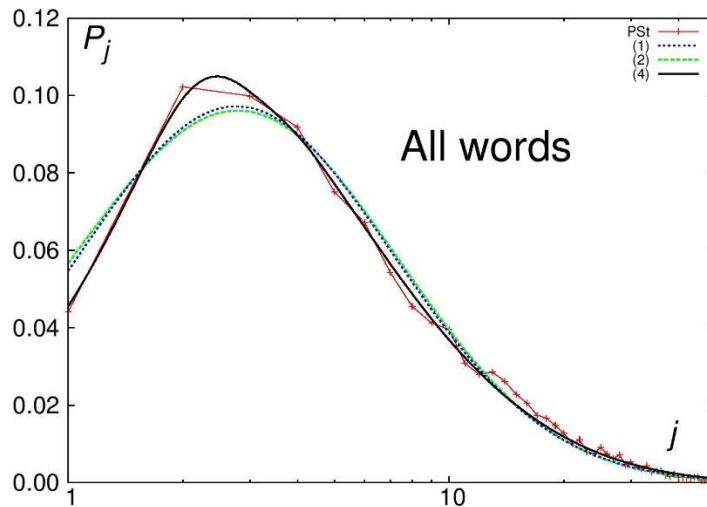


Figure 7 Fitting sentence length measured in all words for PSt.

Note the logarithmic scale on the horizontal axis. The values of fitting parameters are as follows: $a = 0.219$, $b = 0.0456$, $c = 3.34$. The determination coefficient $R^2 = 0.997$.

5. Discussion

We analyzed sentence lengths in Ukrainian texts of one writer, Ivan Franko. The availability of a tagged text corpus of his works made it possible to apply various units to measure the length: in addition to orthographic words, the number of autosemantic parts of speech and the number of finite verbs were calculated.

Three continuous distributions were tested as models for sentence lengths. They are the log-normal distribution and two phenomenologically introduced dependences, $x^{-a-1} \exp(-bx^\beta)$ and $x^{-a-1} e^{-bx-c/x}$ (both of them properly normalized). The second model yields good fits at small values of the β parameter and thus its limiting version ($\beta \rightarrow 0$) corresponds to the log-normal distribution. The third model should be used to obtain good fits in the domain of small sentence lengths, where it works better than the log-normal one.

In general, all the distributions lead to rather good descriptions of available data, with the determination coefficient rarely going below 0.95. The best fits are in the case of sentence lengths measured as the number of finite verbs ($R^2 > 0.99$). The obtained results complement previous studies of the same text material with discrete models (Buk et al. 2019).

In prospect, it would be useful to compare the sentence lengths in Ukrainian with the same characteristics in other Slavonic languages, as well as with data from languages from different language families. While the reported study concerns the texts of one writer, we expect that the tested models could be used not only for the Ukrainian language in general but for other languages due to their simplicity. To what extent the values of parameters are characteristic with respect to author/style/language is yet to be clarified.

In his chapter from *The Oxford Handbook of the Word* about word length (Grzybek 2015), Peter wrote:

Surprisingly, however, the word–clause relation has not to date been empirically studied (Cramer 2005, p. 672)—a research gap soon to be filled (Grzybek & Rovenchak 2014).

Unfortunately, this gap has never been filled in view of complexity of a universal clause definition applicable for various languages. We hope that the results of this work will eventually become a small gap-filler complementing other approaches based on the analysis of syntax, in particular using Universal Dependencies (Nivre et al. 2016). But this is yet another story.

References

- Abramowitz, M. & Stegun, I.A.** (1972). *Handbook of Mathematical Functions*. Washington, D. C.: National Bureau of Standards.
- Aikhenvald, A.Y.** (2018). ‘Me’, ‘us’, and ‘others’: expressing the self in Arawak languages of South America, with a focus on Tariana. In: Huang, M. & Jaszczolt, K.M. (eds.), *Expressing the Self: cultural diversity and cognitive universals: 13-39*. Oxford, UK: Oxford University Press.
- Buk, S.** (2012). The architecture of Polish-Ukrainian and Ukrainian-Polish parallel corpus of Ivan Franko’s self-translations. *Slavia Orientalis* 61, 2: 213-230.
- Buk, S.** (2013). Kvantyatyvna parametryzacija tekstiv Ivana Franka: proekt ta joho realizacija. *Visnyk Lvivskoho Universytetu Serija Filolohichna* 58: 290-307.
- Buk, S., Krynytskyi, Y. & Rovenchak, A.** (2019). Properties of autosemantic word networks in Ukrainian texts. *Advances in Complex Systems* 22, 6, 1950016. DOI: <https://doi.org/10.1142/S0219525919500164>
- Buk, S. & Rovenchak, A.** (2008). Menzerath–Altmann law for syntactic structures in Ukrainian. *Glottology* 1, 1, 10–17. DOI: <https://doi.org/10.1515/glot-2008-0002>
- Buk S. & Rovenchak, A.** (2016). Probing the “temperature” approach on Ukrainian texts: Long-prose fiction by Ivan Franko. In: Kelih, E., Knight, R., Mačutek, J. & Wilson, A. (eds.), *Studies in Quantitative Linguistics 23: Issues in Quantitative Linguistics 4: 160-175*. Lüdenscheid: RAM-Verlag.
- Denysiuk, I.** (2008). Novatorstvo Franka-prozajika. *Ukrajinske literatureznavstvo* 70, 138-152.
- Finch, G.** 2000. Syntax. In: *Linguistic Terms and Concepts: 77–141*. Basingstoke: Palgrave. DOI: https://doi.org/10.1007/978-1-349-27748-3_4
- Fortson, B.W.** (2011). *Indo-European Language and Culture: An Introduction*. New York, NY: Wiley.
- Grzybek, P.** (2012). Close and Distant Relatives of the Sentence: Some Results from Russian. In: Obradović, I., Kelih, E. & Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics: Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 26-29, 2012: 44-58*. Belgrade: Academic Mind.
- Grzybek, P.** (2015). Word length. In: Taylor, J.R. (ed.), *The Oxford Handbook of the Word: 89-119*. Oxford: Oxford University Press. DOI: <https://doi.org/10.1093/oxfordhb/9780199641604.013.37>
- Idiatov, D.** (2010). Person–number agreement on clause linking markers in Mande. *Studies in Language* 34, 4, 832–868. DOI: <https://doi.org/10.1075/sl.34.4.03idi>
- Ishida, M. & Ishida, K.** (2007). On distributions of sentence lengths in Japanese writing. *Glottometrics* 15, 28-44.
- Johnson, N.L., Kotz, S. & Balakrishnan, N.** (1994). 14: Lognormal Distributions. In: *Continuous Univariate Distributions, Vol. 1. Second edition. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics: 207-258*. New York: John Wiley & Sons.

- Kalinina, E. & Sumbatova, N.** (2007.) Clause structure and verbal forms in Nakh-Daghestanian languages. In: Nikolaeva, I. (ed.) *Finiteness. Theoretical and Empirical Foundations: 183-249*. Oxford: Oxford University Press.
- Kelih, E., Grzybek, P., Antić, G. & Stadlober, E.** (2006). Quantitative text typology: The impact of sentence length. In: Myra Spiliopoulou, M. et al. (eds.), *From Data and Information Analysis to Knowledge Engineering. Studies in Classification, Data Analysis, and Knowledge Organization: 382-389*. Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/3-540-31314-1_46
- Khany, R. & Babanezhad Kafshgar, N.** (2016). Analysing texts through their linguistic properties: A cross-disciplinary study. *Journal of Quantitative Linguistics* 23, 3, 278-294. DOI: <https://doi.org/10.1080/09296174.2016.1169848>
- Landolfi, A., Sammarco, C. & Voghera, M.** (2010). Verbless clauses in Italian, Spanish and English: a Treebank annotation. In: Bolasco, S., Chiari, I. & Giuliano, L. (eds.) *Statistical Analysis of Textual Data. Proceedings of the 10th International Conference Journées d'Analyse statistique des données Textuelles, LED: 1187-1194*.
- Limpert, E., Stahel, W. A. & Abbt, M.** (2001). Lognormal distributions across the sciences: Keys and clues. *BioScience* 51, 5, 341-352. DOI: [https://doi.org/10.1641/0006-3568\(2001\)051\[0341:LNDATS\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2)
- Mačutek, J.** (2007). Pairs of corresponding discrete and continuous distributions: Mathematics behind, algorithms and generalizations. In: Grzybek, P. & Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 407-414*. Berlin: de Gruyter. DOI: <https://doi.org/10.1515/9783110894219.407>
- Mačutek, J. & Altmann, G.** (2007). Discrete and continuous modelling in quantitative linguistics. *Journal of Quantitative Linguistics* 14, 1, 81-94. DOI: <https://doi.org/10.1080/09296170600850627>
- Maillart, Ch. & Parisse, Ch.** (2019). Clauses and phrases. In: Damico, J.S. & Ball, M.J. (eds.), *The SAGE Encyclopedia of Human Communication Sciences and Disorders: 334-347*. Thousand Oaks, CA: SAGE. DOI: <https://doi.org/10.4135/9781483380810.n119>
- Martin, J., Johnson, H., Farley, B., & Maclachlan, A.** (2003). Aligning and using an English-Inuktitut parallel corpus. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts. Data Driven Machine Translation and Beyond, Vol. 3, Edmonton, May-June 2003: 115-118*.
- Miller, C.L.** (ed.) (1999). *The Verbless Clause in Biblical Hebrew: Linguistic Approaches*. Winona Lake, Indiana: Eisenbrauns.
- Nivre, J. et al.** (2016). Universal dependencies v1: A multilingual treebank collection. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16): 1659-1666*.
- Özcan, A. & Kuruoğlu, G.** (2018). Sentence length of Turkish patients with schizophrenia. *International Journal of Psycho-Educational Sciences* 7, 1, 68-73.
- Pande, H. & Dhimi, H. S.** (2015). Determination of the distribution of sentence length frequencies for Hindi language texts and utilization of sentence length frequency profiles for authorship attribution. *Journal of Quantitative Linguistics* 22, 4, 338-348. DOI: <https://doi.org/10.1080/09296174.2015.1106269>
- Pastukh, T.** (1996). Roman u systemi prozovykh tvoriv Ivana Franka. *Ukrajins'ke*

literaturoznavstvo 62, 100-108.

- Roukk, M.** (2007). The Menzerath–Altmann law in translated texts as compared to the original texts. In: Grzybek, P. & Köhler, R. (eds.), *Quantitative Linguistics, Vol. 62: Exact Methods in the Study of Language and Text. Dedicated to Professor Gabriel Altmann On the Occasion of His 75th Birthday: 605–610*. Berlin: de Gruyter. DOI: <https://doi.org/10.1515/9783110894219.605>
- Rovenchak, A. & Buk, S.** (2018). Part-of-speech sequences in literary text: Evidence from Ukrainian. *Journal of Quantitative Linguistics* 25, 1, 1-21. DOI: <https://doi.org/10.1080/09296174.2017.1324601>
- Sanada, H.** (2016). The Menzerath-Altmann law and sentence structure. *Journal of Quantitative Linguistics* 23, 3, 256–277. DOI: <https://doi.org/10.1080/09296174.2016.1169850>
- Scarborough, H.S. et al.** (1991). The relation of utterance length to grammatical complexity in normal and language-disordered groups. *Applied Psycholinguistics* 12, 1, 23-46. DOI: <https://doi.org/10.1017/S014271640000936X>
- Sichel, H.S.** (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society: Series A* 137, 25–34.
- Vieira, D.S., Picoli, S. & Mendes, R.S.** (2018). Robustness of sentence length measures in written texts. *Physica A: Statistical Mechanics and its Applications* 506, 749-754. DOI: <https://doi.org/10.1016/j.physa.2018.04.104>
- Williams, C.B.** (1940). A note on the statistical analysis of sentence length as a criterion of literary style. *Biometrika* 31, 3/4, 356–361. DOI: <https://doi.org/10.2307/2332615>; <https://www.jstor.org/stable/2332615>
- Wimmer, G. & Altmann, G.** (1999). *Thesaurus of Univariate Discrete Probability Distributions*. Essen: Stamm.
- Wimmer, G. et al.** (2003). *Úvod do analýzy textov*. Bratislava: Veda.
- Yang, J.** (2019). Syntactic hierarchy depth: distribution, interrelation and cross-linguistic properties. *Journal of Quantitative Linguistics* 26, 2, 129-145. DOI: <https://doi.org/10.1080/09296174.2018.1453962>
- Yaruss, J.S.** (1999). Utterance length, syntactic complexity, and childhood stuttering. *Journal of Speech, Language, and Hearing Research* 42, 2, 329-344. DOI: <https://doi.org/10.1044/jslhr.4202.329>
- Yule, G.U.** (1939). On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. *Biometrika*, 30, 3/4, 363-390. DOI: <https://doi.org/10.2307/2332655>

Evolution of Chinese Word Length Motifs

Heng Chen¹, Haitao Liu^{1,2}, Gabriel Altmann³

Abstract

Word length is an explicit and central property of word, especially in synergetic linguistics. Figuring out how word length sequence evolves in a language is critical for constructing a dynamic lexical control cycle, as well as explaining and predicting language sub-systems and interactions and changes in the language. Based on diachronic written Chinese texts of six periods across 2,000 years, this study employed word length motif to measure word length sequence dynamics in texts. Results show that, in terms of the two fitting models used, i.e. the Zipf–Alekseev function and the Menzerath–Altmann function, the latter yields better fitting results, which reveals evolutionary patterns regarding its parameters. Further analyses indicate that the evolutionary mechanism in written Chinese, particularly word length increase (multi-syllabicity) and emerging coherent rhythmic types, turn out to be the causes of the changes of word length sequences in written Chinese.

Keywords: word length; L-motif; written Chinese; language evolution; word sequence

1. Introduction

Word length, a substantial and central phenomenon in a language’s complex dynamic system (Grzybek & Altmann 2002, Grzybek 2015), serves as a crucial linguistic property for developing a comprehensive theory of language, especially in quantitative linguistics (Grzybek 2006, Altmann 2013). Although word length is subordinate to word, it has extremely strong psychological reality (Baddeley, Thomson & Buchanan 1975), inheriting some intrinsic features of word (Köhler 2008), including cognitive complexity (Baddeley, Thomson & Buchanan 1975), rhythmic features (Chao 1976, Duanmu 2012), and information content (Piantadosi et al. 2011). Moreover, in the lexical control cycle, word length interacts with word frequency, polysemy, and other structural properties, thus forming a scientific linguistic theory, i.e. synergetic linguistics (Köhler 1986, 2005).

The distribution of a given word length in a text is not chaotic, but follows law-like regularities. Taking a text as a bag of words, the distribution of word length adheres to certain model families (Chen & Liu 2016). In the hierarchical dimension, word length abides by the Menzerath–Altmann law with regard to language units length, and in the sequential dimension, “a textual sequence is rather a repetition pattern displaying manifold regularities such as distance, lumping, strengthening or weakening towards the end of sentence or chapter or text, oscillations, cohesion, etc.” (Altmann 2015, p. 1). Psychological studies have shown that a human’s memory span is inverse to word length, that is, longer words are harder to remember and tend to be avoided since they are uneconomic in communication (Baddeley, Thomson, & Buchanan 1975). According to Zipf’s law and the principle of least effort, more frequent words are generally

¹ Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China, chenheng@gdufs.edu.cn.

² Department of Linguistics, Zhejiang University, Hangzhou, China, htliu@163.com.

³ Ruhr-Universität Bochum, Germany. Gabriel Altmann sent us the data on Aug. 5 last year (2019), encouraging us to write a new article, and promising to be a co-author. Now we have completed the paper, but dear Gabriel is no longer with us. May he rest in peace!

much shorter. Conversely, longer words usually contain more information and are more likely to be used to express abstract concepts (Garcia et al. 2012). Interestingly, in a running text, average word length usually increases from beginning to end (Fan et al. 2010), which supposedly can be explained with the information theory and theme-rheme theory: in a sentence new information usually follows known information, and new information tends to be expressed with more complex and longer words (Clark & Clark 1977, Givón 1984).

Word length dynamics underlines lexical diversification and language evolution (Grzybek 2015). Previous studies show that word length co-evolves with the lexical system, resulting in an increase of word length (Lieberman 2011, Bochkarev et al. 2012, Chen & Liu 2014), which is claimed to be an essential regularity of word evolution in written Chinese (Chen, Liang & Liu 2015). Therefore, the modelling of evolutionary patterns of word length in texts is of great importance for language evolution studies (Lian & Li 2019). Earlier studies are more focused on the frequency distributions of word lengths (for a review see Grzybek 2006). Recent years have seen a surge of word length sequence studies, such as word length correlations (Guzmán-Vargas et al. 2015, Kalimeri et al. 2015, Chen & Liu 2018), word length repetitions (Altmann & Köhler 2015), and word length entropies (Ebeling & Pöschel 1994, Papadimitriou et al. 2010, Kalimeri et al. 2015).

In this study, however, we turn to the latest sequence analysis method termed “motif” (Liu & Liang 2017), a syntagmatic approach toward language sequences, which has been suggested by Köhler (2006, 2008) and Köhler & Naumann (2008). A motif is the longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic structural unit (Köhler 2006). Motifs have been widely used in studies such as automatic text classifications (Köhler & Naumann 2010), stylistic analyses, authorship identification (Liu & Liang 2017), literary interpretations (Liang, Lv & Liu 2019), etc. Studies have also shown that motif is a very promising attempt to discover the syntagmatic relations of the word lengths in texts (Milička 2015, Köhler 2006, 2008, 2015). A word length motif (L-motif) is defined as the longest continuous series of words of equal or increasing length (usually measured in syllables). In this paper we will explore how L-motifs have evolved in written Chinese across two millennia.

Our previous study (Chen & Liang 2017) has shown that the rank-frequency distributions of L-motifs in written Chinese can be fitted with the power law function $y=ax^b$, and an increasing trend of parameter a and a decreasing trend of parameter b are observed for diachronic data. However, for length distribution of L-motifs, the hyper-Pascal distribution (Köhler & Naumann 2008) seems unfit for ancient Chinese data. Since “no hypothesis and none of its mathematical expressions (models) are final” (Altmann 2015, p. 5), new data and new models are needed. The Zipf–Alekseev function is a universal model for length distributions of linguistic units (Popescu & Altmann 2014), and the Menzerath–Altmann function has been used to fit length distributions of L-motifs (Mačutek & Mikros 2015). Thus, in this study, the Zipf–Alekseev function and the Menzerath–Altmann function are used to fit the data, and further evolutionary problems are discussed. We will explore three research questions (RQ) in particular:

- RQ 1: Can the diachronic data of length distributions of L-motifs in written Chinese be fitted with the Zipf–Alekseev model and Menzerath–Altmann model?
- RQ 2: How do L-motifs evolve across six periods in written Chinese, and can the evolutionary patterns be captured with value changes of the parameters in the models?
- RQ3: What is the mechanism driving L-motif evolution in written Chinese?

The rest of this paper is organized as follows. Section 2 describes the materials and methods used in this study; Section 3 gives the results of the diachronic investigations, as well as some

discussions. Section 4 is a conclusion. The anticipation is that this study may give us a much more in-depth understanding of word length motif dynamics.

2. Materials and methods

2.1. Materials

To explore the questions above, we need a manageable number of reliable diachronic texts. In the present study, we use a collection of written Chinese texts ranging from around 300 BC to 2010 AD, which is divided into six time periods: 3rd century BC – 2nd century BC, 4th century AD – 5th century AD, 12th century AD – 13th century AD, 16th century AD – 17th century AD, 19th century AD, 21st century AD. The scale of the whole text collection in each time period ranges from about 10,000 to 2 million characters. Since the distribution of L-motifs can be influenced by text size (Mikros & Mačutek 2017), we randomly selected a sample of 10,000 words of text from the text collections in each time period. The details of the texts are shown in **Table 1**. Moreover, to guarantee the impartiality of the results, the segmentation work was done by an expert in old Chinese.

Table 1
Diachronic corpus details of written Chinese

Time period	1	2	3	4	5	6
	Work	Work	Work	Work	Work	Work
		<i>Shìshū ǒxī nyǔ</i>	<i>Niǎn Yùguānyīn</i>	<i>Shìèrlóu</i>	<i>Nàhǎn</i>	<i>Xīndàofó zhī</i>
	<i>MèngZǐ</i> (Mencius)	(A New Account of the Tales of the World)	(Grinding Jade Goddess of Mercy)	(Twelve Floors)	(Yelling)	(The Buddha Knows Your Mind)
Texts	<i>Lǚshìchūn qū</i> (Mister Lv's Spring and Autumn Annals)	<i>Yánshì Jīaxùn Shū</i> (Mister Yan's Family Motto)	<i>Cuòzhǎncuī níng</i> (Wrongfully Accused of Ying Ning)	<i>Wúshēn gxi</i> (A Silence Play)	<i>Pánghuá ng</i> (Hesitating)	<i>Huíménlǐ</i> (A Wedding Present)
Scale (characters)	141,864	94,729	11,220	233,430	91,705	12,980
Time span	3th century B.C.– 2th century B.C.	4 th century A.D.– 5 th century A.D.	12 th century A.D.– 13 th century A.D.	16 th century A.D.– 17 th century A.D.	20 th century A.D.	21 th century A.D.

2.2. Methods

As defined by Köhler (2015), a length-motif is a continuous series of equal or increasing length values (e.g. of morphs, words, or sentences).

An example of Chinese word length motif segmentation is listed as follows (adapted from Chen & Liang 2017). The Chinese sentence

Chinese: 汉语词长动链是如何演化的?

Pinyin: Hànyǔ cícháng dòngliàn shì rúhé yǎnhuà de?

English (correspondence to words): (Chinese) (word length) (motif) (is) (how) (evolve) (particle 'de')?

English: How does Chinese word length motif evolve?

is according to the above-given definition represented by a sequence of three word length motifs: (2-2-2) (1-2-2) (1). It should be noted that the Chinese word length is measured in syllables in this study.

For the data obtained, we used NLREG and fitted the Zipf–Alekseev function with added 1, and Menzerath function with added 1 (the added 1 is to make sure that the predicted values are no less than 1). The goodness of fit evaluations can be seen from the coefficient of determination R^2 . We say the result is accepted for $R^2 > 0.75$, good for $R^2 > 0.80$, and very good for $R^2 > 0.90$. The two models are as follows.

The Zipf–Alekseev function with added 1:

$$y = cx^{a+b\ln x} + 1$$

The Menzerath–Altmann function with added 1:

$$y = ax^b e^{-cx} + 1$$

3. Results and discussions

In the following sections, the results of our diachronic investigations based on two different models, i.e. the Zipf–Alekseev function and the Menzerath–Altmann function, as well as some syntheses are given respectively.

3.1. Results based on the Zipf–Alekseev model

The fitting results from time period 1 to 6 are displayed in Tables 2–7.

Table 2
Length distributions of word length motifs in period 1

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	3	38.1061	20	5	3.2120
2	167	145.7444	21	6	2.7516
3	185	183.9489	22	7	2.3944
4	166	168.9756	23	5	2.1157
5	124	137.1520	24	6	1.8969
6	103	105.4683	25	3	1.7244
7	72	79.2168	26	3	1.5877
8	59	58.9934	27	1	1.4788
9	46	43.9102	28	1	1.3916
10	53	32.8169	30	2	1.2651

Evolution of Chinese Word Length Motifs

11	25	24.6955	31	3	1.2193
12	24	18.7471	33	1	1.1516
13	17	14.3757	34	6	1.1266
14	21	11.1473	38	1	1.0635
15	13	8.7492	39	1	1.0539
16	12	6.9566	40	1	1.0458
17	14	5.6080	42	1	1.0333
18	11	4.5869	45	2	1.0210
19	14	3.8087			
$a=2.838, \quad b=-1.262, \quad c=37.106 \quad R^2=0.9701$					

Table 3
Length distributions of word length motifs in period 2

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	18	37.7332	16	6	4.9432
2	185	194.5639	17	12	3.8826
3	306	258.6839	18	1	3.1231
4	193	233.2088	19	3	2.5752
5	163	180.7093	20	6	2.1769
6	138	131.0708	21	2	1.8852
7	97	92.3669	22	2	1.6701
8	70	64.4193	24	1	1.3911
9	63	44.9094	25	1	1.3013
10	40	31.4794	26	1	1.2334
11	27	22.2710	27	1	1.1818
12	15	15.9472	29	2	1.1118
13	11	11.5843	30	1	1.0883
14	12	8.5550	31	1	1.0701
15	13	6.4364	38	1	1.0155
$a=3.465, \quad b=-1.540, \quad c=36.733, \quad R^2=0.9683$					

Table 4
Length distributions of word length motifs in period 3

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	25	48.8471	11	24	20.8189
2	253	243.1758	12	17	14.6168
3	323	309.2479	13	20	10.4386
4	235	267.5104	14	8	7.6018
5	208	199.6514	15	6	5.6593
6	142	139.9421	16	5	4.3175
7	107	95.5790	19	1	2.2589
8	65	64.7685	20	1	1.9258
9	41	43.9727	21	3	1.6857
10	33	30.0833			
$a=3.440, \quad b=-1.588, \quad c=47.847, \quad R^2=0.9880$					

Table 5
Length distributions of word length motifs in period 4

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	27	22.6291	11	21	14.3507
2	192	215.8910	12	13	9.4015
3	381	329.6897	13	19	6.3376
4	245	295.6052	14	7	4.4255
5	221	214.9563	15	2	3.2210
6	149	142.4649	16	3	2.4548
7	88	90.5995	17	5	1.9626
8	64	56.7321	20	1	1.2952
9	52	35.4927	21	1	1.2027
10	19	22.3935			
$a=4.741, \quad b=-2.061, \quad c=21.629, \quad R^2=0.9693$					

Table 6
Length distributions of word length motifs in period 5

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	20	70.8818	12	13	7.4679
2	381	343.8280	13	13	5.1269
3	367	388.2316	14	4	3.6666
4	292	294.9740	15	8	2.7442
5	172	194.0350	16	2	2.1544
6	129	120.7034	17	1	1.7726
7	99	73.7371	19	1	1.3573
8	62	45.0618	20	2	1.2466
9	35	27.8355	21	1	1.1718
10	33	17.5020	24	1	1.0611
11	17	11.2667	28	1	1.0171
$a=3.553, \quad b=-1.815, \quad c=69.882, \quad R^2=0.9791$					

Table 7
Length distributions of word length motifs in period 6

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	42	88.8559	10	21	15.6470
2	441	399.3015	11	14	9.9347
3	397	426.1774	12	6	6.5264
4	288	309.6668	13	6	4.4660
5	216	196.2568	14	7	3.2036
6	135	118.2429	15	3	2.4194
7	72	70.2444	16	1	1.9259
8	51	41.8889	19	1	1.2754
9	32	25.3318			
$a=3.455, \quad b=-1.838, \quad c=87.856, \quad R^2=0.9817$					

All six fittings are successful, as displayed in the tables. To have a much clearer look at the changes of parameters a , b , and c in the Zipf–Alekseev function, Figure 1 displays the results from time period 1 to 6.

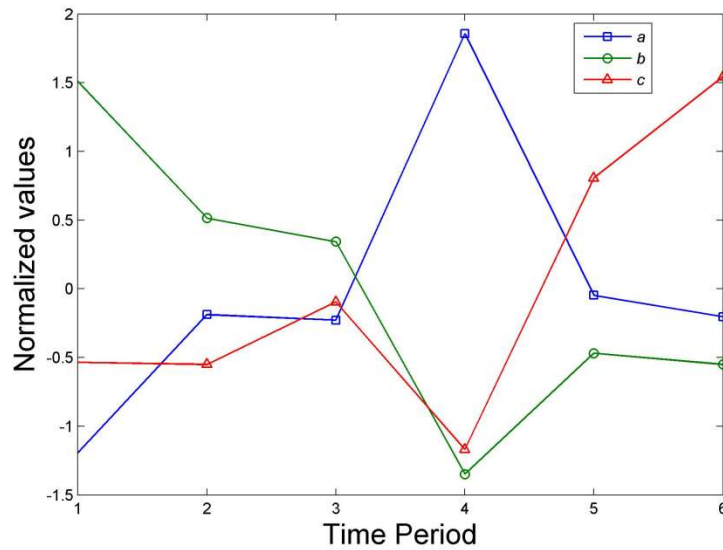


Figure 1 Changes of parameter values based on the Zipf–Alekseev model (with normalized values)

It can be seen from Figure 1 that all three parameters seem to show no evolutionary trend. The linear regressions for each parameter values across the six periods all failed, which corroborates that although the ancient Chinese L-motif length distributions can be fitted with the Zipf–Alekseev function, its parameters show no evolutionary regularities.

3.2. Results based on the Menzerath–Altmann model

The fitting results from time period 1 to 6 based on the Menzerath–Altmann function are displayed in Tables 8–13.

Table 8
Length distributions of word length motifs in period 1

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	3	52.7955	21	6	1.0648
2	167	133.6620	22	7	1.0361
3	185	173.4940	23	5	1.0201
4	166	170.9240	24	6	1.0111
5	124	144.7750	25	3	1.0061
6	103	111.4920	26	3	1.0034

Evolution of Chinese Word Length Motifs

7	72	80.4603	27	1	1.0018
8	59	55.4334	28	1	1.0010
9	46	36.9304	30	2	1.0003
10	53	24.0328	31	3	1.0002
11	25	15.4193	33	1	1.0001
12	24	9.8521	34	6	1.0000
13	17	6.3459	38	1	1.0000
14	21	4.1836	39	1	1.0000
15	13	2.8732	40	1	1.0000
16	12	2.0907	42	1	1.0000
17	14	1.6292	45	2	1.0000
18	11	1.3600	55	1	1.0000
19	14	1.2045	56	1	1.0000
20	5	1.1154	77	1	
$a=103.571,$			$b=2.357,$		
			$c=0.693,$		
			$R^2=0.9302$		

Table 9
Length distributions of word length motifs in period 2

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	18	61.5614	16	6	1.4874
2	185	182.4980	17	12	1.2498
3	306	244.6390	18	1	1.1268
4	193	236.4050	19	3	1.0638
5	163	190.9480	20	6	1.0319
6	138	137.7590	21	2	1.0158
7	97	92.0175	22	2	1.0078
8	70	58.1863	24	1	1.0019
9	63	35.3847	25	1	1.0009
10	40	20.9700	26	1	1.0004
11	27	12.2773	27	1	1.0002

Evolution of Chinese Word Length Motifs

12	15	7.2225	29	2	1.0001
13	11	4.3673	30	1	1.0000
14	12	2.7923	31	1	1.0000
15	13	1.9404	38	1	1.0000
$a=139.936,$		$b=2.792,$	$c=0.838, R^2=0.9386$		

Table 10
Length distributions of word length motifs in period 3

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	25	74.2751	11	24	9.6920
2	253	225.8942	12	17	5.4793
3	323	296.9419	13	20	3.2615
4	235	276.3291	14	8	2.1220
5	208	212.8298	15	6	1.5483
6	142	145.5214	16	5	1.2645
7	107	91.7531	19	1	1.0278
8	65	54.6327	20	1	1.0129
9	41	31.2597	21	3	1.0060
10	33	17.4595			
$a=183.698,$		$b=2.944,$	$c=0.919, R^2=0.9627$		

Table 11
Length distributions of word length motifs in period 4

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	27	49.4536	11	21	6.0397
2	192	212.5791	12	13	3.2660
3	381	315.0771	13	19	1.9926
4	245	300.7452	14	7	1.4252
5	221	225.5672	15	2	1.1787
6	149	145.3606	16	3	1.0739
7	88	84.4960	17	5	1.0301

Evolution of Chinese Word Length Motifs

8	64	45.6990	20	1	1.0019
9	52	23.5568	21	1	1.0007
10	19	11.8649			
		$a=149.346,$	$b=3.751,$	$c=1.126,$	$R^2=0.9510$

Table 12
Length distributions of word length motifs in period 5

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	20	96.0927	12	13	1.8126
2	381	317.3662	13	13	1.3164
3	367	385.6326	14	4	1.1206
4	292	310.6769	15	8	1.0452
5	172	199.9268	16	2	1.0167
6	129	111.7748	17	1	1.0061
7	99	56.8956	19	1	1.0008
8	62	27.2410	20	2	1.0003
9	35	12.6587	21	1	1.0001
10	33	5.9595	24	1	1.0000
11	17	3.0368	28	1	1.0000
		$a=323.199,$	$b=3.499,$	$c=1.223,$	$R^2=0.9463$

Table 13
Length distributions of word length motifs in period 6

Length	Frequency	Frequency'	Length	Frequency	Frequency'
1	42	117.6284	10	21	5.1618
2	441	367.3505	11	14	2.6479
3	397	425.7847	12	6	1.6340
4	288	328.5182	13	6	1.2381
5	216	202.9310	14	7	1.0876
6	135	109.0821	15	3	1.0317
7	72	53.4712	16	1	1.0113

Evolution of Chinese Word Length Motifs

8	51	24.7174		19	1	1.0005
9	32	11.1512				
$a=409.787,$		$b=3.464,$		$c=1.257,$		$R^2=0.9510$

It can be seen from the fitting results in Tables 8–13 that all six periods’ data can be successfully fitted with the Menzerath–Altmann function. Moreover, the changes of parameters a , b , and c in the function with time periods are displayed in Figure 2.

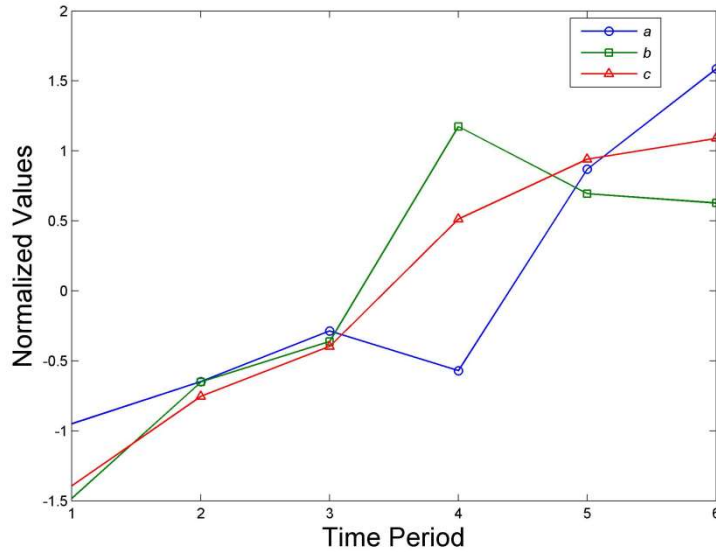


Figure 2 Changes of parameter values based on the Menzerath–Altmann function (with normalized values)

Differently from in Figure 1, in Figure 2 we can see a much clearer change of parameter values. Linear regressions show that, for parameter a , the slope is 0.4842, $R^2= 0.8204$, $P=0.0129$; for parameter b , the slope is 0.4603, $R^2= 0.7416$, $P=0.0276$; for parameter c , the slope is 0.5258, $R^2=0.9677$, $P=0.0004$. All three parameters show a significant trend of increase over the time periods, although there are some oscillations in time period 4, which is a common phenomenon in quantitative linguistics (Grzybek & Altmann 2002).

3.3. Some comprehensive analyses from multiple dimensions

Based on the motifs data, as well as the data of word length per se, we calculated number of L-motifs, entropy, average L-motif length, and average word length in each time period, which are shown in Table 14.

Table 14
 Statistics about the distribution of L-motifs in each period

Time pe- riod	Number of L-mo- tifs	En- tropy	Average L-motif length	Average word length
1	1187	2.7073	7.2072	1.1687
2	1392	2.4089	5.5675	1.2909
3	1517	2.2449	4.8859	1.3603
4	1510	2.1873	4.8046	1.3783
5	1654	2.1197	4.4667	1.3531
6	1733	2.0158	4.0675	1.4163

If we fit a linear function to the number of L-motifs data, the result is successful with a slope of 100.257, $P=0.002$.

The Pearson correlation analyses show that average L-motif length and average word length are highly inversely correlated, which is -0.980 , at the significance level of 0.01 (two-sided). The interrelations between these statistics can be explained as follows: as written Chinese evolves, the average word length increases, and the L-motif types diversify (but L-motif tokens are more clustered); as a result, the number of L-motifs increases, but the average L-motif length decreases over time.

4. Discussions and conclusions

To have a more dynamic understanding of language as evolving system, this study employed word length motif to approach language sequence dynamics in written Chinese across 2,000 years.

Firstly, since the previous study shows that the hyper-Pascal model does not fit ancient Chinese text data (Chen & Liang 2017) well, here we tested two new models, i.e., the Menzerath–Altmann function and the Zipf–Alekseev function. Results show that both of the functions are fit for diachronic data in written Chinese across the six periods. However, as for parameter changes, the Zipf–Alekseev function results display no significant trend. Evolutionary aspects can be seen from the parameter changes in the three parameters of the Menzerath–Altmann model, and all these three parameters increase over time. We suppose the differences between the two fitting results may be that the parameters a and b in the Zipf–Alekseev function are more self-organized (Popescu, Best & Altmann 2014), and the parameters in the Menzerath–Altmann function can capture the hierarchical dynamics of language units in written Chinese evolutions.

Secondly, some evolutionary patterns about the length distribution of L-motifs have been observed. (1) the L-motif length classes decrease from 71 in period 1 to 19 in period 6; (2) the number of motifs increases from 1,187 to 1,733; (3) entropy decreases over time, which indicates that as written Chinese evolves, the L-motifs are becoming more and more aggregated on some specific types; (4) the average L-motif length displays a decreasing trend as average word length goes in the opposite direction.

Thirdly, as for the mechanism driving the evolution of L-motifs in written Chinese, it may be attributed to two main factors, namely the increase of word length, and the forming of specific rhythmic units in written Chinese.

The distribution of L-motifs in a language or text is influenced by many factors. Word length motifs display properties similar to words, and “word meanings are directly related to their recurrence distributions via the permutability of concepts across discourse contexts” (Altmann, Pierrehumbert & Motter 2009, p. 6). Since the length of a word is related to the length of its neighboring words (Mikros & Mačutek 2017), all the factors that influence word collocations in texts will influence L-motif distributions. For example, Mikros & Mačutek (2017) show that word length motifs can be influenced by two important factors, i.e. word length distribution and text length. Moreover, they also claim that the number of motifs is determined by the mean word length (namely, by its inverse proportion) in relatively short texts. Kelih’s (2012) study reports an increase of word length with increasing text length. This is why we choose the same size of text for each time period in this study.

The increase of average word length seems to be a basic regularity in a dynamic lexical control circuit (Altmann 2013, Chen, Liang & Liu 2015), which underlines lexical diversification, and thus enables the expressing and communicating with new concepts more efficiently. So how does word length increase influence word length sequences? According to the definition of the L-motif in section 1, we can reason that the length of motifs in texts can be intensely influenced by the occurrence of long words. In the light of Givón’s quantity principle, less predictable and more important information (usually new information) will be given more coding material, namely longer linguistic units, and vice versa (Givón 1984). Since the sentence-initial part is the position for given information, it certainly needs less coding material and hence shorter linguistic units; sentence-final constituents often convey new information, hence being longer and more complex (Wang & Liu 2014). Therefore, the increase of word length (multi-syllabicity) will certainly have a significant influence on L-motifs, especially the occurrence of long L-motifs, which will interrupt continuous short words with increasing probability over time periods.

L-motifs can also be influenced by rhythmical aspects (Wilson 2019). In Chinese, studies show that there are preferred word length combinations in Chinese (Duanmu 2012). For example, it is found that 1+2 is overwhelmingly disfavored in [Noun+Noun] combinations and 2+1 is overwhelmingly disfavored in [Verb+Object] combinations in modern written Chinese. The results in this study show that the preferences change over time periods.

Language is a complex dynamic system, which displays self-organizing and self-adapting features. The increase of word length results in changes in word length sequences, as well as other changes in associated sub-systems in the language, which needs further research in the future.

Acknowledgements

This study was supported by the Philosophy and Social Science Planning Project of Guangdong Province, China (Grant No.GD19YYY04), the Social Science Foundation of Education Ministry of China (Grant No.20YJC740005), and the project from Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies (Grant No. JDXM1902).

References

- Altmann G.** (2013). Aspects of word length. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics 3*: 23-38. Lüdenscheid: RAM.
- Altmann G.** (2015). Introduction. In: Mikros, G.K., Mačutek, J. (eds.), *Sequences in Language and Text: 1-6*. Berlin: de Gruyter.
- Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E.** (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One* 4, 11, E7678.
- Altmann, G. & Köhler, R.** (2015). *Forms and Degrees of Repetition in Texts*. Berlin: de Gruyter.
- Baddeley, A. D., Thomson, N., & Buchanan, M.** (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning & Verbal Behavior* 14, 575-589.
- Bochkarev, V.V., Shevlykova, A.V., and Solovyev, V.D.** (2012). Average word length dynamics as indicator of cultural changes in society. [<http://arxiv.org/abs/1208.6109>]
- Chao, Y.** (1976). Rhythm and structure in Chinese word Conceptions. In: Chao, Y. (ed.), *Aspects of Chinese Sociolinguistics: Essays by Yuanren Chao*: 275-292. Stanford: Stanford University Press. (selected and introduced by Anwar S. Dil.)
- Chen, H. & Liu, H.** (2014). A diachronic study of Chinese word length distribution. *Glottometrics* 29, 81-94.
- Chen, H., & Liang, J.** (2017). Chinese word length motif and its evolution. In: Liu, H., Liang, J. (eds.), *Motifs in Language and Text: 37-64*. Berlin: de Gruyter.
- Chen, H., & Liu, H.** (2016). How to Measure Word Length in Spoken and Written Chinese. *Journal of Quantitative Linguistics* 23, 1, 5-29.
- Chen, H., & Liu, H.** (2018). Quantifying evolution of short and long-range correlations in Chinese narrative texts across 2000 years. *Complexity* 2018, 1-12.
- Chen, H., Liang, J., Liu, H.** (2015). How does word length evolve in written Chinese? *PLoS One* 10, 9, e0138567.
- Clark, H.H., Clark, E.V.** (1977). *Psychology and language: An introduction to psycholinguistics*. New York: Harcourt Brace Jovanovich.
- Duanmu, S.** (2012). Word-length preferences in Chinese: a corpus study. *Journal of East Asian Linguistics* 21, 1, 89-114.
- Ebeling, W., Pöschel, T.** (1994). Entropy and long-range correlations in literary English. *EPL (Europhysics Letters)* 26, 4, 241-246.
- Fan, F., Grzybek, P., Altmann, G.** (2010). Dynamics of word length in sentence. *Glottometrics* 20, 70-109.
- Garcia, D., Garas, A., & Schweitzer, F.** (2012). Positive words carry less information than negative words. *EPJ Data Science* 1, 1-12.
- Givón, T.** (1984). *Syntax: A functional-typological introduction*. Amsterdam: Benjamins.
- Grzybek, P.** (2006). History and methodology of word length studies: the state of the art. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues: 15-90*. Dordrecht: Springer.
- Grzybek, P.** (2015). Word length. In: Taylor, J. R. (ed.), *The Oxford Handbook of the Word*: 89-119. Oxford University Press.
- Grzybek, P., Altmann, G.** (2002). Oscillation in the frequency-length relationship. *Glottometrics* 5, 97-107.
- Guzmán-Vargas, L., Obregón-Quintana, B., Aguilar-Velázquez, D. et al.** (2015). Word-length correlations and memory in large texts: A visibility network analysis. *Entropy* 17, 12, 7798-7810.
- Kalimeri, M. et al.** (2015). Word-length entropies and correlations of natural language written texts. *Journal of Quantitative Linguistics* 22, 2, 101-118.

- Kelih, E.** (2012). Systematic interrelations between grapheme frequencies and the word length: empirical evidence from Slovene. *Journal of Quantitative Linguistics* 19, 3, 205-31.
- Köhler, R.** (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* [= Synergetic Linguistics. Lexical Structure and Dynamics]. Bochum: Brockmeyer.
- Köhler, R.** (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative linguistics. An international handbook: 760-774.* Berlin: de Gruyter.
- Köhler, R.** (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bučková, M. (eds.), *Favete linguis. Studies in honour of Viktor Krupa: 145-152.* Bratislava: Slovak Academic Press.
- Köhler, R.** (2008). Word length in text. A study in the syntagmatic dimension. In: Mislovicová, S. (ed.), *Jazyk a jazykoveda v pohybe: 416-421.* Bratislava: Veda.
- Köhler, R.** (2015). Linguistic Motifs. In: Mikros, G. K., Mačutek, J. (eds.), *Sequences in Language and Text: 89-108.* Berlin: de Gruyter.
- Köhler, R., Naumann, S.** (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, B., Schidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg: 637-646.* Berlin, Heidelberg: Springer.
- Köhler, R., Naumann, S.** (2010). A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language. Structures – Functions – Interrelations – Quantitative Perspectives: 81-89.* Wien: Praesens.
- Lian, F., Li, Y.** (2019). Word Length Distribution in German Texts during the 17th-19th Century. *Journal of Quantitative Linguistics* 2, 1-21.
- Liang, J., Fang, Y., Lv, Q., Liu, H.** (2017). Dependency distance differences across interpreting types: implications for cognitive demand. *Frontiers in Psychology* 8, 2132.
- Lieberman, M.** (2011). Real trends in word and sentence length. [<http://languagelog.ldc.upenn.edu/nll/?p=3534>].
- Liu, H., Liang, J.** (eds.) (2017). *Motifs in language and text.* Berlin: de Gruyter.
- Mačutek, J., Mikros, G.K.** (2015) Menzerath-Altmann Law for Word Length Motifs. In: Mikros, G.K., Mačutek, J. (eds.), *Sequences in Language and Text: 125-132.* Berlin: de Gruyter.
- Mikros, G.K., Mačutek, J.** (2017). Word length distribution and text length: two important factors influencing properties of word length motifs. In: Liu, H., Liang, J. (eds.), *Motifs in language and text: 151-164.* Berlin: de Gruyter.
- Milička, J.** (2015). Is the Distribution of L-Motifs Inherited from the Word Length Distribution? In: Mikros, G.K., Mačutek, J. (eds.), *Sequences in Language and Text: 135-145.* Berlin: de Gruyter.
- Papadimitriou, C. et al.** (2010). Entropy analysis of natural language written texts. *Physica A: Statistical Mechanics and its Applications* 389, 16, 3260-3266.
- Piantadosi, S., Tily, H., Gibson E.** (2011). Word lengths are optimized for efficient communication. *PNAS* 108, 9, 3526-3529.
- Popescu, I.-I., Best, K.-H., Altmann, G.** (2014). *Unified modeling of length in language.* Lüdenscheid: RAM.
- Wang, H., Liu, H.** (2014). The effects of length and complexity on constituent ordering in written English. *Poznan Studies in Contemporary Linguistics* 50, 477- 494.
- Wilson, A.** (2019). Lengths and L-motifs of Rhythmical Units in Formal British Speech. *Glottometrics* 48, 37-51.

Language Identification by Simple Character Profiles

Eric S. Wheeler¹, with²
Sheila Embleton, Dorin Uritescu³

Dedicated to Peter Grzybek, a scholar, colleague and friend. He will be missed. We think he might have enjoyed this little project of ours.

Abstract:

A simple measure of a text, using letter frequency, can be used to distinguish the language of a text. The inter-language differences in a sample of texts are very much greater than the intra-language differences. Multidimensional scaling can be used to give a visually satisfying demonstration of these differences.

Keywords: language identification; letter frequency; character profile; visualization; multi-dimensional scaling;

1. Introduction

It seems interesting (and even useful) to be able to identify the language of a text by making only a simple calculation.

The calculation we examine here is the frequency of alphabetic characters in each text from a set of texts, in English, French and some other languages. The character frequencies are set in a fixed order, and the resulting vector of n numbers for each text become coordinates in n -dimensional space. We find that texts in the same language are separated by small distances, whereas the distance between languages (represented by the centroid or average position of a set of texts) is many times greater.

To visually show these differences, we re-purpose some of the tools developed for our other research (in dialectometry, cf. Embleton et al. 2007, 2008, 2018 and in particular, the 3-d viewer in Embleton et al. 2013), and in doing so, discover both an inviting way to present this kind of data, and also many more questions that could be investigated.

¹ York University, Toronto, Canada, eric.wheeler@sympatico.ca.

² We acknowledge the support of the Social Sciences and Humanities Research Council of Canada for their funding of our research projects. While Wheeler must take primary responsibility for what is reported here, all three authors developed the underlying research in dialectometry, and all three are pleased to honour Peter Grzybek.

³ Between the time when this paper was accepted and its final publication, our friend and colleague Dorin Uritescu passed away on 15 April 2020. May he rest in peace.

2. First count

Initially, we selected some texts that were in English with a corresponding translation into French and German. Project Gutenberg (<https://www.gutenberg.org> "... a library of over 60,000 free eBooks") provided some of Charles Dickens' novels in English, and in French and German translations. Figure 1, Figure 2 and Figure 3 show the distribution of characters in the English, French and German texts of Dicken's "Oliver Twist".

Although the patterns are not exactly the same, they are similar and match an exponential trend line (calculated by the spreadsheet) with r-squareds above 0.95 (not surprising, given what we know about Zipf's law, Zipf 1949). It seems unlikely then that this naive approach would give a useful way to distinguish one language from another.

Comparing only the first few, most frequent letters did not seem like a promising measure either, because they were so much the same (see Table 1). The count of characters on these texts showed that the first 6 or 7 types accounted for about half the tokens in each language; in each case "e" was the most frequent and the next few were similar if not identical in sequence. On the other hand, the letters that were distinctive to a language (such as "â" for French, or "ö" in German) did not occur frequently enough and might not be in any particular text. It seemed they might not provide a reliable indicator either.

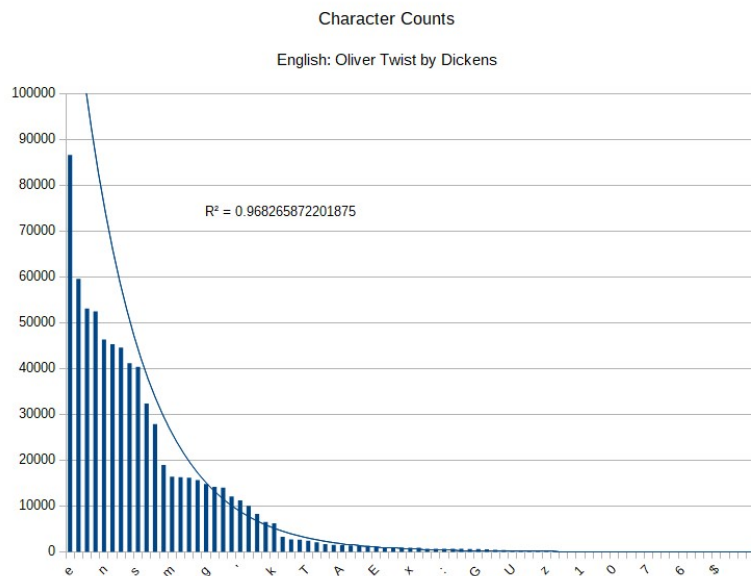


Figure 1 Character counts for the English text of "Dickens: Oliver Twist" in ranked order, with an exponential trend line

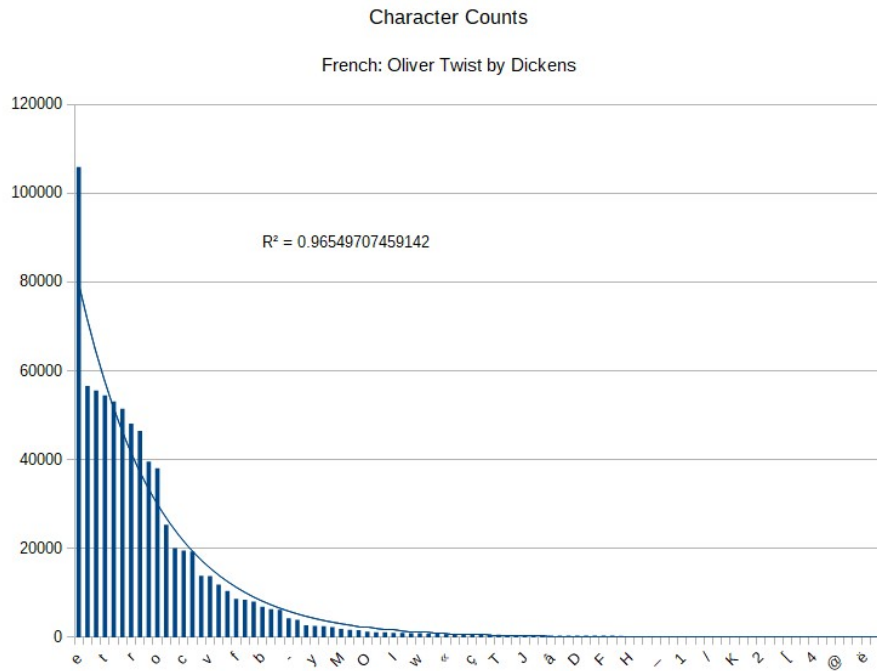


Figure 2 Character counts for the French translation of "Dickens: Oliver Twist" in ranked order, with an exponential trend line

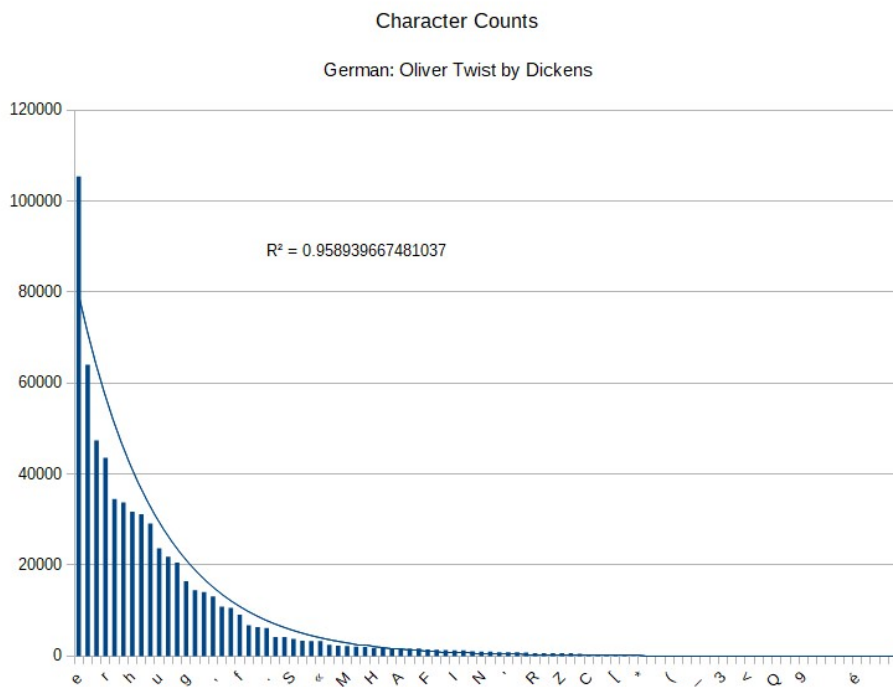


Figure 3 Character counts for the German translation of "Dickens: Oliver Twist" in ranked order, with an exponential trend line

Table 1

The most frequent characters in the English, French and German versions of Dicken's Oliver Twist

English	e t a o n i h r s d l u m
French	e a i t s n r u l o d m c
German	e n i r t s h a d u l c g

3. Character Profiles

Instead, we elected to make what we call a Character Profile, by counting from a long list of letters (alphabet = "0123456789_-,;:?.')[]@*/&#% aAbBcCdDeEfFgGhHiIjJkKlLmMnNoOöpPqQrRsStTuUvVwWxXyYzZ") and ignoring any occurrences that were not in this alphabet (cf. Wheeler 2003 for a similar approach).

To make a Character Profile of a given text, using a given alphabet, we:

1. Saved the text in a text file using an encoding of UTF-16-BE (the sixteen bit, Unicode, "big end" format) so that all the texts were interpreted the same.
2. Counted the occurrences of each character in the text, and the total number of characters.
3. Assigned the counts to a vector in the order of the given alphabet.
4. Normalized the counts by dividing each count by the total number of characters, and then multiplying by a scaling factor (10 000) so that the result is a n-tuple of integers (ranging from 0 to 10 000), where n is the length of the given alphabet. (See Figure 4 for an example.)

Character counts for Dicken's "Oliver Twist" in English

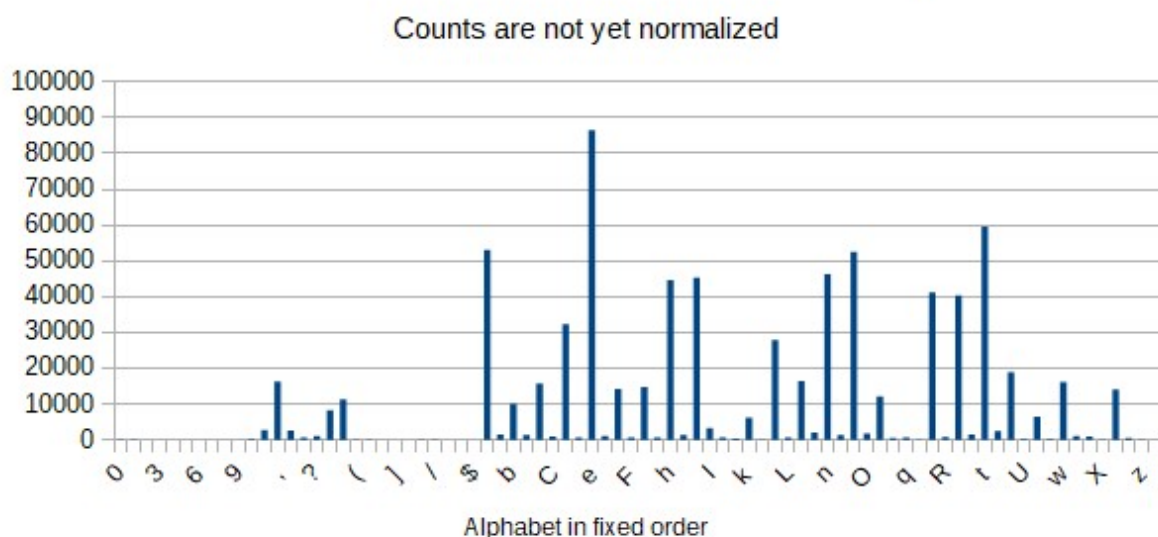


Figure 4 A sample Character Profile as a bar graph

Characters that were not in the alphabet were assigned to a total for "odd" characters, but in subsequent work, they were ignored. However, their impact was still present in the total number of characters, so that a text with a large number of odd characters moderated the normalized

size of the other characters. An alternative approach, then, would be to use a total of only the characters in the alphabet.

4. Visualizing

A Character Profile can be seen as a point in n -space, and texts can be compared by measuring the distance between points. We use the standard Euclidean distance (the square-root of the sum of the squares of the difference between each component of the vectors being compared) which allows us to assume there is an orthogonal basis (axes at right angles to one another).

Using multidimensional scaling (MDS), and the distance matrix for p points, it is possible to create a p -dimensional space in which the p points make a "cloud" in that space, and the cloud can be centred on the origin of the space (Wheeler 2005). Note that p (the number of texts) can be much smaller than n (the length of the alphabet), so it becomes more practical to use MDS to visualize the relationship among the texts rather than trying to visualize the points in n -dimensional space directly.

In either case, though, the points in the high-dimensional space are exactly the right distance from one another; the challenge is to project these high-dimensional points to a 2-dimensional picture or an interactive 3-dimensional picture and still preserve as much of the relationship among them as possible. For two points, a line is sufficient to picture the points at exactly their right distance apart; for 3 points a plane may be needed; and in general p points could require a $(p-1)$ - space in the extreme case.

To make our picture, we look for the point that is the most distant from the origin, and take the line from the origin to it as our x - axis. Then we pick a point that is most distant from the x -axis, and drop a perpendicular to the axis. Our y -axis is chosen to be from the origin, parallel to that perpendicular; the z -axis is orthogonal to the x - y plane, etc. Unless the cloud has some very particular structure, the x -axis is going to account for the most of the distances among the points, the y -axis for the second most, and so on. Mindful of how Zipf's Law works everywhere (Wheeler 2002), it is not surprising then that the first two or three axes give us a good picture of a p -dimensional cloud. In fact, in the experiments we did here, the cloud turned out to be a very flat (2-dimensional) pancake; using the 3-dimensional viewer, we could look at the cloud edgewise, and see that there was little or nothing to the 3rd and higher dimensions. That means that points that are far apart in our pictures are indeed far apart, and that points that are close together are probably close; they might be accidentally close (because of the way we project it) but if they stay close when we rotate them interactively in our 3-d viewer, it is unlikely to be other than as they appear.

5. First Results

In Figure 5 is the picture of the English, French and German versions of a group of Dicken's novels (5 texts each in English and French, 3 in German; they are not necessarily translations of one another). Each text is marked by a solid circle, and a group of texts are connected by an open circle around them, centred on the centroid (the average Character Profile) of the group.

What is striking about this picture is that each language group (marked by the open circle) is very compact, while the distances between language groups are very much larger. We *have* a measure that distinguishes language from language (with issues such as genre and author under control). But this is only one example. Will the measure serve for more languages? and more genres?

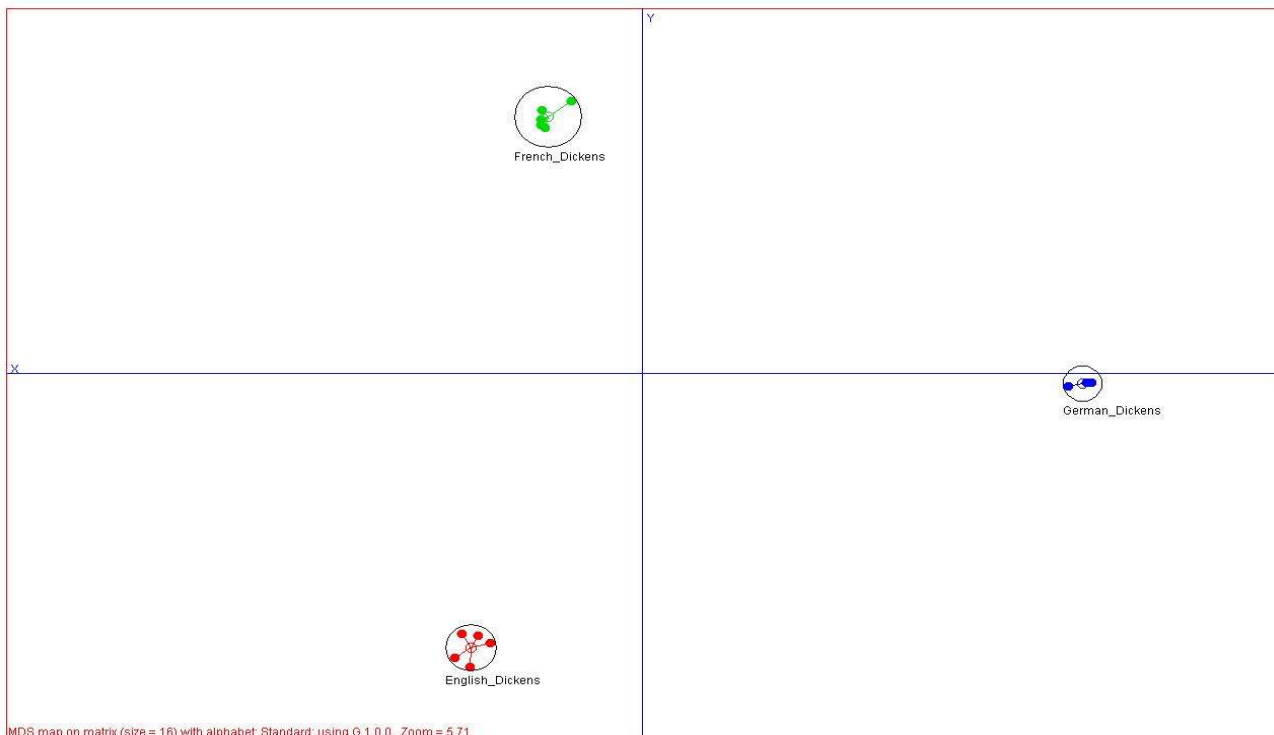


Figure 5 Dicken's novels (5 English, 5 French, 3 German versions) in Character Profile space.

6. Further Results

The Canadian parliamentary proceedings are recorded in Hansard, in both English and French. Here is a source of language texts, originally spoken, but perhaps spoken from written notes, perhaps in the speaker's first or second language, and then written, translated if need be, and edited by the Hansard staff and the speaker; thus, not spontaneous language, but certainly language reflective of what the language is expected to be. We selected the proceedings of 5 consecutive days from October 2004, and compared the English and French.

As expected, the intra-language distances are relatively small compared to the inter-language distances (see Figure 6). In particular, the French texts are all in the same region (one Hansard outlier in both English and French makes the Hansard circles somewhat larger than if the point had been omitted).

The English Dickens group is separate (by a small amount) from the Hansard group. Possibly this reflects the date of the language: Dickens in English was written in the mid-1800's whereas the French translations of Dickens (presumably), and the English and French Hansard are all early twenty-first century. Perhaps over the course of a century and a half, the profile has shifted.

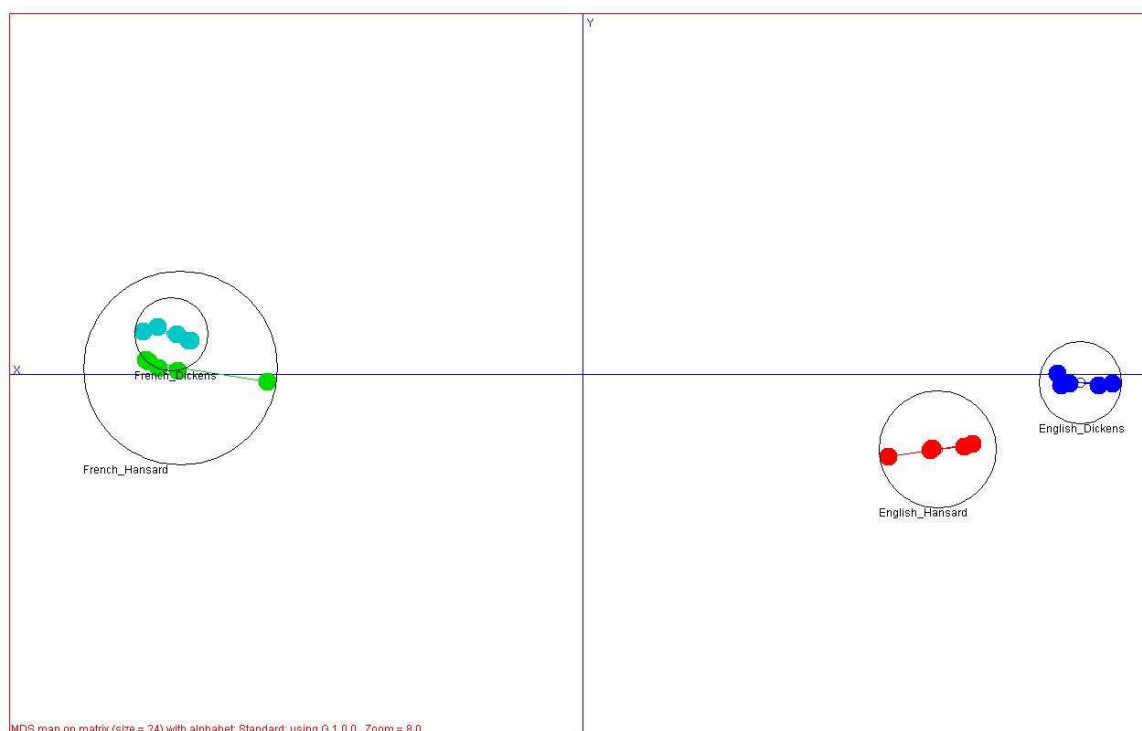


Figure 6 English and French texts from Hansard and Dickens. Note the overlap of the French texts (left), and the small separation of the English groups (right).

The success of these first trials opens up many possibilities. We have downloaded some European Union laws, duly translated into English, French and German (to match Hansard in genre; more languages are available), and added some Dickens in Finnish (a non-Indo-European language), and an Old English text (to expand the time range). The results get a little more crowded with the additional points, but the basic result remains: interlanguage distances exceed intralanguage distances. See Figure 7.

The Finnish group is well removed from the others. Is this an indicator of being in a different language family? Further research, with more texts and more languages, could explore the distances between language families.

The EU groups (admittedly, with a small sample of text) for English and German are removed from the other samples of those languages (but not the French EU sample). Does this reflect some aspect of EU translation? Are French writers more standardized in their language in some sense?

The Old English point is based on Aelfric's grammar, and contains a large amount of Latin in the form of word lists, etc. Does the Latin make a difference to the profile? According to the Dictionary of Old English, the entire corpus of Old English is about 3 million words of Old English and nearly a million words of Latin (see DOE and VARIENG). Perhaps this mixture is still an appropriate way to make the Old English profile.

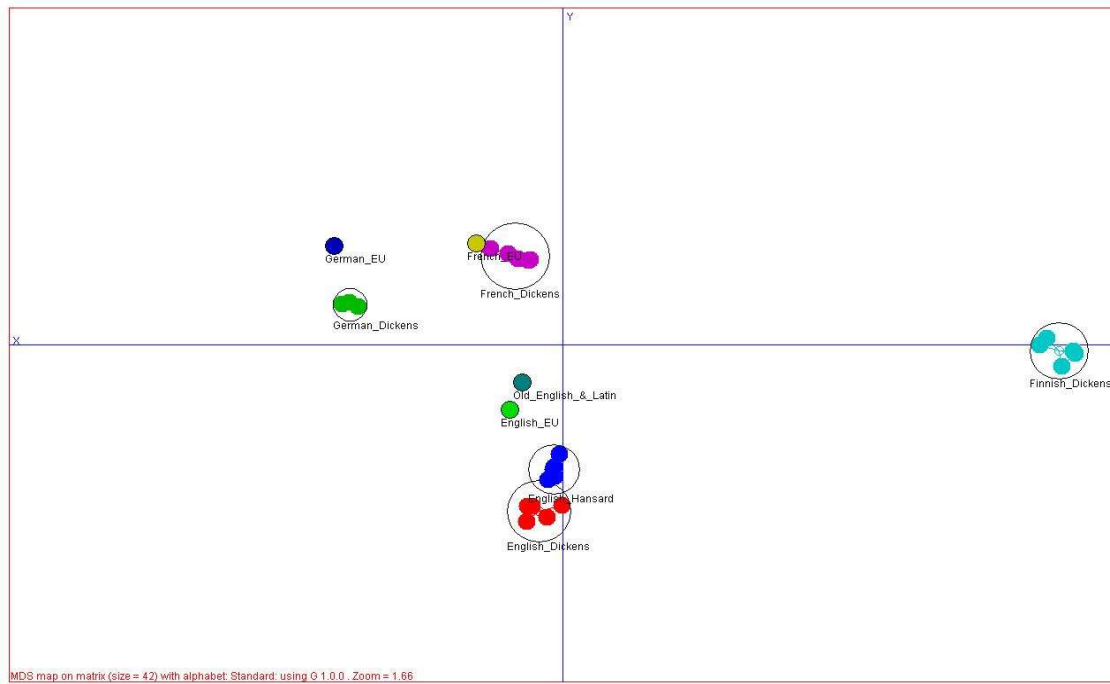


Figure 7 Various texts from English, French, German, Finnish and Old English & Latin. The French Hansard is omitted because it would be with and hide the other two French groups

7. Future directions

Clearly, Character Profiles and their placement in Character Profile space can be both an interesting and useful approach to characterizing language varieties. The approach invites looking at more texts, of various kinds. That involves work, gathering and formatting the texts, and grouping them appropriately, but it also invites us to find explanations for the differences. Is it simply a matter of accident that languages use the alphabet differently? Or is there some correlate (such as language family, genre or age) that goes with it? To achieve valid results, more data and perhaps more sophisticated statistics are needed. But in the meanwhile, the simple measure of texts as Character Profiles, and the visual presentation of Character Profiles gives us a convincing, and intuitively satisfying view of how different languages can be.

References

- DOE.** Dictionary of Old English. <https://www.doe.utoronto.ca/pages/index.html>
- Embleton, Sh.M., Uritescu, D., Wheeler, E.S.** (2007). Romanian Online Dialect Atlas: Data Capture and Presentation. In: Grzybek, P.; Köhler, R. (eds), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*: 87-96. Berlin and New York: de Gruyter.
- Embleton, Sh.M., Uritescu, D., Wheeler, E.S.** (2010). Identifying Dialect Regions: Specific features vs. overall measures using the Romanian Online Dialect Atlas and Multidimensional Scaling. In: Heselwood, B., Upton, C. (eds.), *Proceedings of Methods XIII. Papers from the Thirteenth International Conference on Methods in Dialectology, 2008*: 79-90. Frankfurt/Main: Peter Lang.
- Embleton, Sh., Uritescu, D., Wheeler, E.S.** (2013). Defining dialect regions with interpretations. Advancing the multidimensional scaling approach. *Literary and Linguistic Computing* 28, 1, 13-22. <https://doi.org/10.1093/lilc/fqs048>
- Embleton, Sh., Uritescu, D., Wheeler, E.S.** (2018). “An Expanded Quantitative Study of Linguistic vs Geographic Distance Using Romanian Dialect Data”. In: Wang, L., Köhler, R., Tuzzi, A. (eds.), *Structure, Function and Process in Texts, Proceedings of Qualico 2016*: 25-33. Lüdenscheid: RAM.
- VARIENG.** The Research Unit for the Study of Variation, Contacts and Change in English. <http://www.helsinki.fi/varieng/CoRD/corpora/DOEC/> (accessed 9 Dec 2019)
- Wheeler, E.S.** (2002). Zipf's Law and Why It Works Everywhere. *Glottometrics* 4, 45-48.
- Wheeler, E.S.** 2003. Multidimensional Scaling for Visualizing Text Separation Methods. *Glottometrics* 6, 65-69.
- Wheeler, E.S.** (2005). Multidimensional Scaling for Linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds), *Quantitative Linguistics. An International Handbook*: 548-553. Berlin: de Gruyter.
- Zipf, G.K.** (1949). *Human Behavior and the Principle of Least Effort, An Introduction to Human Ecology*. facsimile 1965. New York: Hafner.

A Classification of the Celtic Languages Based on Grapheme Frequencies

Andrew Wilson¹, Ján Mačutek²

Abstract

Grapheme frequencies from a small parallel corpus of psalms translated into Breton, Cornish, Irish (both the Early Modern and present-day versions of the language), Manx, Scottish Gaelic, and Welsh are analyzed. They can be modelled – as with many other languages – by the negative hypergeometric distribution. Based on the modified Ord graph constructed from the grapheme frequencies, the Celtic languages can be divided into two groups which differ slightly from the traditional Celtic language classification (Manx is placed among the P-Celtic rather than the Q-Celtic languages); however, the difference can be explained by the English (or Scots) influences on the Manx orthography.

Keywords: Celtic languages; grapheme frequencies; language classification

1. Introduction

The aim of this paper is to undertake a mathematical modelling of grapheme frequencies in the six surviving Insular Celtic languages (including two periods in the history of Irish), with a view to establishing how far they support the traditional genetic classification into Q-Celtic (or Goidelic) and P-Celtic (or Brythonic). In this introduction, we set out the context for our investigation by presenting the reader with a brief orientation to the history of the Celtic languages and their orthographies.

The Celtic languages form a subset of the larger Indo-European group of languages. Celtic is generally considered to be an entirely independent branch of Indo-European, although a minority of scholars have grouped it instead, along with the Italic languages, into an Italo-Celtic branch.³ Historically, Celtic seems to have been quite widely dispersed, being spoken not only in the British Isles and on the island of Ireland, but also in parts of continental Europe (Eska & Evans 1993). However, the continental varieties of Celtic died out quite early, by around 1 CE, leaving behind only a scanty corpus of short inscriptional texts. The Celtic languages that have been spoken in more recent history all belong to the Insular Celtic branch; they are: Irish, Scottish Gaelic, Manx, Welsh, Cornish, and Breton.⁴ These six languages divide further into two

¹ Department of Linguistics and English Language, County South, Lancaster University, Lancaster, United Kingdom, LA1 4YL, a.wilson@lancaster.ac.uk.

² Mathematical Institute, Slovak Academy of Science, Štefánikova 49, 814 73 Bratislava, Slovakia, and Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Mlynská dolina, 842 48 Bratislava, Slovakia, jmacutek@yhoo.com.

³ Fife (1993, p.5) describes this latter hypothesis as being "out of fashion", although a few researchers still support it.

⁴ There seems also to have been an Insular P-Celtic Cumbric language, though the evidence base for it is rather limited and there are no surviving texts in Cumbric (Schmidt 1993: 67).

sub-branches: Goidelic (consisting of Irish, Scottish Gaelic, and Manx) and Brythonic (consisting of Welsh, Cornish, and Breton). The two sub-branches are also commonly known, respectively, as Q-Celtic and P-Celtic (Schmidt 1993), owing to the different developments of the hypothesized Proto-Indo-European labiovelar phoneme */kʷ/; in Q-Celtic, the labial element disappeared, leaving behind a velar phoneme (/k/), whereas in P-Celtic it was the velar element that was lost, resulting in a bilabial phoneme (/p/). In the remainder of this paper, we will use the terminology of Q-Celtic and P-Celtic.

The earliest examples of Celtic texts are sparse and often fragmentary, consisting of short inscriptions or merely a few words (mostly proper names) within texts written in other languages. This absence of written evidence makes precise dating of early developments in Celtic difficult. However, the period from around 500 to 600 CE appears to be the time during which the early stages of the modern Celtic languages began to diverge recognizably from their common ancestors.

The Q-Celtic languages show a clear attested history, with the three modern languages all going back to an Early Irish base from which they gradually diverged, following the ‘export’ of the Irish language to Scotland and the Isle of Man between around 500-600 CE (Broderick 1993, Gillies 1993, MacKinnon 1993). The earliest Irish inscriptions are written in a runic script (known as Ogam or Ogham) and date from around the fourth to seventh centuries CE (Russell 1995, pp. 209-211). There are then, following this period, numerous manuscript texts written in the variety known as Old Irish, beginning in around the eighth century. Indeed, despite the ultimate divergence of the three spoken Q-Celtic languages, a literary form of Irish continued to unite the written cultures of the Q-Celtic lands until around the late seventeenth century, when printed texts in the Scottish Gaelic and Manx languages began to appear (MacCoinnich 2008).

There is somewhat less evidence for the earliest period of P-Celtic, with the appearance of texts beginning only in around the eighth to ninth centuries. Nevertheless, it is assumed that all three modern P-Celtic languages are descended from a common intermediate ancestor, Common Brythonic (or Brittonic). By around 500 to 600 CE, we begin to see the divergences that characterize the three modern P-Celtic languages (George 1993, Stephens 1993, Watkins 1993). Note, incidentally, that Breton, although its native-speaker base is restricted to north-western France (Brittany), is most certainly descended by migration from Insular P-Celtic and is therefore not a ‘Continental Celtic’ language, despite its geographical location.

All of the six Celtic languages are still spoken, written, and printed today, although their speaker numbers, and the everyday extent of their use, vary considerably. Welsh is almost certainly the most widely spoken on a daily basis, with Manx, Cornish, and Breton being the least widely used. Welsh, Irish, Scottish Gaelic, and Breton all have continuous native-speaker traditions dating back to their origins, with Welsh and Irish also having joint official status (alongside English) in Wales and the Republic of Ireland respectively. Scottish Gaelic does not have full official language status in Scotland, but it does nevertheless have a certain degree of legal recognition and is strongly supported, including the existence of Gaelic-medium schools and colleges, as well as publicly supported broadcast media. Breton does not have official language status in France, but it is nevertheless recognized and is available (within its region) as a subject at both school and university levels. Manx also had a native-speaker tradition until 1974, when the presumed last native speaker died; however, the language has continued to be taught and

used since then, so there has been virtually no hiatus in its tradition. Cornish has perhaps been the least fortunate of these languages. It was effectively a dead language by the end of the nineteenth century (and probably, to a large extent, even earlier than that), but it has since been revived by active enthusiasts and enjoys some recognition from bodies such as the BBC and the Church of England; the MP Andrew George, in 1997, was the first to speak some Cornish in the House of Commons (Mills 2010), and the language has even featured more recently in a UK-wide television advertisement for Kelly's Cornish ice cream.

As with any language, the orthographies of the six Celtic languages have evolved over time, and a detailed history of any of them would require an entire book, or even a series of books. Here, we present only the most basic historical points to support our analyses; further details may be found in the references cited, and in the standard reference books on the Celtic languages (such as Ball & Fife 1993, and Russell 1995).

The Irish orthography has its foundations in the early medieval Old Irish period, and is probably the most conservative of the six orthographies when compared with the phonology of the modern spoken language. The modern orthographic standard, which emerged from the linguistic revival movement towards the end of the nineteenth century, drew heavily on models from the seventeenth century Early Modern Irish period, giving it a somewhat archaic character (Ó Cearúil 1999). A series of spelling reforms during the middle of the twentieth century (with an official standard - *An Caighdeán Oifigiúil* - finally published in 1958) made some changes and simplifications to this. The example below (taken from Ó Cearúil 1999) shows a sentence from the Irish constitution in its original 1937 orthography (which is based on the older standard) and also from a 1960 edition (in the modernized official standard); the words which have been affected by the modernized spelling are highlighted in bold type in both versions:

[1937] *Deimhnigheann náisiún na hÉireann leis seo a gceart do-shannta, do-chlaoidhte, ceannasach chun cibé cinéal Riaghaltais is rogha leo féin do bhunú, chun a gcaidreamh le náisiúnaibh eile do chinneadh, agus chun a saoghal poilitidheachta is geilleagair is saoidheachta do chur ar aghaidh do réir dhúthchais is gnás a sinsear.*

[1960] *Deimhníonn náisiún na hÉireann leis seo a gceart doshannta, dochloíte, ceannasach chun cibé cineál Rialtais is rogha leo féin a bhunú, chun a gcaidreamh le náisiúin eile a chinneadh, agus chun a saol polaitíochta is geilleagair is saíochta a chur ar aghaidh de réir dhúchais is gnás a sinsear.*

The orthography of Scottish Gaelic evolved directly from – and was therefore strongly influenced by – the Irish orthography, although, over time, it took its own course in adapting to various current and local pronunciations (Ross 2016). However, unlike Irish, there has never been an authoritative central standard for Scottish Gaelic orthography, although, in recent years, reference has sometimes been made to the conventions adopted by Scotland's school examinations board, the Scottish Qualifications Authority (2009). The use of accents, in particular, has been rather variable across the history of Scottish Gaelic (see further below).

The Welsh orthography, which is strongly phonemic in character, has been largely settled since around the turn of the seventeenth century. It draws primarily on the spelling conventions of the 1588 Welsh Bible (Price 1984) and on the alphabetic inventory used by Davies in 1621

(Morris Jones 1913). A number of remaining uncertainties were dealt with in a set of recommendations published in 1928 (Price 1984).

Native-speaker Cornish had died out without ever arriving at a standardized orthography. In the course of the twentieth century revival movement (and continuing into the present century), several competing orthographies have evolved, including Kernewek Kemmyn, Unified Cornish, Tudor Cornish, Revived Late Cornish, and, most recently, the Standard Written Form. The texts processed in our study were published using Kernewek Kemmyn, which was adopted as the official orthography of the Cornish Language Board in 1987. This is a phonemic orthography devised by Ken George, based on his reconstructions of Middle Cornish phonology (for which see, e.g., George 1983).

Earlier Breton orthography was influenced by French (Russell 1995), though later revisions adapted it more closely to the phonology of Breton itself. Moves towards a standardization gradually began in the early nineteenth century, with the emergence of two standards in the early twentieth century: the KLT orthography of 1908-1911 and the ‘G’ orthography of 1902 (Hewitt 2017, p. 190). Three orthographies are now current: Peurunvan (which unified the earlier KLT and ‘G’ orthographies in 1941), Skolveurieg (1953), and Assimileg (1975). Of these, Peurunvan is the most widespread. Based on the presence of the form *frouezh* (which would be spelled *frouez* in Skolveurieg and *frwezh* in Assimileg), we can say that our texts are printed in the Peurunvan orthography.

It is a moot point whether Manx was ever written down earlier than the seventeenth century; certainly, the earliest surviving Manx text is Bishop Phillips’ translation of the Book of Common Prayer, dating from around 1610, although it was never published at the time (Thomson 1969). For this work, Phillips devised his own orthography that drew on both English and Welsh. However, it did not find widespread favour, and the orthography used today represents instead an evolution from that used for the earliest printed Manx book of 1707. This is generally assumed to have been based on the English orthography of the time, rather akin to the sorts of pronunciation transcriptions that one can find in a tourist phrase-book today; however, it has also been suggested that it may instead have a Scots base and may therefore have been imported from south-western Scotland, perhaps a century or so prior to the publication of that book (Ó hÍfearnáin 2007).⁵ Either way, it is very different from the Irish-based Gaelic orthographies used for the other Q-Celtic languages and exhibits a clear dominance of English grapheme-phoneme correspondences.

The following extracts (showing verse 10 of Psalm 51, from the texts we have used as our data) will give the reader some impression of the similarities and differences in the orthographies of the seven language varieties, when viewed in context. The reader will note, in particular, a number of identical or similar vocabulary items shared among the first three Q-Celtic languages (e.g. *spiorad*; *Dhia/Dhé*); these also occur in the Manx text, but are somewhat ‘hidden’ behind the English-based orthography (e.g. *spiorad* shows up as *spyrryd*, and *Dhia/Dhé* as *Yee*).

⁵ Ó hÍfearnáin (2007) quotes a lecture by Williams as suggesting that the Manx orthography has close affinities with the early sixteenth-century ‘Book of the Dean of Lismore’, which is a Gaelic manuscript written down using a Scots-based orthography. (Note that Scots is not the same thing as Scottish Gaelic, but is essentially a dialect of English.) This manuscript shows that Manx was not unique in using an English-based orthography to represent a Q-Celtic language, and there are also other examples from both Scotland and Ireland. However, unlike in the case of Manx, they never supplanted the traditional Irish-based orthography as the standard.

The word for ‘clean’ or ‘pure’ is, incidentally, shared by all of the Celtic languages, though it appears slightly differently in each case (*glan, glen, lân, lan, c'hlan*).

Early Modern Irish:

Cruthaigh ionnam croidhe glan, a Dhé; agus athnúaghaidh ann mo mheadhón spiorad iomlán.

Present-Day Irish:

Cruthaigh ionam croí glan, a Dhia, agus cruthaigh spiorad daingean as an nua ionam.

Scottish Gaelic:

Dean dhomh cridhe glan, O Dhia agus ath nuadhaich spiorad ceart an taobh a stigh dhiom.

Manx:

Croo aynym cree glen, O Yee: as jean ass-y-noa spyrryd cairagh cheu-sthie jee'm.

Welsh:

Crea galon lân ynof, O Dduw: ac adnewyddda yspryd uniawn o'm mewn.

Breton:

O Doue! Krou ennon ur galon c'hlan, nevesa em c'hreizon ur spered eeun.

Cornish:

Gwra kolonn lan ragov, A Dhyw: ha daswra ynnov vy spyrys len.

2. Language material and data

In order to control for text contents, length, type, etc., we based our analysis on parallel translation texts. We used the same set of parallel psalm translations that was used by Wilson & Harvey (2020). This gave us a sample of thirty psalms for each Celtic language, viz., Welsh, Cornish, Breton, Manx, Scottish Gaelic, Early Modern Irish, and Present-Day Irish, resulting in a total sample size of 210 individual texts. The thirty parallel psalms were selected randomly from psalms 1 to 111, as the psalms with numbers greater than 111 were not readily available in two of the languages (Breton and Early Modern Irish). The psalms included in the sample were numbers 1, 2, 3, 6, 12, 17, 20, 21, 27, 28, 29, 32, 41, 42, 43, 51, 54, 56, 79, 81, 84, 85, 90, 91, 95, 96, 97, 98, 99, and 101.

As mentioned by Wilson & Harvey (2020), it was not possible to control for every possible source of variation other than the language itself. Specifically, it was necessary to use texts which were composed and published across a fairly wide range of dates. Although publishing in Irish and Welsh had been established by the beginning of the seventeenth century (and our Early Modern Irish and Welsh samples date originally from ca.1640-1685 and ca.1620 respectively), Scottish Gaelic and Manx did not begin to be used as literary languages until around the eighteenth century. (Our data for these languages are, respectively, from 1794 and 1765.) A Cornish translation of the psalms was not produced until as late as 1997 (the one which we use here), and Breton translations were not published until the nineteenth century (with our data being originally from 1893, but revised somewhat in ca.2004-2011). To provide some indications regarding language change, it was nevertheless possible to include a later Irish translation from 2004 (which we call here ‘Present-Day Irish’). Some later translations do exist for some of the other languages (e.g. Welsh), but they were not readily available in machine-readable form.

The grapheme inventories that we used for this experiment follow the commonly documented alphabet inventories for the languages in question, plus any additional letters that occurred in these particular text samples. The latter largely occur in proper names of foreign origin (for instance, the grapheme <z> occurs in the Present-Day Irish texts as part of a single proper name), although they may also appear, very rarely, in other contexts. Digraphs were only included in our inventories if they are usually recognized as part of the alphabet when preparing published dictionaries, etc. Hence, for example, the Breton digraphs <ch> and <c'h> were included, but no attempt was made to consider digraphs (or other multigraphs) in the strict sense of one-to-one phoneme-grapheme correspondences; this would make the processing of texts difficult, especially with languages such as Irish, which has a lot of ‘silent’ letters. Letters which did not occur in the present corpus of texts were disregarded.

In all of these languages, the umlaut accent (i.e. two dots over a vowel) is always a diaeresis marker. In other words, it merely tells a reader, in a given context, not to pronounce a sequence of two vowels as a diphthong. Arguably, therefore, one can say that it is not an accent as such, and we therefore disregarded it when counting.

The orthographies of Breton, Cornish, Manx, and Present-Day Irish⁶ posed no other methodological problems. However, a few brief notes are necessary regarding the other languages.

Early Modern Irish. The first printed edition of these texts (from 1685) was typeset using a special Irish alphabet, based on earlier manuscript hands, which included an additional 16 special characters representing digraphs and trigraphs.⁷ Many later reprints, however, converted the texts to the ordinary Roman alphabet (though without otherwise changing the spelling); for instance, the early edition prepared by Robert Kirk in 1690, for use in Scotland, was such a conversion into Roman type (Durkacz 1978). As there is not - as far as we have been able to find - a full electronic edition of the original text, we did not seek to re-create the original Irish character inventory for counting purposes, at the risk of introducing errors. Irish and Roman characters have been used in parallel almost throughout the history of Irish printing (Ó Ciosáin 2004/2006).

Scottish Gaelic. The letter ‘J’ is not generally recognized as part of the Scottish Gaelic alphabet. In our texts, however, it occurred at the beginning of three proper names: *Jacob*, *Joseph*, and *Juda*. We considered this to be a free variant of <i> rather than a separate grapheme, so we merged these cases into the counts for <i>. The same texts in the Early Modern Irish version do use <i> in these contexts.

As regards accents, their use in Scottish Gaelic is not entirely standardized, even today. Ross (2016, pp. 171-206) has thoroughly researched the history of Gaelic accentuation between around 1750 and 2007, and has identified five practices that have been used during that time: using the acute accent only; using the grave accent only; using the grave accent with the acute accent for <e>; using the grave accent with the acute on both <e> and <o>; and using no accents at all. Our texts tend very much towards the ‘no accents’ end of the scale, with very low frequencies of accented forms (both á and à occur only once, and ò occurs four times). We therefore took the decision not to count these as separate graphemes, but to combine them with the corresponding unaccentuated forms.

⁶ As was mentioned above, grapheme <z> occurs once in the present-day Irish corpus as a part of a proper name; it was disregarded in our analyses.

⁷ For a history of early Irish-type printing, see Lynam (1924).

Welsh. The recognized inventory of the Welsh alphabet, especially as regards digraphs, has been somewhat fluid over the years. (A brief overview of the history can be found on pp. 9-11 of Morris Jones 1913.) Here, we have used the most widely recognized set of digraphs, as used to prepare present-day Welsh dictionaries; some earlier authors accepted additional ones.

As with Scottish Gaelic, the letter ‘J’ is not generally recognized as part of the Welsh alphabet, but it again occurred in our texts at the beginning of some proper names. We merged these cases with the counts for <i>.

The grave accent (unlike the circumflex and acute accents) is hardly ever used in modern Welsh, but it does occur in some of these older texts (i.e. it is not a digitization error or a printing artifact). In the Book of Psalms overall, it occurs 88 times on <y> and just 4 times on <o>. Here, we merged the counts for ÿ with the counts for <y>; ò did not occur.

Table 1
Grapheme frequencies in Celtic languages

Breton	Cornish	Irish (EM)	Irish (PD)	Manx	SG	Welsh							
a	3407	a	2689	a	4031	a	4236	a	2891	a	4608	a	2638
b	336	b	302	á	483	á	475	b	308	b	597	â	79
ch	25	c	14	b	555	b	509	c	838	c	1101	b	327
c’h	285	d	1367	c	1113	c	1035	d	955	d	1374	c	355
d	1238	e	2365	d	1623	d	1251	e	3622	e	1661	d	1148
e	3474	f	169	e	1318	e	1053	f	141	f	242	e	1754
ê	12	g	859	é	545	é	507	g	1040	g	1107	ê	8
f	111	h	1927	f	390	f	440	h	2209	h	3152	f	861
g	525	i	483	g	1302	g	983	i	1550	i	3235	g	785
h	722	j	48	h	3187	h	2170	j	289	l	852	h	344
i	1108	k	394	i	2662	i	2576	k	141	m	1060	i	1671
j	71	l	984	í	427	í	497	l	1321	n	2141	l	730
k	507	m	513	l	883	l	970	m	836	o	1334	m	517
l	913	n	1962	m	1027	m	966	n	2345	p	79	n	1870
m	530	o	1546	n	2049	n	2151	o	1983	r	1594	o	1086
n	1987	p	169	o	1468	o	1313	p	88	s	1370	ô	5
ñ	196	r	1667	ó	334	ó	288	q	15	t	1263	p	102
o	2097	s	1251	p	75	p	89	r	1854	u	1312	r	1578
p	267	t	649	r	1609	r	1550	s	1870			s	560
r	1746	u	298	s	1287	s	1166	t	1099			t	519
s	609	v	592	t	1179	t	1195	u	426			u	497
t	1124	w	1220	u	1147	u	681	v	304			û	1
u	1243	y	1613	ú	218	ú	252	w	63			w	1332
ù	186							y	2804			y	2387
v	642											ÿ	31
w	118											ng	93
y	17											ch	293
z	892											ph	10

rh	100
th	317
dd	879
ff	56
ll	199

3. Model for grapheme rank-frequency distributions

When one models linguistic data for which there is no “natural ordering” (such as, e.g., graphemes⁸ or words), one of the most reasonable approaches is to rank them – i.e., to order them decreasingly according to their frequencies (the most frequent item is then labelled 1, the second most frequent 2, etc.).

The negative hypergeometric distribution, shifted to the right by 1 (for its classic definition see Wimmer & Altmann 1999, pp. 465-468), with

$$P_x = \frac{\binom{M+x-2}{x-1} \binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}}, \quad x = 1, 2, \dots, n+1,$$

fits the grapheme rank-frequency distributions in many languages sufficiently well ($n+1$ is the inventory size, i.e., the number of graphemes a language uses; K and M are free parameters). This topic was studied extensively and systematically for Slavic languages: see Grzybek et al. (2009), Grzybek & Rusko (2009), and the references therein. Other languages for which some results were obtained include German (Best 2005; Grzybek 2007a), Irish and Manx (Wilson 2013), and several West African languages (Rovenchak & Vydrin 2010).

The goodness of fit of the model is evaluated, as has become usual in quantitative linguistics, in terms of the discrepancy coefficient

$$C = \frac{\chi^2}{N} = \frac{1}{N} \sum_{i=1}^k \frac{(f_i - NP_i)^2}{NP_i},$$

where f_i is the observed frequency of value i , P_i is the theoretical probability of value i (i.e. the one from the model), k is the number of classes into which the data are divided (often, but not always, the number of different values observed, or, in other words, the number of types), and N is the total number of observations (i.e. the number of tokens). Most often, the fit is considered satisfactory if $C \leq 0.02$; sometimes a “more tolerant” threshold with $C \leq 0.05$ is used (see Mačutek & Wimmer 2013). We note that p-values become useless for (very) large sample sizes, as one rejects virtually all null hypotheses for such samples (see e.g. Browne & Cudeck 1993 in general, and Mačutek & Wimmer 2013 specifically for models in quantitative linguistics). Parameter values and the discrepancy coefficients are presented in Table 2.

⁸ Specifically, for graphemes see a discussion in Koščová et al. (2016).

Table 2

Fitting the negative hypergeometric distribution to data from Table 2

	Breton	Cornish	Irish (EM)	Irish (PD)	Manx	SG	Welsh
<i>K</i>	3.3777	3.3652	2.5059	2.2746	3.7066	2.2674	4.0033
<i>M</i>	0.7545	0.9314	0.7326	0.6663	0.9427	0.7067	0.8782
<i>n</i>	27	22	22	22	23	17	32
<i>C</i>	0.0161	0.0049	0.0099	0.0092	0.0073	0.0245	0.0066

The fit of the negative hypergeometric distribution is very good for six out of the seven Celtic languages. Scottish Gaelic is the only exception, with the discrepancy coefficient slightly exceeding the threshold of 0.02 (which might be caused by the ‘pooling’ of the infrequent accented forms in these texts with their unaccentuated counterparts; see Section 2), but nevertheless the fit remains acceptable.

4. Towards a classification

Grapheme frequencies served as the basis of a classification of the Slavic languages, suggested by Koščová et al. (2016). We follow the same approach here.

Three indices of so-called qualitative variation are used to characterize the data, namely, the variance analogue

$$VA = 1 - \frac{\sum_{i=1}^K \left(f_i - \frac{N}{K}\right)^2}{\frac{N^2(K-1)}{K}},$$

the standard deviation analogue

$$SDA = 1 - \sqrt{\frac{\sum_{i=1}^K \left(f_i - \frac{N}{K}\right)^2}{\frac{N^2(K-1)}{K}}},$$

and the relative entropy

$$RE = \frac{-\sum_{i=1}^K \frac{f_i}{N} \log \frac{f_i}{N}}{\log K},$$

where N is the sample size, K the number of categories (i.e. the grapheme inventory size here), and f_i the observed frequency of the i -th category (log here denotes the natural logarithm).

These indices (Gadrich et al 2015 present an overview of several related data characteristics) differ from the classical variance (or standard deviation), as they do not depend on the order in which the categories (i.e., in this paper, the particular graphemes) are presented. All of them are normalized, so they take values from the interval $[0,1]$, with value 0 corresponding to exploiting one grapheme only, while value 1 is attained if all graphemes appear with the same frequency (i.e. in the case of the uniform distribution). Thus, they neutralize the differences between grapheme-inventory sizes in the languages under study, which have a strong impact on the variance (and other properties) of the grapheme rank-frequency distributions (see Grzybek et al. 2005 and Grzybek & Kelih 2005).

Then, the three above-mentioned indices are used to create the coordinates I_m (the x -coordinate) and S_m (the y -coordinate) of the modified Ord graph (introduced by Koščová et al. 2016), with

$$I_m = \frac{SDA}{VA},$$

and

$$S_m = \frac{RE}{SDA}.$$

The resulting modified Ord graph is presented in Figure 1. The ellipses indicate two clusters obtained by the k-means method applied to the coordinates I_m and S_m (a short description of the method can be found in Izenman 2008, pp. 423-424).

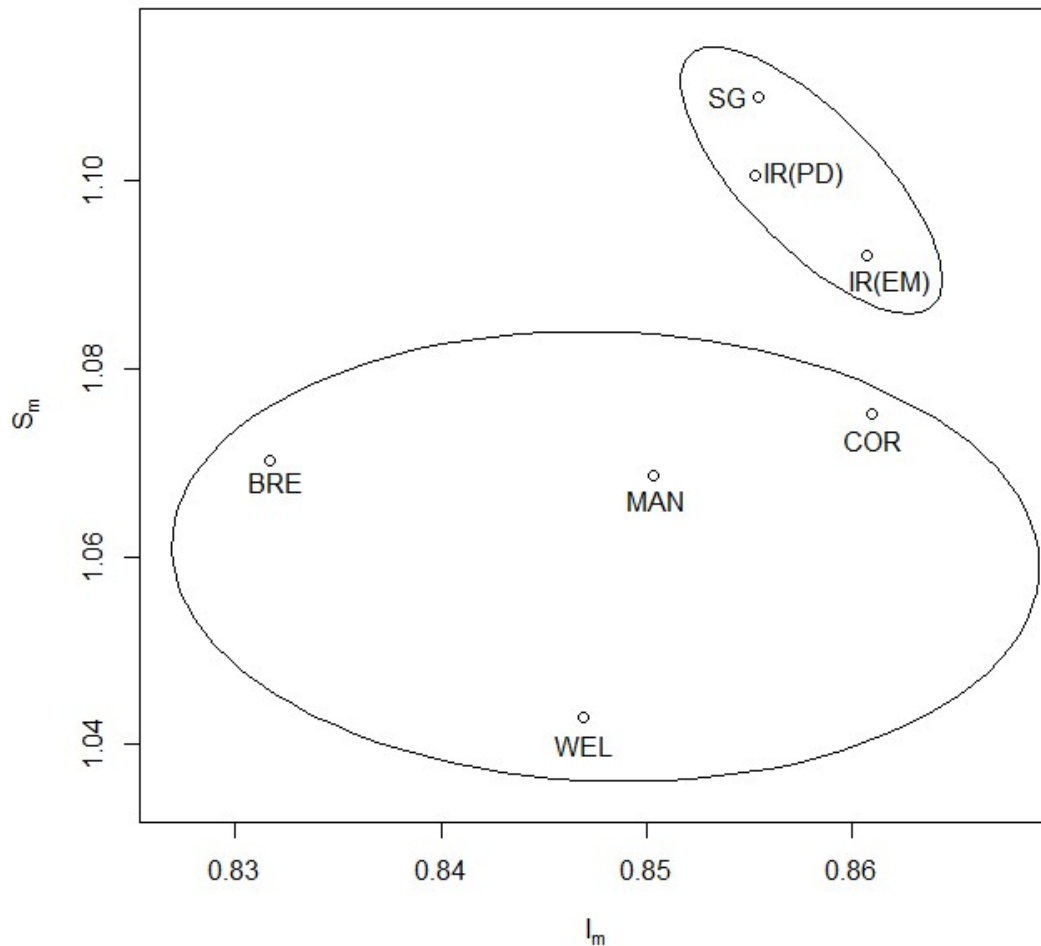


Figure 1 Modified Ord graph for grapheme frequencies from Table 1 (BRE – Breton, COR – Cornish, IR(EM) – Early Modern Irish, IR(PD) – Present-Day Irish, MAN – Manx, SG – Scottish Gaelic, WEL - Welsh), with cluster analysis applied to graph coordinates.

The two clusters closely resemble the traditional classification of the Celtic languages. One of them contains three Q-Celtic (or Goidelic) languages (both versions of Irish and Scottish

Gaelic), whereas the P-Celtic (or Brythonic) languages (Breton, Cornish, and Welsh) are contained in the other, together with Manx (from the Q-Celtic branch). The cluster containing Scottish Gaelic and the two varieties of Irish is quite compact, consistent with the common origin of both the languages and their orthographies. The fact that accent marks were very sparsely used in our Scottish Gaelic texts, and consequently disregarded in our grapheme counting, has had little effect on the clustering. Nor does date appear to have resulted in much variation: the two varieties of Irish occur quite closely together, despite a gap of more than three hundred years between them. In contrast, the cluster containing the three P-Celtic languages is quite diffuse. This can perhaps be accounted for by the very different histories of their orthographies, and the influences on them. For instance, Breton orthography developed alongside, and clearly under the influence of, French, whereas the primary linguistic contact of both Cornish and Welsh would have been English (although Latin and Norman French were also used, at various dates and in varying amounts, within the institutions of religion, education, and the law). Likewise, whereas the Welsh and Breton orthographies evolved gradually in stages since the middle ages, the Cornish Kernewek Kemmyn orthography is based on a thorough linguistic reconstruction of earlier Cornish phonology during the late twentieth century. Furthermore, Welsh recognizes a number of digraphs as forming a part of its alphabet inventory (as does Breton, though to a lesser extent), whereas Cornish does not. The ‘wrong’ position of Manx, grouped here with the P-Celtic rather than the other Q-Celtic languages, can also be explained by the history of its orthography, which – as outlined in Section 1 – developed on a basis of English (or Scots) phoneme-grapheme correspondences, as opposed to the traditional Irish-based orthographies of the other Q-Celtic languages.⁹

The clusters from Figure 1 are validated using silhouettes (see Rousseeuw 1987). A silhouette takes values from the interval $[-1,1]$. Objects with silhouette values close to 1 are considered well clustered, while values close to zero (or even negative values) suggest that an object is an outlier. Admittedly, our two clusters are not very well separated, with Cornish being very close to the Goidelic languages (its silhouette value is -0.01).

The unstable position of Cornish is highlighted also by the fact that it ‘migrates’ to the Q-Celtic group if the letter <z> is considered to be a part of the Irish alphabet (it occurs in the texts under analysis in a single proper name, see Section 2). However, given that we consider only grapheme frequencies here, and all other language units and properties are neglected, the results, although not ideal from the cluster-analysis point of view, are remarkably close to the traditional linguistic classification of the Celtic languages (and the only difference – Manx being classified with the P-Celtic languages – can be reasonably explained). In addition, it was shown that a small change in a grapheme inventory can cause a significant shift in the characteristics of the grapheme rank-frequency distribution (see Grzybek et al. 2005, Grzybek 2007b).

⁹ The importance of the orthographical principles was emphasized also by Koščová et al. (2016), where Belarusian was not included in the analysis, because its orthography (phonetically based) substantially differs from the other Slavic languages (in which letters code phonemes and partly morphophonemes). A discussion of the peculiar position of the Belarusian orthography among the orthographies of the other Slavic languages, and its consequences for mathematical modelling of its grapheme frequencies, can be found also in Kelih (2009).

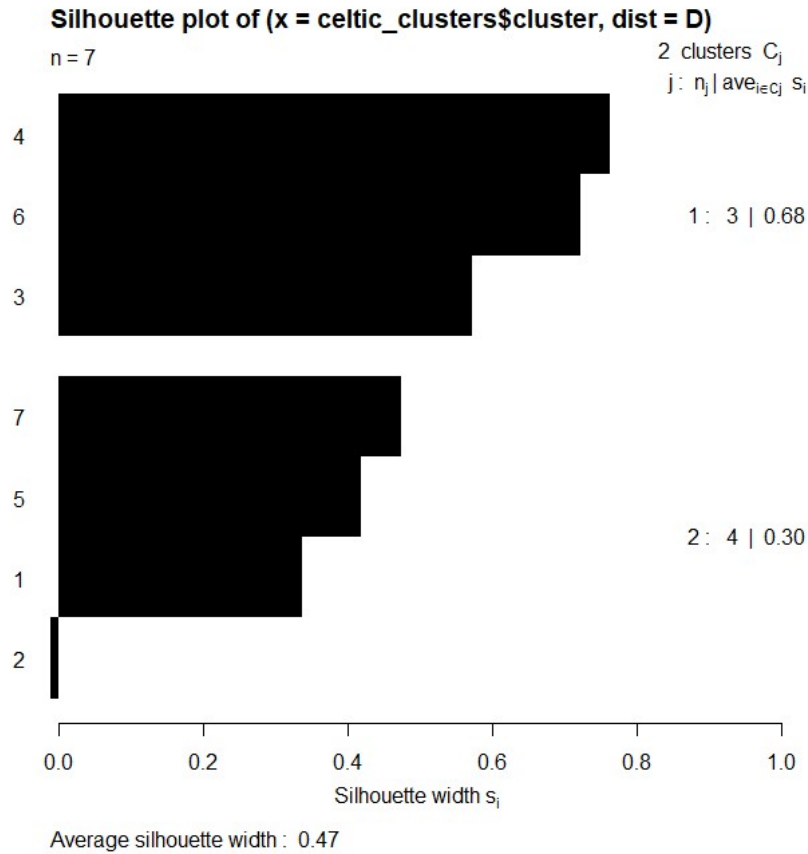


Figure 2 Silhouettes of clusters from Figure 1 (1 – Breton, 2 – Cornish, 3 – Early Modern Irish, 4 – present-day Irish, 5 – Manx, 6 – Scottish Gaelic, 7 - Welsh).

We note that the same clusters are obtained if the k-means clustering method is applied to the parameters K and M of the negative hypergeometric distribution from Table 2 (which is the mathematical model for grapheme rank-frequency distributions, see Section 3).

5. Conclusion

Our study produced two main results. First, we showed that the negative hypergeometric distribution is a good model for ranked grapheme frequencies in all of the Celtic languages. (This was already shown by Wilson 2013 for Irish and Manx.) We thus enlarge the set of languages for which the distribution achieves a good fit. We are able to suggest that the negative hypergeometric distribution is a promising candidate for a language law in the sense of Altmann (1993), since it is (a) a special case of a mathematically formulated language theory (Wimmer & Altmann 2005) and (b) corroborated on several languages from different language families. Most probably, it is a model for grapheme frequencies in languages which use alphabets or abjads. On the other hand, inventories of syllabaries and abugidas (see Daniels 1996 for an overview of writing systems) are too large, and Zipf-like distributions could be used to model their rank-frequency distributions. The paper by Radojičić et al. (2019) can serve as an indirect hint; it showed that the Zipf-Mandelbrot distribution (Wimmer & Altmann 1999, p. 666) fits the ranked

frequencies of syllables in Serbian.¹⁰ An hypothesis that syllable rank-frequency distributions are similar to word-frequency distributions (i.e. the Zipf distribution or one of its generalizations) was formulated by Strauss et al. (2008, p. 11).

Second, based exclusively on grapheme frequencies (which is a huge reduction of information – no other language units and properties were taken into account), we obtained a classification of Celtic languages which is linguistically interpretable if the peculiarities of the Manx orthography are considered. We followed the method (the modified Ord graph) which was introduced and applied to the Slavic languages by Koščová et al. (2016) with reasonable (i.e. linguistically justifiable) results. The approach must be tested on other language families before a conclusion on its general applicability can be reached.

Acknowledgement

Supported by research grant VEGA 2/0054/18 (J. Mačutek).

¹⁰ Serbian uses both the Cyrillic and Latin alphabets. But the inventory sizes of syllabic scripts should be closely related to the number of syllables which occur in the language.

References

- Altmann, G.** (1993). Science and linguistics. In: Köhler, R., Rieger, B.B. (eds.), *Contributions to Quantitative Linguistics: 3-10*. Dordrecht: Kluwer.
- Ball, M.J., Fife, J.** (eds.) (1993). *The Celtic Languages*. London: Routledge.
- Best, K.-H.** (2005). Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics 11*, 9-31.
- Broderick, G.** (1993). Manx. In: Ball, M.J., Fife, J. (eds.), *The Celtic Languages: 228-286*. London: Routledge.
- Browne, M.W., Cudeck, R.** (1993). Alternative ways of assessing model fit. In: Bollen, K.A., Long, J.S. (eds.), *Testing Structural Equation Models: 136-161*. Newbury Park (CA): SAGE.
- Daniels, P.T.** (1996). The study of writing systems. In: Daniels, P.T., Bright, W. (eds.), *The World's Writing Systems: 3-17*. Oxford: Oxford University Press.
- Durkacz, V.** (1978). The source of the language problem in Scottish education, 1688-1709. *The Scottish Historical Review 57(163)*, 28-39.
- Eska, J.F., Evans, D.E.** (1993). Continental Celtic. In: Ball, M.J., Fife, J. (eds.), *The Celtic Languages: 26-63*. London: Routledge.
- Fife, J.** (1993). Introduction. In: Ball, M.J., Fife, J. (eds.), *The Celtic Languages: 3-25*. London: Routledge.
- Gadrich, T., Bashkansky, E., Zitikis, R.** (2015). Assessing variation: A unifying approach for all scales of measurement. *Quality & Quantity 49*, 1145-1167.
- George, K.** (1983). A computer model of sound changes in Cornish. *Journal of the Association of Literary and Linguistic Computing 4*, 39-48.
- George, K.** (1993). Cornish. In: Ball, M.J., Fife, J. (eds.), *The Celtic Languages: 410-468*. London: Routledge.
- Gillies, W.** (1993). Scottish Gaelic. In: Ball, M.J., Fife, J. (eds.), *The Celtic Languages: 145-227*. London: Routledge.
- Grzybek, P.** (2007a). On the systematic and system-based study of grapheme frequencies: A re-analysis of German letter frequencies. *Glottometrics 15*, 82-91.
- Grzybek, P.** (2007b). What a difference an <<E>> makes: Die erleichterte Interpretation vom Graphemhäufigkeiten unter erschwerten Bedingungen. In: Deutschmann, P. (ed.), *Kritik und Phrase. Festschrift für Wolfgang Eismann zum 65. Geburtstag: 105-128*. Wien: Praesens.
- Grzybek, P., Kelih, E.** (2005). Towards a general model of grapheme frequencies in Slavic languages. In: Garabík, R. (ed.), *Computer Treatment of Slavic and East European Languages: 73-87*. Bratislava: Veda.
- Grzybek, P., Kelih, E., Altmann, G.** (2005). Graphemhäufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfang – eine Nebenbemerkung zur Diskussion um das ë. *Anzeiger für Slavische Philologie 33*, 117-140.
- Grzybek, P., Kelih, E., Stadlober, E.** (2009). Slavic letter frequencies: A common discrete model and regular parameter behavior? In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 17-33*. Lüdenscheid: RAM-Verlag.

- Grzybek, P., Rusko, M.** (2009). Letter, grapheme and (allo-)phone frequencies: The case of Slovak. *Glottology* 2(1), 30-48.
- Hewitt, S.** (2017). Breton orthographies: an increasingly awkward fit. In: Jones, M.C., Mooney, D. (eds.), *Creating Orthographies for Endangered Languages: 190-234*. Cambridge: Cambridge University Press.
- Izenman, A.J.** (2008). *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning*. Berlin: Springer.
- Kelih, E.** (2009). Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle. *Glottometrics* 18, 53-68.
- Koščová, M., Mačutek, J., Kelih, E.** (2016). A data-based classification of Slavic languages: Indices of qualitative variation applied to grapheme frequencies. *Journal of Quantitative Linguistics* 23, 177-190.
- Lynam, E.W.** (1924). The Irish character in print, 1571-1923. *The Library, series 4*, 4(4), 286-325.
- MacCoinnich, A.** (2008). Where and how was Gaelic written in late medieval and early modern Scotland? Orthographic practices and cultural identities. *Scottish Gaelic Studies* 24, 309-356.
- Mac Eoin, G.** (1993). Irish. In: Ball, M.J., Fife, J. (eds.), *The Celtic Languages: 101-144*. London: Routledge.
- MacKinnon, K.** (1993). Scottish Gaelic today: social history and contemporary status. In: Ball, M.J., Fife, J. (eds.), *The Celtic Languages: 491-535*. London: Routledge.
- Mačutek, J., Wimmer, G.** (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics* 20, 227-240.
- Mills, J.** (2010). Genocide and ethnocide: The suppression of the Cornish language. In: Partridge, J. (ed.), *Interfaces in Language: 189-206*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Morris Jones, J.** (1913). *A Welsh Grammar, Historical and Comparative*. Oxford: Clarendon Press.
- Ó Cearúil, M.** (1999). *Bunreacht na hÉireann: A Study of the Irish Text*. Dublin: The Stationery Office.
- Ó Ciosáin, N.** (2004/2006). Print and Irish, 1570-1900: An exception among the Celtic languages? *Radharc* 5/7, 73-106.
- Ó hIfearnáin, T.** (2007). Manx orthography and language ideology in the Gaelic continuum. In Eloy, J.-M., Ó hIfearnáin, T. (eds.), *Langues Proches-Langues Collatérales/Near Languages-Collateral Languages. Actes du Colloque International Réuni à Limerick, du 16 au 18 Juin, 2005 : 159-170*.
- Price, G.** (1984). Welsh as a literary, standard, and official language. In: Ball, M.J., Jones, G.E. (eds.), *Welsh Phonology: 262-269*. Cardiff: University of Wales Press.
- Radojičić, M., Lazić, B., Kaplar, S., Stanković, R., Obradović, I., Mačutek, J., Leššová, L.** (2019). Frequency and length of syllables in Serbian. *Glottometrics* 45, 114-123.
- Ross, S.** (2016). *The Standardisation of Scottish Gaelic Orthography 1750-2007: A Corpus Approach*. PhD thesis, University of Glasgow.
- Rousseuw, P.J.** (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53-65.

- Rovenchak, A., Vydrin, V.** (2010). Quantitative properties of the Nko writing system. In: Grzybek, P., Kelih, E., Mačutek, J. (eds), *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives: 171-181*. Wien: Praesens.
- Russell, P.** (1995). *An Introduction to the Celtic Languages*. London: Longman.
- Scottish Qualifications Authority** (2009). *Gaelic Orthographic Conventions*. Retrieved April 20, 2020, from https://www.sqa.org.uk/sqa/files_ccc/SQA-Gaelic_Orthographic_Conventions-En-e.pdf
- Schmidt, K.-H.** (1993). Insular Celtic P- and Q-Celtic. In: Ball, M.J., Fife, J. (eds.), *The Celtic Languages: 64-98*. London: Routledge.
- Stephens, J.** (1993). Breton. In: Ball, M.J., Fife, J. (eds.), *The Celtic Languages: 349-409*. London: Routledge.
- Strauss, U., Fan, F., Altmann, G.** (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM-Verlag.
- Thomson, R.L.** (1969). The study of Manx Gaelic. *Proceedings of the British Academy* 55, 177-210.
- Watkins, T.A.** (1993). Welsh. In: Ball, M.J., Fife, J. (eds.), *The Celtic Languages: 289-348*. London: Routledge.
- Wilson, A.** (2013). Probability distributions of grapheme frequencies in Irish and Manx. *Journal of Quantitative Linguistics* 20, 169-177.
- Wilson, A., Harvey, R.** (2020). Using rank-frequency and type-token statistics to compare morphological typology in the Celtic languages. *Journal of Quantitative Linguistics* 27, 159-186.
- Wimmer, G., Altmann, G.** (1999). *Thesaurus of Univariate Discrete Probability Distributions*. Essen: Stamm.
- Wimmer, G., Altmann, G.** (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807*. Berlin, New York: de Gruyter.

Data sources

Welsh

<http://justus.anglican.org/resources/bcp/>

Breton

<http://bibl.monsite-orange.fr/>

Manx

<http://mannin.info/MHF/>

Cornish

Courtesy of Keith Syed, Cornish Bible Project

Scottish Gaelic

Digital Archive of Scottish Gaelic, Text No. 152 (<http://www.dasg.ac.uk>)

Present-Day Irish

<https://www.ireland.anglican.org/prayer-worship/book-of-common-prayer/2004-texts>

Early Modern Irish

<http://macmate.macace.net/~macfhionn@macace.net/index.html/Psailm.html>

Finding the Author of a Translation. An Experiment in Authorship Attribution Using Machine Learning Methods in Original Texts and Translations of the Same Author

George Mikros¹

Abstract

The aim of this paper is to investigate whether authorship attribution methods can be used effectively in order to discover the author of an anonymous translation. We constructed a corpus of Modern Greek newswire texts and a corpus of translations written from the same author/translator. In addition, we created a corpus of original Modern Greek texts from many different authors matching the genre, topic and medium of the author's/translator's corpus. In all the corpora we measured a wide variety of stylometric features ("classic" stylometric features, word and character n-grams). We performed three different experiments of authorship attribution in order to test whether machine learning techniques (support vector machines (SVMs)) could reliably a) discriminate the author's/translator's texts from the other authors and b) recognize that the translations included in the corpus were indeed written by the author/translator even when the training corpus does not have translation samples. The results seem promising since our method using feature selection and class balance resampling achieved a 0.89 recall, recognizing 97 out of 109 chunks of translated text using an SVM model trained only in original texts from the author/translator and the other authors.

Keywords: authorship attribution; translation; Modern Greek; n-grams; support vector machines

1. Introduction

Authorship identification refers to the connection of a text of unknown authorship to a specific author using a set of quantifiable text features as indicators of the author's style. Language usage has long been recognized as a carrier of various extralinguistic information such as historical period, dialect (both geographical and social), author's sex and age, ideology etc. The history of stylometry is multilinear and as Grzybek (2014, p. 62) claims, "it has been a combination and overlapping of different lines, converging only partly, at particular points of time or periods". Lorenzo Valla (1407–1457) was one of the first philologists who exploited stylometric evidence to disprove the originality of *Donation of Constantine*, a Roman imperial decree issued by Emperor Constantine I in order to transfer authority over Rome and the western part of the Roman Empire to the Pope. Investigations that linked stylometry to authorship identification started to appear in the 19th century with the extensive quantitative study of Shakespeare's plays by a number of scholars clustered around the "New Shakspeare Society" (Sawyer, 2006).

Significant progress has been noticed since then both in the statistical methods employed for this task and the textual features used to identify each author's style. Major landmarks in the field were the authorship analysis of *The Federalist Papers* performed by Mosteller & Wallace (1984) and the multivariate statistical methods introduced by Burrows (1987, 1989, 1992) and his associates (Burrows & Craig, 1994; Burrows & Hassal, 1988).

¹ Hamad Bin Khalifa University, Qatar, gmikros@gmail.com.

Since the late 1990s authorship identification has known a new impetus based on developments in a number of key research areas such as Information Retrieval, Machine learning and Natural Language Processing (Stamatatos, 2009). Furthermore, a vast amount of text is now available online and Web 2.0 has added to the now standard internet genres of email, web page and online forum message new forms of online expression such as blogs, tweets and instant messaging.

2. Stylometric analysis and authorship attribution in translations

Authorship identification techniques have been extensively used for the attribution of texts in specific authors as long as these texts are produced originally from one of them in his/her mother tongue. However, there is little experience in testing authorship identification methods in cases where the author is not producing his own text but translates a text of a different author written in a foreign language. Stylometric theory assumes that each author possesses a distinct, unique “writeprint” which is expressed quantitatively through the idiosyncratic occurrence variation of its most frequent linguistic structures and various indices of unconscious linguistic behavior such as lexical “richness” formulas, word and sentence lengths etc. If such a “writeprint” exists, then it should be text topic and genre neutral. Translations test the theory of “writeprint” in its extreme.

If the identity of the translator survives through the process of translation and can be traced in a text that was originally written in another language and by a different author then stylometric authorship attribution would increase its methodological robustness and reliability. Authorship attribution in a translation is actually a decomposition problem where the researcher must find the optimum way to identify and extract four different effects exerted in the linguistic structure of the translated text:

- a) Text topic and genre
- b) Original author
- c) Translation process
- d) Translator

Each one of the above factors shapes the stylometric profile of the text in a unique way contributing an unknown amount of variation to the final script. In order to accurately identify the translator of a text we should be able to estimate and isolate the effects of all the above parameters. There is a significant amount of research in factors (a) and (b). Topic and genre classification has been an active research field with amazing results and standard datasets (e.g. Lewis, Yang, Rose, & Li, 2004) and methodology (e.g. Kessler, Numberg, & Schütze, 1997; Santini, 2004; Sebastiani, 2002). The same is true for classic authorship attribution, which has recently attracted significant attention in the broader context of information retrieval and text mining (Juola, 2007, 2008; Stamatatos, 2009).

The translation process has long been studied both as theory and as praxis. However, stylometric analyses of translations are not common. Only recently a number of studies have been published examining whether translated texts are stylometrically “normal” texts or have specific peculiarities that constitute different genres, a separate “dialect” within a language commonly referred to as the “third code” (Frawley, 1984) or “translationese” (Gellerstam, 1986).

Baroni & Bernardini (2006) use machine learning algorithms (support vector machines (SVMs)) and a variety of features including word unigrams, bigrams and trigrams in order to predict the status of a text (original or translation) in two parallel monolingual (Italian and English) corpora from the geopolitical domain. An ensemble of SVMs reached 86.7% accuracy, outperforming the average recognition rates of ten human subjects. Their results support the theory that translations are a distinct genre and can be recognized using algorithms and features commonly employed in other text classification tasks.

Related research aims were pursued by Koppel & Ordan (2011), who tested two different claims: a) Translations from different source languages into the same target language are sufficiently different that a classifier can accurately identify the source language of a given translated text; b) Translations from a mix of source languages are sufficiently distinct from texts originally written in the target language for a learned classifier to accurately determine if a given text is translated or original. The corpus used was EuroParl, a comparable corpus which consists of transcripts of addresses given in the European Parliament translated from 11 different languages to English as well as original English. The researchers counted the frequency of 300 function words in 200 chunks of translated text from five corpora corresponding to five different source languages. Using Bayesian logistic regression as a learning method they achieved a rate of 92.7% correctly recognizing translated from original English texts. In a subsequent experiment using the same features and learning algorithm they tried to predict the status of a text in a specific source language using a model trained in different source languages. In this scenario the accuracy of the classification was lower than in the previous experiment and was analogous to the degree of similarity of the specific language pair. The above result shows clearly that although a general “translationese” exists, there are also present strong interference effects produced by the source language and sometimes these can be very effective.

There are a few studies that investigate the stylometric properties of translated texts and whether these can be used in order to extract the translator’s authorship. One of the oldest was an investigation of the authorship of an anonymous translation of the military history of Charles XII (King of Sweden, 1697–1718) from French to English. The study was undertaken by Michael and Jill Farringdon (Farringdon & Farringdon, 1979) and the candidate author was Henry Fielding, a distinctive novelist. The study was based on frequency counts of idiomatic words existing in both Fielding’s writings and the anonymous translation. Furthermore, a corpus of authors contemporary to Fielding was created in order to construct a baseline of word usage. The main conclusion was that Fielding was indeed the anonymous translator, although the corpus analysis and the quantitative methodology employed were somehow simplistic.

In another study Burwick & McKusick (2007) attempt to show that the author of a specific anonymous translation of Goethe’s *Faustus* was indeed the famous romantic English poet Samuel Taylor Coleridge (1772–1834). The corpus used for this study was compiled using the anonymous 1821 *Faustus* and five other translations of the play by known authors. Furthermore they used two plays by Coleridge and another two plays translated by him. The features used were the frequencies of words of two letters, three letters and so on, up to words of eight letters, and the frequencies of ten particular words (he, in, now, of, shall, then, this, to, which, and your) which they find are used at different rates in a Coleridge play and in a group of translations of *Faustus* by other writers. In order to test whether there are statistically significant differences between the frequencies of the features in the anonymous translation and the translations with certain authorship they used the chi-squared test. Burwick & McKusick’s study concludes that Coleridge was indeed the author of the specific translation. From the methodological point of view the specific study has been criticized for both the corpus and the features used (Craig, 2008; Uhlig, 2010; Whissell, 2011).

Recently Hedegaard & Simonsen (2011) demonstrated that the translator influences heavily the text, and his or her contribution can be traced using the most frequent words as features. Using a mixed corpus of original and translated texts they used semantic features (frames from the semantic net FrameNet), frequent word and character n-grams in order to identify the authors behind the translations and the translators behind the authors. In the first case the author attribution accuracy using his/her translation reached 90%. In the second case, the researchers identified the translator with an accuracy of 90.9% using the 400 most frequent words in the corpus.

3. Corpus development

We are interested whether we can identify whether a specific author is at the same time the translator of text given the following prerequisites:

- We have a corpus of texts from this author written originally in his/her mother tongue.
- We have a corpus of translations which have been produced by this author.
- We have a corpus of texts written from other candidate authors in their mother tongue.

In order to research the above authorship identification scenario we needed a corpus from a person who is an active author in his/her mother tongue and at the same time a professional translator. We identified an author who regularly publishes articles for culture topics in a wide-circulation Greek newspaper (*Kathimerini*) and at the same time is a professional translator (French to Greek) and an academic with specialization in translation theory and methodology. She offered us 151 articles she had published in the newspaper and four book-wide translations she had created for a Greek publishing house. In order to equalize the text length among the different corpora the four translations were segmented into 109 equally sized text chunks of 800 words each. Furthermore, we developed a corpus of candidate authors with texts originally produced in their mother tongue (Greek) on the same topic (culture) and genre (newspaper articles) as our author/translator. The basic descriptive statistics of the three corpora developed are shown in Table 1 below:

Table 1
Descriptive statistics of the corpora used in the present research

<i>Corpora</i>	<i>Texts (N)</i>	<i>Corpus size (in words)</i>	<i>Average text length (in words)</i>	<i>Std dev of text length</i>	<i>Min. text length</i>	<i>Max. text length</i>
Author A (Texts originally produced in mother tongue)	151	146,424	969.7	776	232	6,696
Candidate authors (Texts originally produced in mother tongue)	618	509,078	823.8	534	34	5,156
Author A (Translations)	109	87,292	800.0	0	800	800
Total	878	742,794				

The quantitative profile of the three corpora is very similar since it contains texts of nearly equal average size and with wide variation between the minimum and maximum length (high standard deviation). Furthermore it is highly homogeneous in terms of topic and genre as far as the original texts are concerned.

4. Stylometric features and classification algorithms

Authorship identification research has used an impressive array of stylometric features (Grieve, 2007; Juola, 2008; Stamatatos, 2009) ranging from characters to syntactic and semantic units. We selected our features taking into consideration the best practices established in authorship identification research published from the 1990s till today. After initial experimentation with many different feature groups we decided to use both classic stylometric features (e.g. word length, lexical “richness” indices etc.) and features borrowed from information retrieval research (e.g. character and word n-grams). We used single feature groups and in a later stage we combined different feature groups, a methodology that gave us the best results and is generally accepted as the better strategy (Juola, 2008, p. 269; Zheng, Li, Chen, & Huang, 2006, p. 380). In all our frequency-based features we calculated their normalized frequency in order to avoid

text length bias in subsequent calculations. Feature normalized frequency (*fnf*) of a feature *i* in a document *j* is defined as follows:

$$fnf_{i,j} = \frac{frf_{i,j} \cdot 100}{\sum_k frf_{k,j}} \quad (1)$$

where *frf_{i,j}* is the raw frequency of the feature *i* in the document *j* multiplied by 100 and divided by the sum of number of occurrences of all features (*k*) in the document *j*, that is, the size of the document $|j|$. The features we finally used in our system were the following:

1) Common stylometric features (CSF) (22 features)

a) Lexical “richness”

- (1) Yule’s K: Vocabulary richness index that exhibits stability in different text sizes (Tweedie & Baayen, 1998).
- (2) Lexical density: The ratio of functional to content words in the text, also known as functional density (Miranda & Calle, 2007).
- (3) % of hapax and dis legomena: The percentage of words with frequency 1 and 2 in the text segment.
- (4) Dis/hapax legomena: The ratio of dis legomena to hapax legomena in the text segment, indicative of authorship style (Hoover, 2003).
- (5) Relative entropy: Is defined as the quotient between the entropy of the text and its maximum entropy multiplied by 100. Maximum entropy for a text is calculated if we assume that every word appears with frequency 1 (Oakes, 1998, p. 62).

b) Word level measures

- (1) Average word length (per text) measured in letters.
- (2) Word length distribution: The normalized frequency of words 1, 2, 3 ... 14 letters long.

2) Information retrieval features (IRF) (2,131 features)

a) Word unigrams

Word frequency is considered among the oldest and most reliable indicators of authorship, outperforming sometimes even the n-gram features (Allison & Guthrie, 2008; Coyotl-Morales, Villaseñor-Pineda, Montes-y-Gómez, & Rosso, 2006; Diederich, Kindermann, Leopold, & Paass, 2003). The number of words used in authorship identification studies varies from the 100 most frequent words of the training corpus (Burrows, 1987) to many thousands (Madigan et al., 2005) depending on the classification algorithm and the dataset size. We decided to calculate the normalized frequencies of the 100 most frequent words (unigrams).

b) Word bigrams

Word n-grams with n=2 have long been used in authorship attribution with success (Coyotl-Morales et al., 2006; Gehrke, 2008; Peng, Schuurmans, & Wang, 2004). We detected the 1,000 most frequent word bigrams in the training corpus and calculated their normalized frequency in all corpora.

c) Character unigrams

The normalized frequency of each letter in the text segment. We measured in total 31 letters (we calculated separately the stressed and the unstressed vowels).

d) Character bigrams

Character n-grams provide a robust indicator of authorship and many studies have confirmed their superiority in large datasets (Grieve, 2007; Koppel, Schler, & Argamon,

2011; Luyckx & Daelemans, 2011). We extracted the 1,000 most frequent character bigrams of the training corpus and calculated their normalized frequency in all corpora. The total feature vector used contained 2,153 features. The most frequent unigrams were detected using a custom PERL script which identified tokens as a sequence of alphanumeric characters using the regular expression `\w+`. Later a custom PERL script took as input a list of the most frequent tokens in the training corpus and produced a vector containing normalized frequency of occurrence of each token in all the texts contained in the datasets.

The most frequent n-grams were detected using the Ngram Statistics Package (NSP) (Banerjee & Pedersen, 2003), a PERL module designed for word and character n-gram identification. Tokenization in n-gram identification followed the following rules:

- A token was identified as any sequence of alphanumeric characters using the following regular expression: `\w+`
- As tokens were identified also the punctuation marks were defined in the following regular expression: `[\.,;\:\?\!]`. Punctuation usage often reflects author-related stylistic habits (Mikros, 2007) and n-grams with punctuation can better capture these stylistic idiosyncrasies.
- All tokens were converted to lowercase.

Output files from NSP were converted to vectors using a custom PERL script which aggregated n-gram counts from each text file and normalized them using feature normalized frequency.

Given the rather large dimensionality of the extracted features, the task of training models is time- and memory-consuming, even for moderate numbers of training instances. Therefore a method for efficiently solving the large-scale classification problem was required.

For the purposes of the classification tasks, we used the sequential minimal optimization (SMO) algorithm (Platt, 1999), an efficient implementation of support vector machines (SVMs) (Vapnik, 1995) in both memory and speed issues. SVMs are considered one of the most robust algorithms in text classification tasks (Joachims, 1998) and have been applied with success in many authorship identification problems (Diederich et al., 2003; Escalante, Solorio, & Montes-y-Gómez, 2011; Houvardas & Stamatatos, 2006; Zheng et al., 2006).

5. Attribution experiments

In order to validate the attribution efficiency of our approach we ran three different attribution experiments using different combinations of the above-mentioned corpora. All classification experiments which didn't use a specific test sample were evaluated using a ten-fold cross-validation procedure. Evaluation metrics used are borrowed from information retrieval research and are described below:

- Precision: Is defined as the number of all correctly attributed documents to an author divided by the total number of documents the classifier considers that belong to the author.
- Recall: Is defined as the number of all correctly attributed documents to an author divided by all the documents that he/she has actually written.
- F_1 : Harmonic mean of the precision and recall values. Its formula is presented below:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

a. Original texts of the author/translator vs. the texts of the other candidate authors

This is a classic authorship attribution experiment with the primary aim the best discrimination of the author's/translator's style from the other candidate authors. This experiment can be seen as a two-class classification problem. More specifically we treat all the candidate authors as

one class (O) and the original texts of the author/translator as the other class (A). The average cross-validated accuracy of the classification was excellent (98.8%). Precision, recall and F_1 values for each class are reported in the following table (Table 2):

Table 2
Evaluation metrics for the SVM classification with O and A classes

Classes	Precision	Recall	F_1
<i>O</i>	0.992	0.994	0.993
<i>A</i>	0.963	0.954	0.959

From the above results we conclude that the style of the author/translator can be nearly perfectly discriminated by the style of the other candidate authors using the specific combination of the selected features and SVM.

b. Original texts and translations of the author/translator vs. the texts of the other candidate authors

In this experiment we augment the corpus of the original texts of the author/translator with her translations. We perform again a two-class classification taking as one class (A/T) all the written samples of the author/translator (original articles + translations) and as the other class (O) all the articles of the other candidate authors. The average cross-validated accuracy of the classification was again excellent (97.4%). Precision, recall and F_1 values for each class are reported in the following table (Table 3):

Table 3
Evaluation metrics for the SVM classification with O and A classes

Classes	Precision	Recall	F_1
<i>O</i>	0.989	0.974	0.981
<i>A/T</i>	0.941	0.973	0.957

The addition of the translations in the corpus of the author/translator had a small negative impact on the precision and a reverse small positive impact in the recall of the respective class (A/T). This can be interpreted as an improvement in the sensitivity of the machine learning since the algorithm can now recognize better all the relevant texts written by the author/translator (higher recall) at the cost of making more misclassifications and losing (small) discriminatory power between the two classes. Concluding, the stylometric information provided by the translations in general didn't decrease the overall accuracy of the classification and helped the algorithm to better capture the stylometric profile of the author.

c. Original texts of the author/translator and texts from the other candidate authors vs. her translations.

In this experiment we ran two different sub-experiments constructing two different training corpora and two different test corpora as follows:

- a) A training corpus of the original texts of the author/translator and the texts of the other candidate authors enriched with a random sample of translation chunks (50%) and a test corpus with the remaining 50% of the translation chunks.
- b) A training corpus of the original texts of the author/translator and the texts of the other candidate authors and a test corpus consisting only of the translations of the author/translator.

In these experiments we are investigating whether our model can recognize that specific texts which are translations belong to the same author as texts written originally in her mother tongue. In sub-experiment (a) we include in the training corpus some examples of translations while in

sub-experiment (b) we test whether translations can be attributed to a person even when we haven't included translation examples in the training corpus. Since in both cases the test corpus contains only the class Translation (T), we use only recall as evaluation metric.

The recall value for sub-experiment (a) is 0.927, which means that if we include samples of translations in the training corpus we can reliably detect if a translation has been written by a specific author. In sub-experiment (b) however the recall value was 0.23. This dramatic decrease in the performance of the algorithm can be interpreted by the complete absence of translated text samples in the training corpus.

In order to enhance translation attribution, we performed feature selection using a genetic search algorithm (Goldberg, 1989) and retested the now reduced feature vector (735 features). The recall value increased and reached 0.323.

Since our training corpus has many more texts from the class "other candidate authors" (618) compared with the class "author/translator" (151 original texts/109 translation chunks), we considered reducing the size of the former class. This issue is widely known in machine learning literature as the "class imbalance problem" of the training data and can diminish the performance of most machine learning methods, including those that present robustness in noisy data such as the SVMs (Wu & Chang, 2003). For this reason we made a random resampling of the class "other candidate authors" preserving 25% of the original data (156 texts). We trained once again the SVM model in the original texts of both classes and we measured its recall in the 109 translated text segments. Its recall value without any feature selection was 0.37, which was higher than the other attempts but still low for practical purposes. The last step in this optimization procedure was to apply feature selection in the equal class size dataset. We reapplied the genetic search algorithm and obtained an impressive increase in the recall of the translated text segments. The SVM model recognized as a translation of the author/translator 97 of the 109 texts, representing a recall value of 0.89. The increase of the recall value as a function of the various optimizations can be seen in the following diagram (Figure 1):

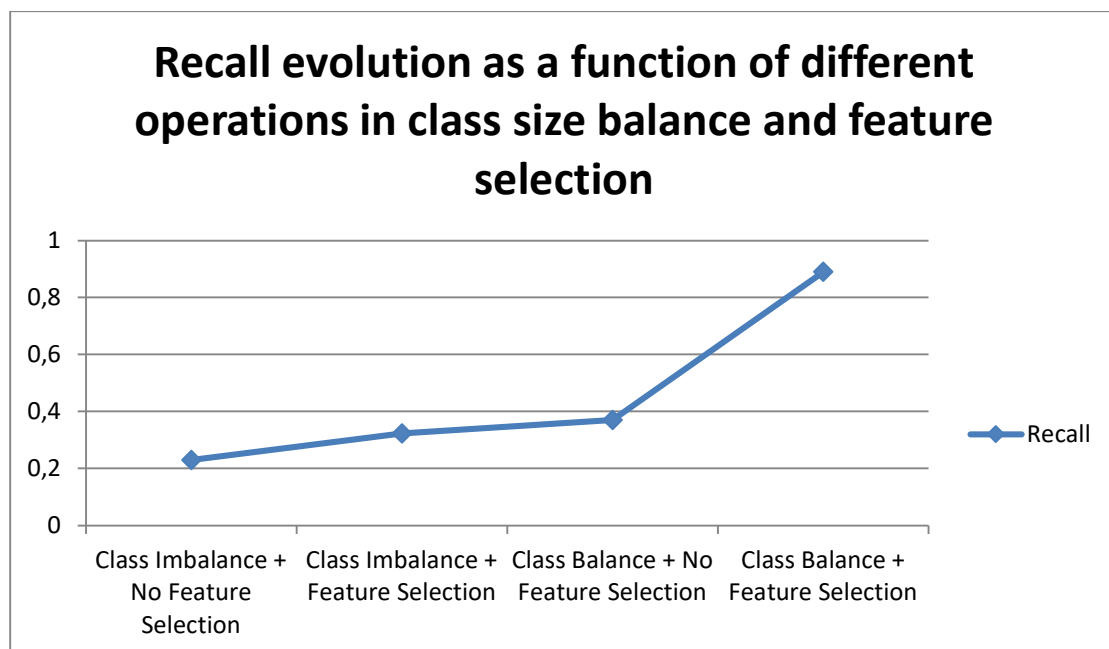


Figure 1 Recall of translated text segments using SVMs and varying class size balance and feature selection in the training corpus

As can be seen the best recall value was obtained when both feature selection and class balance were applied to the training corpus. It seems that optimal results are gained only when both

optimizations are applied. On the contrary, when we use each one alone the recall increase is rather small compared to their simultaneous application.

6. Conclusions

In the present study we investigated whether we can reliably attribute translations in an author using the original texts he/she produces in his or her mother tongue as a training corpus. In order to answer our main research question we compiled a corpus which contained original texts from an author who at the same time was a professional translator. Furthermore, we enriched the corpus with texts from other authors in the same topic, genre and medium of the original texts of the author/translator. We calculated a wide array of stylometric features and using support vector machines as classification algorithm we investigated the attribution accuracy in different scenarios. The main conclusions from the above-described experimental procedure are:

- We can obtain high accuracy in authorship attribution even in the case where the one class is a specific author and the other class is a merged corpus of texts from different authors.
- We can obtain equally high accuracy in authorship attribution when we add translations to the corpus of the original texts of the author/translator. Translations seem to carry extra stylistic information related to the author/translator since the recall of attribution increased after adding the translations in the training corpus.
- We can reliably attribute translations to an author if we have a training corpus that contains at least some samples of translations that belong with certainty to the specific author.
- The attribution of anonymous translations to a specific author when we don't have any certain translation samples from him is a much more difficult task. We obtained very good results when we applied feature selection and resampled our dataset in order to have an equal number of texts from both the author/translator and the other authors.

Moreover, our results confirm previous research findings (Arun, Suresh, & Madhavan, 2009) indicating that the writeprint of the translator was significantly greater than that of the original author. Future research will be directed to evaluate the source language effect and the impact of specific pairs of author/translator in the accuracy of authorship attribution.

Acknowledgments

The author would like to thank Professor Lena Marmarinou-Politou (Department of Philology, National and Kapodistrian University of Athens) for bringing him the idea of investigating the authorship of Alexandros Papadiamantis, one of the greatest Modern Greek writers in anonymous translations of the 19th century. The present study is part of the pilot investigation whether translation authorship is feasible and to what extent. Many thanks also to the author/translator of this study, Dr. Titika Dimitroulia, Professor of the Theory of Translation in the Department of French Language and Literature of the Aristotelian University of Thessaloniki. Without her unconditional offer of all her original texts and translations the present study could not have been completed.

References

- Allison, B., Guthrie, L.** (2008). Authorship attribution of e-mail: Comparing classifiers over a new corpus for evaluation. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA): 2179-2183.
- Arun, R., Suresh, V., Madhavan, V. C. E.** (2009). Stopword graphs and authorship attribution in text corpora *Proceedings of 2009 IEEE International Conference on Semantic Computing (ICSC 2009), 14-16 September 2009, Berkeley, CA. USA:192-196*.
- Banerjee, S., Pedersen, T.** (2003). The design, implementation, and use of the Ngram Statistic Package. In: *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico, February 2003: 370-381*.
- Baroni, M., Bernardini, S.** (2006). A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing 21*, 3, 259-274. DOI: <https://doi.org/10.1093/lc/fqi039>
- Burrows, J. F.** (1987). Word patterns and story shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing 2*, 61-70.
- Burrows, J. F.** (1989). 'A vision' as a revision. *Eighteenth Century Studies 22*, 551-565.
- Burrows, J. F.** (1992). *Computers and the study of literature Computers and Written Texts*. Oxford: Blackwell.
- Burrows, J. F., Craig, D. H.** (1994). Lyrical drama and the “turbid mountebanks”: Styles of dialogue in romantic and renaissance tragedy. *Computers and the Humanities 28*, 2, 63-86.
- Burrows, J. F., Hassal, A. J.** (1988). Anna Boleyn and the authenticity of Fielding's feminine narratives. *Eighteenth Century Studies 21*, 427-453.
- Burwick, F., McKusick, J. C.** (eds.) (2007). *Faustus: From the German of Goethe. Translated by Samuel Taylor Coleridge*. Oxford: Clarendon Press.
- Coyotl-Morales, R., Villaseñor-Pineda, L., Montes-y-Gómez, M., Rosso, P.** (2006). Authorship Attribution Using Word Sequences. In: Martínez-Trinidad, J., Carrasco Ochoa, J., Kittler, J. (eds.), *Progress in Pattern Recognition, Image Analysis and Applications: 844-853*. Springer Berlin / Heidelberg.
- Craig, H.** (2008). Hugh Craig reviews The Stylometric Analysis of Faustus, from the German of Goethe. *The Coleridge Bulletin 32*, 85-88.
- Diederich, J., Kindermann, J., Leopold, E., Paass, G.** (2003). Authorship Attribution with Support Vector Machines. *Applied Intelligence 19*, 1, 109-123.
- Escalante, H. J., Solorio, T., Montes-y-Gómez, M.** (2011). Local histograms of character N-grams for authorship attribution *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1: 288-298*. Portland, Oregon: Association for Computational Linguistics.
- Farrington, M., Farrington, J.** (1979). A computer-aided study of the prose style of Henry Fielding and its support for his translation of The military history of Charles XII. In: Alger, D. E., Knowles, F. E., Smith, J. (eds.), *Advances in Computer-Aided Literary and Linguistic Research. Proceedings of the Fifth International Symposium on Computers in Literary and Linguistic Research: 95-105*. Birmingham: John Goodman & Sons.
- Frawley, W.** (1984). *Translation: Literary, linguistic, and philosophical perspectives*. New York: University of Delaware Press.
- Gehrke, G. T.** (2008). *Authorship discovery in blogs using Bayesian classification with corrective scaling*. Masters, Naval Postgraduate School, Monterey, CA.

- Gellerstam, M.** (1986). Translationese in Swedish novels translated from English. In: Wollin, L., Lindquist, H. (eds.), *Translation Studies in Scandinavia: 88-95*. Lund: CWK Gleerup.
- Goldberg, D. E.** (1989). *Genetic algorithms in search, optimization, and machine learning*. Boston, MA: Addison-Wesley.
- Grieve, J. W.** (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing* 22, 3, 251-270.
- Grzybek, P.** (2014). The Emergence of Stylometry: Prolegomena to the History of Term and Concept. In: Kroó, K., Torop, P. (eds.), *Text within Text - Culture within Culture: 58-75*. Budapest, Tartu: L'Harmattan.
- Hedegaard, S., Simonsen, J. G.** (2011). Lost in translation: authorship attribution using frame semantics *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, 19-24 June 2011, Portland, Oregon, USA*, Vol. 2: 65-70. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Hoover, D.** (2003). Another perspective on vocabulary richness. *Computers and the Humanities* 37, 151-178.
- Houvardas, J., Stamatatos, E.** (2006). N-Gram Feature Selection for Authorship Identification. In: Euzenat, J., Domingue, J. (eds.), *Artificial Intelligence: Methodology, Systems, and Applications: 4183, 77-86*. Berlin/Heidelberg: Springer.
- Joachims, T.** (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.), *Proceedings of the 10th European Conference on Machine Learning, 21-24 April 1998, Dorint-Parkhotel, Chemnitz, Germany: 137-142*. Berlin: Springer.
- Juola, P.** (2007). Future trends in authorship attribution. In: Craiger, P., Sheno, S. (eds.), *Advances in Digital Forensics III: 242, 119-132*. Springer Boston.
- Juola, P.** (2008). Authorship attribution. *Foundations and Trends in Information Retrieval* 1, 3, 233-334. DOI: <https://doi.org/10.1561/1500000005>
- Kessler, B., Numberg, G., Schütze, H.** (1997). *Automatic detection of text genre*. Paper presented at the Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain: 32-38.
- Koppel, M., Ordan, N.** (2011). *Translationese and its dialects*. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11), Portland, Oregon, USA, June 19-24, 2011: 1318-1326.
- Koppel, M., Schler, J., Argamon, S.** (2011). Authorship attribution in the wild. *Language Resources and Evaluation* 45, 1, 83-94. DOI: <https://doi.org/10.1007/s10579-009-9111-2>
- Lewis, D. D., Yang, Y., Rose, T. G., Li, F.** (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *The Journal of Machine Learning Research* 5, 361-397.
- Luyckx, K., Daelemans, W.** (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing* 26, 1, 35-55. DOI: <https://doi.org/10.1093/lc/fqq013>
- Madigan, D., Genkin, A., Lewis, D., Argamon, S., Fradkin, D., Ye, L.** (2005). Author identification on the large scale. *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America: Theme: Clustering and Classification*. Washington University School of Medicine, St. Louis, Missouri: Classification Society of North America: 1-32.

- Mikros, G. K.** (2007). Stylometric experiments in Modern Greek: Investigating authorship in homogeneous newswire texts. In: Köhler, R., Altmann, G. Grzybek, P. (eds.), *Exact methods in the study of language and text: 445-456*. Berlin/New York: Mouton de Gruyter.
- Miranda, G. A., Calle, M. J.** (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing* 22, 1, 49-66.
- Mosteller, F., Wallace, D. L.** (1984). *Applied Bayesian and classical inference. The case of The Federalist Papers* (2nd ed.). New York: Springer-Verlag.
- Oakes, M. P.** (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
- Peng, F., Schuurmans, D., Wang, S.** (2004). Augmenting Naive Bayes Classifiers with Statistical Language Models. *Journal of Information Retrieval* 7, 3-4, 317-345. DOI: <https://doi.org/10.1023/B:INRT.0000011209.19643.e2>
- Platt, J. C.** (1999). Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C. J. C., Smola, A.J. (eds.), *Advances in kernel methods: 185-208*. Cambridge: MIT Press.
- Santini, M.** (2004). *State-of-the-Art on Automatic Genre Identification*. Information Technology Research Institute, University of Brighton.
- Sawyer, R.** (2006). The New Shakspeare Society, 1873-1894. *Borrowers and Lenders. The Journal of Shakespeare and Appropriation* 2, 2, 1-11.
- Sebastiani, F.** (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)* 34, 1, 1-47. DOI: <https://doi.org/10.1145/505282.505283>
- Stamatatos, E.** (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60, 3, 538-556. DOI: <https://doi.org/10.1002/asi.21001>
- Tweedie, F. J., Baayen, H. R.** (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32, 5, 323-352.
- Uhlig, S. H.** (2010). Review of Frederick Burwick and James C. McKusick (eds.). *Faustus: From the German of Goethe*. Translated by Samuel Taylor Coleridge. *The Review of English Studies* 61, 251, 645-648. DOI: <https://doi.org/10.1093/res/hgq052>
- Vapnik, V.** (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.
- Whissell, C.** (2011). Challenging an authorial attribution: Vocabulary and emotion in a translation of Goethe's *Faust* attributed to Samuel Taylor Coleridge. *Psychological Reports*, 108,2, 358-366. DOI: <https://doi.org/10.2466/28.pr0.108.2.358-366>
- Wu, G., Chang, E. Y.** (2003). Class-boundary alignment for imbalanced dataset learning *ICML 2003 Workshop on Learning from Imbalanced Data Sets (II)*. Washington, DC.
- Zheng, R., Li, J., Chen, H., Huang, Z.** (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology* 57, 3, 378-393. DOI: <https://doi.org/10.1002/asi.20316>

Some Aspects of a Sign Language Quantitative Analysis

Jiří Langer^{1,2}, Jan Andres³, Martina Benešová⁴, Dan Faltýnek⁵

Dedicated to the memory of Professor Peter Grzybek (1957 – 2019).

Abstract

National sign languages of the Deaf have been considered natural languages by linguists for more than fifty years. Judging by the available professional publications it does not appear, however, that this conclusion could be verified in the light of the foreseeable validity of the Menzerath-Altmann law (MAL), which is one of the key laws of quantitative linguistics. The MAL might therefore be used as a touchstone for detecting whether the communication system has the character of a full-fledged, natural language. It requires, among other things, identifying a related analogy of the sign language levels of the given “text” structures under consideration, segmented appropriately in terms of constructs and constituents. The desired positivity of the shape parameters in the MAL is calculated numerically and verified statistically by means of the coefficient of determination R^2 , jointly with the homoscedasticity and normality of random errors. Our concrete experiments concern the analysis of three speeches performed by three Deaf users of Czech Sign Language (český znakový jazyk = ČZJ) on given topics in ČZJ.

Keywords: Czech Sign Language; Natural Language; Quantitative Linguistics; Menzerath-Altmann Law; Sign Language Study

1. Introduction

The main aim of this chapter is to present current research outcomes from quantitative analysis of Czech Sign Language. A multidisciplinary research project entitled “The Theoretical Basis for Teaching Czech Sign Language Tested through Quantitative Linguistic Methods”, funded by the Czech Science Foundation, was carried out by researchers from three faculties of Palacký University in Olomouc (Czech Republic). It assumes that a deeper linguistic analysis of ČZJ (especially its hierarchical structure), obtained through quantitative linguistic methods, will contribute to the development of the teaching theory and practice of ČZJ. In terms of its concept and research methods, it is a seemingly unique research project which applies newly created quantitative linguistic methods to Czech Sign Language for Deaf people, which has (in contrast to spoken languages) a simultaneous and polysynthetic (see for example Servusová 2008, Macurová 2011, Okrouhlíková 2015) structure.

Our task is to apply and validate quantitative linguistic methods and analyse the relationships between language levels in sign language. A number of studies in this field have been already done, but from other perspectives (see, for example, Malaia 2017; Malaia et al. 2016; Uras & Verri 1995; Handouyahia et al. 1999; Borneman et al. 2018). However, our tools and approaches are completely different from those applied in the studies mentioned above (Andres et al. 2019; Andres et al. 2020).

¹ Faculty of Education, Palacký University Olomouc, Czech Republic, ORCID: 0000-0001-5804-5066

² Address correspondence to: Jiří Langer, Institute of Special Education Studies, Faculty of Education, Palacký University Olomouc, Žižkovo nám. 5, 771 40 Olomouc, Czech Republic, jiri.langer@upol.cz.

³ Faculty of Science, Palacký University Olomouc, Czech Republic, ORCID: 0000-0001-5405-3827

⁴ Faculty of Arts, Palacký University Olomouc, Czech Republic, ORCID: 0000-0002-5319-1870

⁵ Faculty of Arts, Palacký University Olomouc, Czech Republic, ORCID: 0000-0002-5771-1976

Our chapter is organized as follows. For those who are not familiar with the elements of sign language and quantitative linguistics studies, some basic historical and theoretical facts will be briefly introduced in the next two sections. We will then present the methodological tools (MAL and statistical criteria) which will be applied in our investigations. We will additionally describe and segment the “text” structure under considerations in quantitative linguistic terms. Finally, the main results will be presented and discussed.

2. Sign language

For the sake of completeness, we will present a short characteristic of sign language, which is the natural language of the Deaf (the capital D refers to the cultural, language minority of the Deaf as a group of natural users of sign language). In specialized literature, the term “sign language” collectively belongs to various non-vocal language systems used by Deaf communities of various nationalities. Sign languages differ from spoken languages by the mode of their existence in particular; unlike spoken languages, sign languages are of a visual-motor nature (Johnston & Schembri 2007).

Throughout human history views on the significance of sign language have differed considerably, and at the turn of the twentieth century the view even arose that a deaf individual could learn, more or less, on his or her own if they could read spoken speech from lips and if there was no other possibility of communication. This would be the case if sign language was forbidden and no natural signs were allowed. A frequent, and still enduring, argument of opponents of sign language was the statement that sign languages were not full-fledged natural language systems. American linguist William C. Stokoe published a paper in 1960, however, presenting the results of an extensive linguistic analysis of American Sign Language and, at the same time, proved that sign languages of the Deaf had all the necessary characteristics of natural languages and were, thus, full-fledged languages (Stokoe 1960).

Historically, the research methods for the investigation of sign languages were based on an analogy with natural language. Equivalentents for concepts used in the description of natural language (such as phonemes, words, sentences, etc.) were found for sign language. This analogy was then grounded further with a focus on grammar and generally the semiotic nature of sign language. The greatest difference is seen in the simultaneous presence of phonemes forming signs, which does not correspond to natural language speech, whether they be spoken, written or otherwise. In contrast, the varying individual signs of the phonemic analogy, in relation to other particular signs, correspond to allophonic variability and phonotactic rules which are represented in more or less every natural language. There is, however, also the possibility of considering sign simultaneity (simultaneous production of signs) in the language of the Deaf from the point of view of polysyntheticity. This kind of perspective requires, however, an initial grammar model based on the phonemic structure of the simultaneously used signs.

Schematically, the comparison of the hierarchical structure of spoken and sign languages can be roughly expressed in the Table 1.

Table 1
Language levels in the text/speech (constructs and constituents).

Spoken language	Sign language
sentence	sentence
clause	clause
word	sign
syllable	?
phoneme (letter)	? (phoneme)

Although in the case of the first four stages (the text – sentence – clause – sign), where spoken languages and sign languages are structurally analogical, it is impossible to identify syllables in sign languages due to their simultaneousness, and the determination of the number of phonemes used is also extremely complicated. In the following text, we have therefore attempted to design a way of calculating phonemes which are used while articulating individual signs.

The text, apart from the above-mentioned, will have to satisfy the criterion of homogeneity to the highest attainable degree (Grzybek 2015). The delimitation of sentences, clauses and signs will be solely performed by Deaf users of ČZJ.

Linguistic research reveals that Czech Sign Language, just as American or British Sign Language, is a full-fledged, natural language meeting all language system attributes, independent of Czech structures, and completely linguistically comparable with it. An opinion frequently mentioned by supporters of oral communication, that the use of sign language in the communication of the hearing-impaired leads to the inability to acquire the spoken language, has been refuted because *“knowledge of the sign language does not prevent the acquisition of the majority language and knowledge of structural differences between the majority (spoken) language and the respective sign language can contribute to effecting teaching of the majority language”* (Macurová 2001). Based on knowledge obtained from research into Czech Sign Language, essential differences on the phonological, morphological, and lexical, and syntactic levels between ČZJ and Czech language were recognized (Servusová 2008).

Czech Sign Language is therefore completely independent from the Czech language and has its own grammar and verbal (sign) vocabulary. Unlike spoken languages, sign language has three types of expressional means of communication: the verbally non-vocal component (individual signs — hand moves and positions, face and body moves); the non-verbally non-vocal component (gestures, mimics); and the non-verbally vocal component (spoken and oral components accompanying signs). Another type of classification divides expressional means into manual factors (signs and information expressed using hands) and non-manual factors (mimics, head moves, body moves, etc.).

Alongside the differences stated above, Czech Sign Language also has some specific characteristics which the Czech language does not feature and which participate, along with other grammatical characteristics, in the final form of the sign expression. They include, in particular:

- incorporation processes — in which the simultaneousness and existence of the sign language in a three-dimensional room are used and in which the originally isolated signs are merged and interpenetrated;
- existence of classifiers — substitute signs (with the nature of morphemes) which merge with others and represent a certain meaning;
- mimetic description (space visualization) — by means of which it is possible to describe the characteristics of the given subject and space context easily, quickly and in detail.

3. Quantitative linguistics and the universality of the considered laws

We will use the same standard terminology and methodology used in quantitative linguistics throughout the text (see for example Altmann 1980; Andres et al. 2011; Hřebíček 1997; Hřebíček 2007; Köhler & Rieger 1993). We will now attempt to identify certain key quantitative linguistic concepts and methods.

There have been many ways to study natural languages throughout history: qualitative as well as quantitative approaches. Based on our previous analyses, experiments and their

outcomes, we decided to employ methods of quantitative linguistics for our research, in particular testing ČZJ texts for the MAL and complexity of text manifestation.

The MAL enunciates heuristically that *the longer a language construct is, the shorter its constituents are* (see Altmann 1980), where the language construct is a unit on a higher language level and is composed of units on the immediately lower level, meaning its constituents. The construct length is usually measured in the number of its constituents: the constituent length in the number of its subconstituents. The law reflects the universal nature of language to behave in an economical manner. Every natural language is therefore expected to maintain this quality (for example, see Köhler 2008).

It is well-known that the effect of the MAL can even be detected in the various domains of linguistics amongst other things — for example, Zipf-Mandelbrot's law (Mandelbrot 2000) — but also in non-linear biology, sociology, psychology, economics (see for example the Pareto law, in Mandelbrot 2000) and, more recently, in architecture (Lorenz et al. 2017). Of course, in each of these disciplines, it must have an appropriate interpretation in its own terms.

The MAL was originally formulated as a language law. Its manifestations are, nevertheless, studied in a wide range of phenomena. This makes research into its manifestations, resembling Zipf's law (Hřebíček 2002), show that Zipf's law can be considered as a variation of the MAL in a certain sense. Menzerath (1928, 1954) formulated the law as a relationship between particular language units; it was the relation of word lengths and syllable lengths measured in sound time lengths. Altmann (1980) reformulated the law as a relationship between language constructs and constituents, as was already pointed out. Due to this reformulation of the law by Altmann, quantitative linguistic research has focused more on other language units, including very distinct units, for example those related to semantic text arrangement (see Hřebíček 2002; Ziegler & Altmann 1997). This change allowed for the monitoring of relations among language levels and analysing the complex nature of the text structure (see Andres 2009, 2010; and Andres et al. 2011). It also leads to a discussion about the choice of language units used in analysing the law (see Benešová et al. 2016). The MAL analysis on a wide scale of language levels also complies with the program of synergetic linguistics which studies the relationships of language subsystems (see Köhler 1986, 1990; Fenk & Fenk-Oczlon 1993). In a similar way as in the case of Zipf's law, attention shifted from a natural language text to different structures which may be represented as a text or which can demonstrate the interrelation of their constituents; from the point of view of Zipf's law, this related to the frequency of certain elements in the structure, while the case of the MAL comes from the point of view of the relationship between construct and constituent lengths (size) in a particular structure.

The MAL manifestations were also tested in genetic text (see Bolshoy 2003). Certain linguistic methods are utilized traditionally in the analyses of genetic texts, for example for taxonomic purposes (Damerau-Levenshtein distance, in Damerau 1964). This field also hosts research of linguistic laws, and Zipf's law usage for prediction of non-coding DNA function may serve as an example, such as junk DNA (see Havlin et al. 1995; Mantegna et al. 1995). In linguistic analyses of genetic texts, Zipf's law research anticipates the MAL research and has a long-term tradition (see Niyogi & Berwick 1995; Tsonis, Elsner & Panagiotis 1997). The universality of Zipf's law manifestations in text has been discussed for a long time — see for example the question of law manifestations in random texts in Ferrer-i-Cancho & Elvevåg (2010); a review of the history of the current state of Zipf's law research and its prospects, including research overlaps outside the natural language, as introduced in Piantadosi (2014).

Due to the expanding interest in the MAL manifestations outside natural language, even this law is very likely to be discussed. The MAL and the Zipf's law analyses still represent one of the most significant linguistic approaches in genetic text analysis. They were employed, for example, to describe the hierarchy of the genetic code and protein structure (see Matlach & Faltýnek 2016). Both laws have been tested on many structures different from natural language

texts, we give examples of analyses of genetic texts, animal communication (see references), and Piantadosi (2014) shows more. Andres (2014) explains that both laws are universal, which are formulated isomorphically in many areas — including the Pareto Principle in economics, and the Hurst exponent (see Andres, Langer & Matlach 2020; Ferrer-i-Cancho et al. 2013; Gustison et al. 2016; Li 2012; Nikolaou 2014; Shahzad, Mittenthal & Caetano-Anollés 2015; etc.).

The MAL was tested in texts by patients with speech disabilities, for example with patients suffering from Broca aphasia (see Jašíčková et al. 2014). Zipf’s law manifestations made up a platform to describe texts by patients suffering from schizophrenia, as supported by Ferrer-i-Cancho.

The above-mentioned research of the MAL is often performed assuming that the manifestations of the law are universals of natural language or a structure in general. Related to animal language or to speech disabilities, it is understood as a means of language detection. Testing the MAL in sign language texts therefore appears to be a logical subsequent step, both from the point of view of understanding sign language as natural (which cannot be argued with) and, first and foremost, from the point of view of its structure description.

4. Methodology: application of the Menzerath-Altmann law

As a main tool, we will employ the truncated formula of the MAL, namely

$$y = Ax^{-b}, \text{ resp. (for } b \neq 0), \frac{1}{b} = \frac{\log x}{\log \frac{A}{y}}, \quad (1)$$

where A, b are real parameters.

The parameters A, b in the MAL can be calculated (from the lengths of constructs and constituents) via the standard regression (statistical) method. The formula of the MAL is namely transformed in a logarithmic way into a linear algebraic equation with the coefficients $a := \ln A$ and b , by means of numerical approximations based on the (minimization) least squares technique. For the accuracy of these approximations and more details, see for example Andres et al. (2014).

In the particular case, when $A = y(1) = y_1$, we arrive at the formula

$$y = y_1 x^{-b}, \text{ resp. (for } b \neq 0), \frac{1}{b} = \frac{\log x}{\log \frac{y_1}{y}}, \quad (2)$$

with only one free parameter b , usually called a *shape parameter*.

Formula (1), and its particular case Formula (2), express the relationship between the length x of a construct and the length y of its constituents. On both levels, the lengths can be calculated either in the nominal number of constituents, respectively in time duration in seconds. It is even allowed that both possible ways (called “motifs”) of calculations of lengths (in nominal numbers, that is, in seconds) can be combined (see Köhler 2015).

Heuristically, as has already been pointed out, it says that *the longer a language construct, the shorter its components (constituents) are*. Mathematically, it means that the parameters A, b , characterizing the given structure under consideration, should be positive, that is, $A > 0, b > 0$. Geometrically, the related graphs of the respective functions, expressing the dependence between construct and constituents, should show decreasing. This formula was applied in the frame of quantitative linguistics for the first time by Altmann (1980).

Our task will therefore be to verify the situations in which the calculated shape coefficient b is positive. We will then try to verify the obtained results statistically, using the following standard criteria:

- The *coefficient of determination* R^2 , the values of which belong to the interval $[0,1]$, in soft sciences should be at least around 0.5 or higher; naturally the higher the coefficient is, the better (see Andres et al. 2014);
- *Homoscedasticity*: if the dispersion of random errors, related to a given process of approximation, is constant, then we speak about *homoscedasticity*; otherwise, we speak about *heteroscedasticity*. Homoscedasticity can be tested by means of the White test.
- *Normality*: if a data set is well modelled by a normal (i.e. Gaussian) distribution, then we speak about *normality*. Thus, normality can also be either confirmed or rejected. It can be tested by the Shapiro-Wilk test, or by the Kolmogorov-Smirnov test. If the normality of the random errors is not rejected, then a confidence interval can be determined in a standard way by Wald statistics.

For more details about homoscedasticity/heteroscedasticity and normality, including the above-mentioned tests, see for example Andres et al. (2014) and the references therein.

5. Research material and segmentation tools

The collection of research data of ČZJ for the following quantitative analysis was performed by the elicitation method, as is standard for ČZJ analyses (see Macurová & Bímová 2001; Macurová 2008). Research data were collected solely from Deaf users of ČZJ (the prelingually deaf for whom ČZJ is a natural language) in the shape of natural monologues on given topics (my hobbies, recent holiday, my family, my job, etc.). The recordings were done in places which were common for signers (for example in their homes, workplaces). Due to the careful selection of suitable respondents, there were three native Deaf ČZJ users among the team members who are highly respected in the community of the Czech Deaf and who supervised the grammar and lexical levels of the recorded speeches. Thirty-one speeches (approximately ten minutes long) were recorded in ČZJ from various respondents (older than fifteen years of age with different types and levels of education, different ages, social status, and both gender groups).

Recorded speeches were documented by means of digital video cameras with high resolution and were further analysed by Deaf native ČZJ users.

For the annotation and segmentation of research data the application ELAN language annotator (ELAN 2020), which enables the synchronization of the coding transcription record with the dynamic video material, was used. In this application, it is possible to define the lengths of individual constituents using the timeline, including the addition of possible annotation notes (see Figure 1). Finally, four randomly selected records were segmented. As mentioned above, Deaf native ČZJ users identified the beginnings and ends of all sentences, clauses, and signs on the timeline.

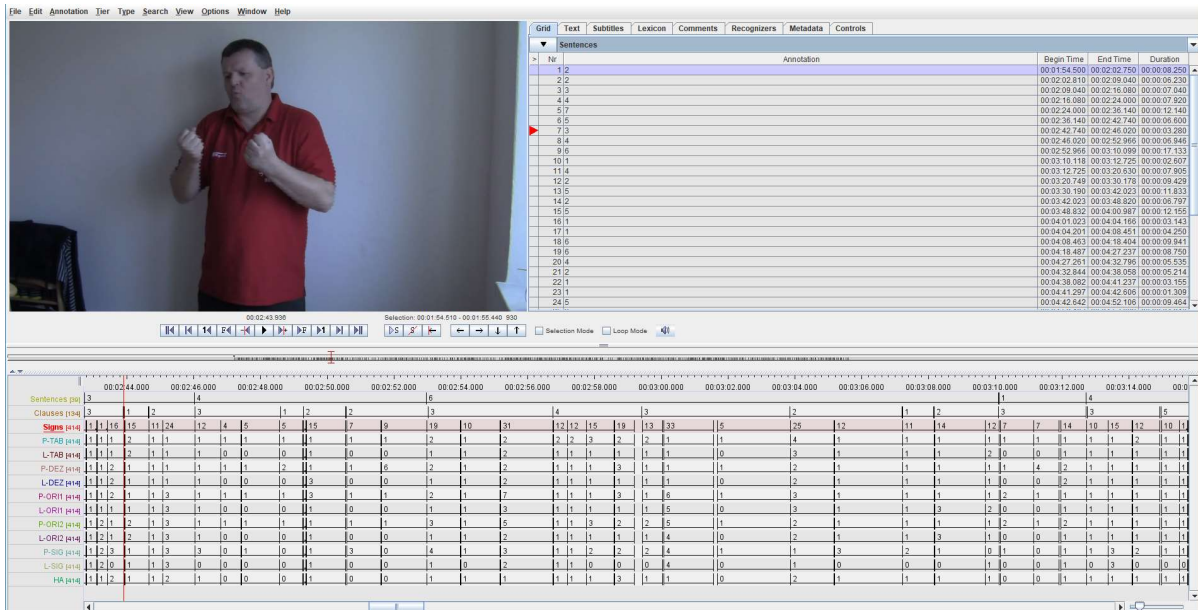


Figure 1 Software tool for the annotation and segmentation of fixed research data (ELAN Language Annotator).

There are two possible methods of counting the number of sign constituents. A key step for segmentation is the definition of constituents and constructs, specifically, in which values they will be further processed. As already pointed out in Section 4, one way is to count the number of individual units in a given language level (the nominal number of constituents constituting the parent construct), which is used in quantitative linguistic studies to verify the MAL. Another way is to calculate the time length of individual constituents and constructs made of them. The third option is a combination of both. We will therefore present the interim results of the analyses performed on the basis of the above-mentioned segmentation methods.

6. Results

To inspect the quality of the relationship of language units on the aforementioned two levels, we decided to test it by means of monitoring the manifestation of the MAL. First, we mined the data from the quantified ČZJ texts. We specifically identified the relationship between sentences (constructs) on *Level 1* and clauses (constituents). Furthermore, we also identified the relationship between clauses (this time, constructs) and signs (constituents) on *Level 2*.

6.1. Level 1 – sentence in constituents vs average length of clauses in subconstituents

As we will see, the following three cases will be considered.

Table 2
Overview of units of constructs and constituents at *Level 1*.

Case	Construct (in constituents)	Constituent (in subconstituents – average)
1	Sentence (in number of clauses)	Clause (in signs)
2	Sentence (in seconds)	Clause (in seconds)
3	Sentence (in number of clauses)	Clause (in seconds)

Level 1, Case 1

In Table 3, we present the quantities of constructs and constituents which were mined from our ČZJ sample. These data are visualized in the form of the data points in Figure 2. The data points are interlaid with a curve which represents the manifestation of the tendency held by the data points in terms of the MAL. If the curve is a downward sloping hyperbolic, the MAL assumptions are fulfilled (this is also verified by the positivity of the calculated parameter b in Table 4. The quality of such a MAL relationship has to therefore be verified by means of statistics (see Table 4); the higher the coefficient of determination R^2 , the better the curve fits the tendency of data points.

Table 3

The relationship between the lengths of sentences x (measured in the number of clauses), occurring with the frequency of z , and the average lengths of clauses y (measured in the number of signs).

x	y	z
1	3.33	9
2	4.29	7
3	3.23	3
4	2.83	6
5	2.97	7
6	2.97	6
7	2.29	1

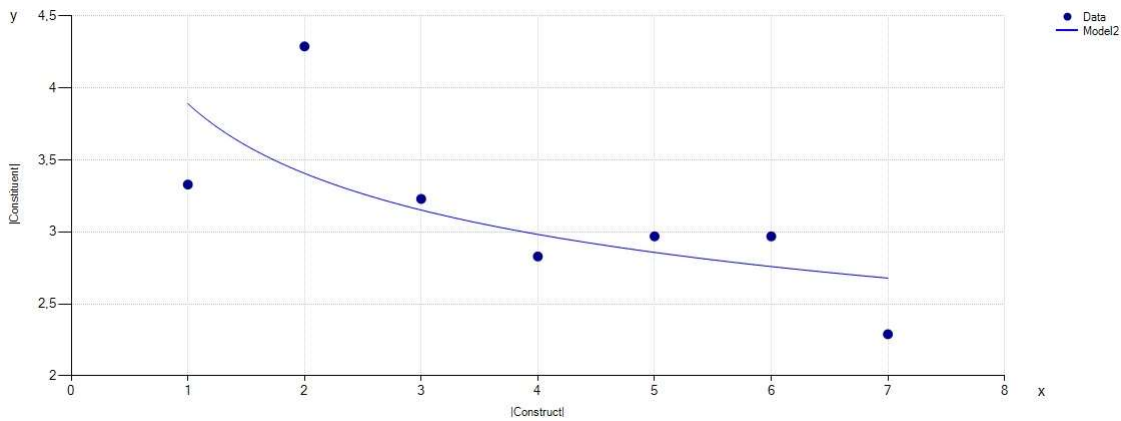


Figure 2 The graph illustrating the MAL curve calculated from the input data presented in Table 3.

Table 4

Statistical evaluation of the experiment; A , b are the parameters of the MAL, R^2 is the coefficient of determination, homoscedasticity, and normality (explained in Section 4).

A	b	R²	Homo.	Normal.
3.8923	0.1918	0.4797	not rejected	not rejected

In this case, the MAL tendency of data points has been proved. We have to add, however, that the goodness-of-fit of the curve is slightly below average (47.97 %), but still mostly acceptable. Moreover, the number of data points is not particularly convincing. This cannot, however, be improved even by extending the sample, for example, because the number of data points reflects the length of constructs.

Level 1, Case 2

In this case, we chose to measure units in the relationship by time units. The data are presented in Table 5 and their visualization in Figure 3.

Table 5

The relationship between the lengths of sentences x (measured in seconds), occurring with the frequency of z , and the average lengths of clauses y (measured in seconds).

x	y	z	x	y	z
1.369	1.369	1	7.92	1.98	1
2.06	2.048	1	7.947	1.99475	1
2.524	2.524	1	8.25	4.14	1
2.631	2.631	1	8.433	1.692	1
3.19	3.19	1	8.443	2.0825	1
3.214	3.214	1	8.511	4.267	1
3.28	1.08	1	8.75	1.428667	1
3.786	3.786	1	9.429	4.667	1
4.25	4.215	1	9.464	1.888	1
4.643	4.547	1	9.941	1.661	1
4.916	1.646667	1	11.833	2.348	1
5.214	2.589	1	12.14	1.725714	1
5.307	2.63	1	12.155	2.42	1
5.535	1.35725	1	12.477	2.50	1
6.23	3.095	1	13.147	2.181833	1
6.60	1.318	1	14.627	2.419	1
6.797	3.387	1	16.633	3.318	1
6.946	1.74	1	17.133	2.83	1
7.04	2.34	1	17.935	2.984833	1
7.905	1.95525	1			

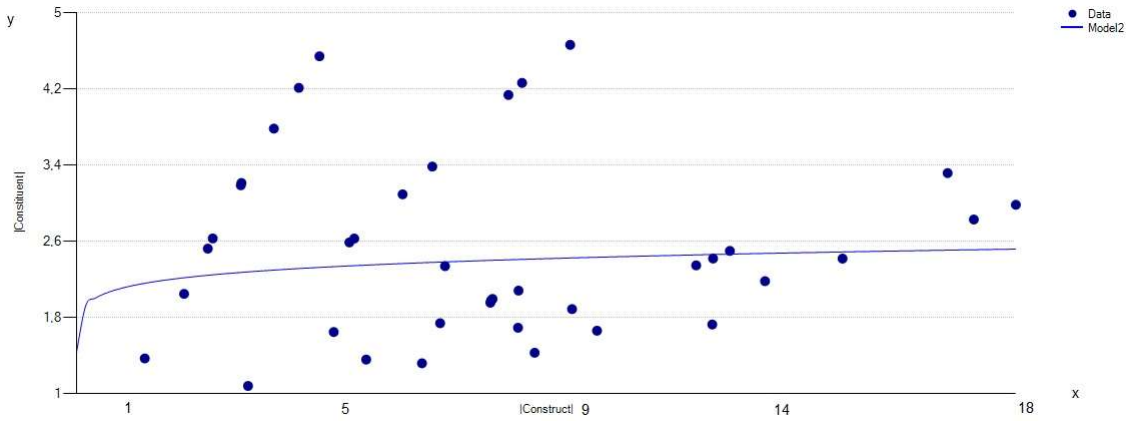


Figure 3 Graph illustrating the MAL curve calculated from the input data presented in Table 5.

Table 6

Statistical evaluation of the experiment; A , b are the parameters of MAL, R^2 is the coefficient of determination, homoscedasticity, and normality (explained in Section 4).

A	b	R²	Homo.	Normal.
1.4157	-0.0588	0.0094	not rejected	not rejected

The obtained results are, unfortunately, not satisfying in terms of the MAL qualities. The MAL represents an inversely proportional relationship to the expected one, that is, its curve to be hyperbolically decreasing, which is, at first sight not the case of the curve in Figure 3. The relationship of sentences in seconds and clauses in seconds does not therefore have the MAL properties as expected from two neighbouring language levels. Finally, we consequently adjusted our original assumptions and tested the relationship between sentences measured in clauses and clauses in seconds for MAL manifestations. The related results are presented in Table 7 and visualized in Figure 4.

Level 1, Case 3

Table 7

The relationship between the lengths of sentences x (measured in the number of clauses), occurring with the frequency of z , and the average lengths of clauses y (measured in seconds).

x	y	z
1	3.0582	9
2	3.5386	7
3	1.6889	3
4	1.8508	6
5	2.2115	7
6	2.2515	6
7	1.7257	1

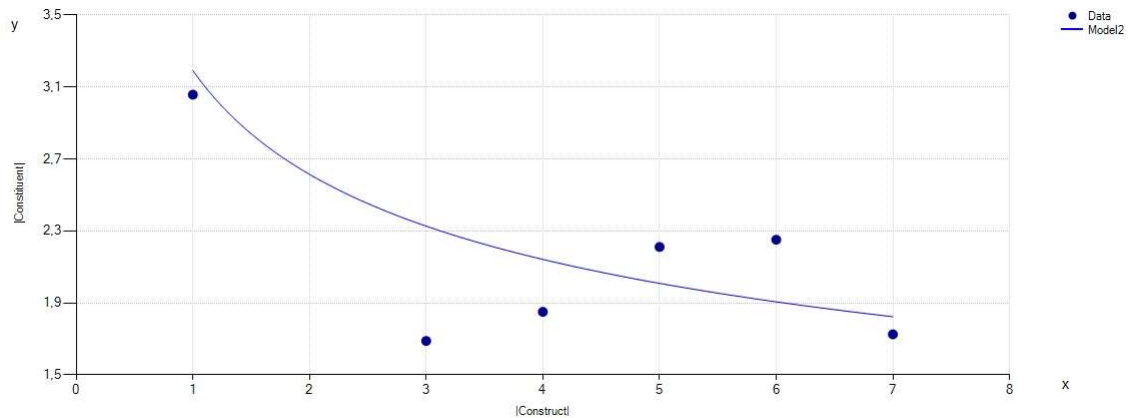


Figure 4 Graph illustrating the MAL curve calculated from the input data presented in Table 7.

The tendency of the curve in Figure 4 clearly proves the MAL. We need to know, however, “how good and how close” the relationship is. Therefore, we also checked it from the point of view of statistics.

Table 8

Statistical evaluation of the experiment; A , b are the parameters of the MAL, R^2 is the coefficient of determination, homoscedasticity, and normality (explained in Section 4).

A	b	R^2	Homo.	Normal.
3.1922	0.2877	0.4750	not rejected	not rejected

There is a short statistical assessment of the experiment in Table 8. It is apparent that the strength of the experiment is slightly below average again (47.50 %) and the decreasing tendency is proven by the positivity of parameter b .

6.2. Level 2 – clause in constituents versus average length of signs in subconstituents

The same procedure was applied to the relationship between clauses and signs.

Table 9

Overview of units of constructs and constituents at *Level 2*.

Case	Construct (in constituents)	Constituent (in subconstituents – average)
1	Clause (in number of signs)	Signs (in number of phonemes)
2	Clause (in seconds)	Signs (in seconds)
3	Clause (in number of signs)	Signs (in seconds)

Level 2, Case 1

Table 10

The relationship between the lengths of clauses x (measured in the number of signs), occurring with the frequency of z , and the average lengths of signs y (measured in the number of signs).

x	y	z
1	11.32	19
2	12.39	42
3	12	33
4	11.8	15
5	11.4	14
6	10.6	4
7	13.25	2
8	15	3
10	10.65	2

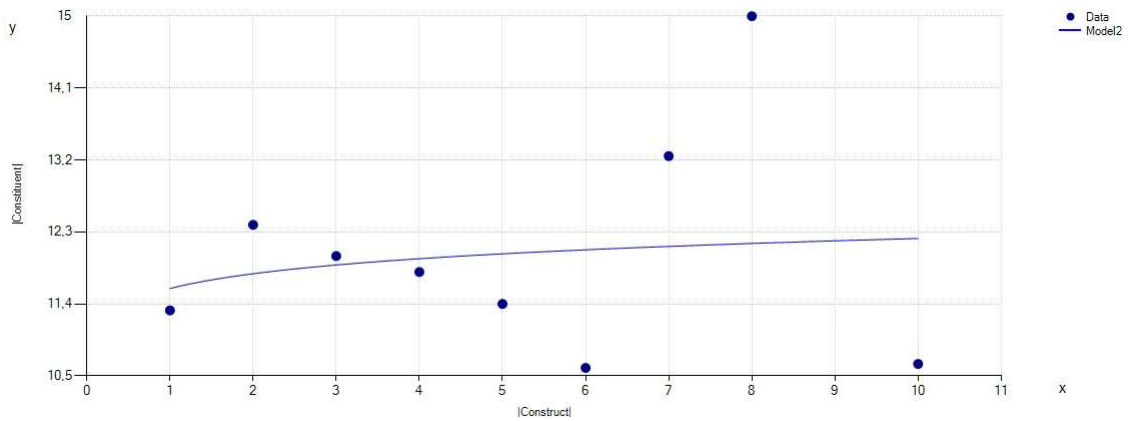


Figure 5 Graph illustrating the MAL curve calculated from the input data presented in Table 10.

Table 11

Statistical evaluation of the experiment; A , b are the parameters of the MAL, R^2 is the coefficient of determination, homoscedasticity, and normality (explained in Section 4).

A	b	R²	Homo.	Normal.
11.5932	-0.0228	0.0233	not rejected	not rejected

It is apparent at first sight from Figure 5 that the curve is upward sloping. Parameter b is therefore not positive which implies that the data points do not satisfy the MAL.

Level 2, Case 2

Table 12

The relationship between the lengths of clauses x (measured in seconds), occurring with the frequency of z , and the average lengths of signs y (measured in seconds).

x	y	z	x	y	z	x	y	z
0.460	0.46	1	1.655	0.798	1	2.648	0.59825	1
0.464	0.244	1	1.662	0.841	1	2.676	0.432333	1
0.547	0.194667	1	1.682	0.538	1	2.679	0.658	1
0.62	0.62	1	1.690	0.8515	1	2.680	0.536	1
0.679	0.679	1	1.728	0.572	1	2.726	0.5406	1
0.71	0.71	1	1.770	0.352	1	2.739	0.5328	1
0.702	0.702	1	1.774	0.863	1	2.750	1.345	1
0.717	0.635	1	1.850	0.61	1	2.761	0.920333	1
0.720	0.72	1	1.865	0.46275	1	2.798	0.908333	1
0.740	0.73	1	1.898	0.47725	1	2.834	0.912667	1
0.750	0.74	1	1.900	0.4775	1	2.840	0.466667	1
0.830	0.796	1	1.907	0.9645	1	3.040	1	1
0.917	0.428	1	1.944	1.944	1	3.150	0.7825	1
0.94	0.94	1	1.932	0.625	1	3.190	1.5	1
0.977	0.5235	1	1.941	0.639	1	3.191	1.565	1
1.000	0.494	1	1.946	0.9525	1	3.214	0.7855	1
1.011	0.989	1	2.012	0.48275	1	3.216	0.78675	1
1.120	0.5655	1	2.013	1.0135	1	3.228	1.076	1
1.130	0.746667	1	2.019	0.976	1	3.290	0.8075	1
1.149	1.202	1	2.023	0.983	1	3.331	0.668	1
1.150	0.45	1	2.048	0.666667	1	3.364	1.25	1
1.158	0.6025	1	2.068	0.685667	1	3.405	1.047333	1
1.170	0.58	1	2.083	1.042	1	3.420	0.662	1
1.250	0.6545	1	2.103	0.704333	1	3.590	0.596667	1
1.260	0.63	1	2.136	1.0605	1	3.595	0.7094	1
1.274	0.655	1	2.179	0.722333	1	3.599	0.7252	1
1.318	1.285	1	2.189	0.5305	1	3.730	1.242333	1
1.341	0.867333	1	2.190	1.089	1	3.786	1.869	1
1.345	0.667	1	2.214	0.629143	1	3.796	0.75	1
1.369	0.444667	1	2.226	0.716	1	4.155	1.00275	1
1.370	0.685	1	2.239	0.4364	1	4.215	1.401	1
1.405	0.468	1	2.267	0.742	1	4.250	0.8432	1
1.417	0.6965	1	2.351	0.765667	1	4.284	0.845	1
1.428	0.685	1	2.386	0.826	1	4.419	1.425	1
1.430	0.715	1	2.387	0.591	1	4.547	0.632429	1
1.464	1.464	1	2.410	0.39	1	4.690	0.67	1
1.488	0.7145	1	2.416	0.781667	1	5.119	1.23825	1
1.500	0.7505	1	2.440	0.684857	1	5.306	0.5218	1
1.512	0.75	1	2.452	0.821667	1	5.534	0.674625	1
1.520	1.51	1	2.524	0.857	1	5.750	0.5717	1
1.540	0.745	1	2.546	1.267	1	6.284	0.7785	1
1.607	0.524	1	2.570	0.836667	1	6.536	0.8095	1
1.648	0.79	1	2.631	0.837	1			

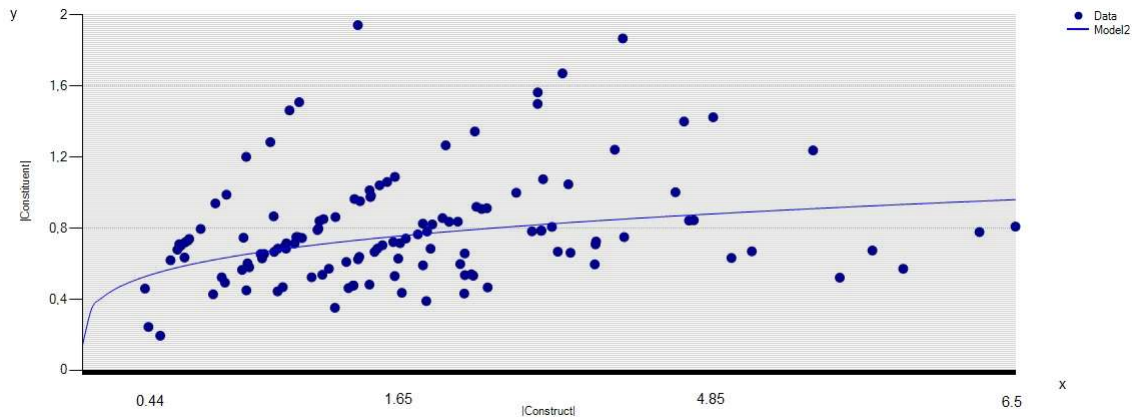


Figure 6 Graph illustrating the MAL curve calculated from the input data presented in Table 12.

Table 13

Statistical evaluation of the experiment; A , b are the parameters of the MAL, R^2 is the coefficient of determination, homoscedasticity, and normality (explained in Section 4).

A	b	R^2	Homo.	Normal.
0.1386	-0.2198	0.1143	not rejected	not rejected

The data points in this case behave in an even more random way than in the previous case. The goodness-of-fit cannot therefore be all that high. In addition, the MAL curve is upward sloping. Thus, in this case, the MAL is not satisfied.

Level 2, Case 3

Table 14

The relationship between the lengths of clauses x (measured in the number of signs), occurring with the frequency of z , and the average lengths of signs y (measured in seconds).

x	y	z
1	0.9767	19
2	0.8381	42
3	0.7680	33
4	0.6823	15
5	0.6376	14
6	0.4714	4
7	0.6512	2
8	0.7542	3
10	0.5468	2

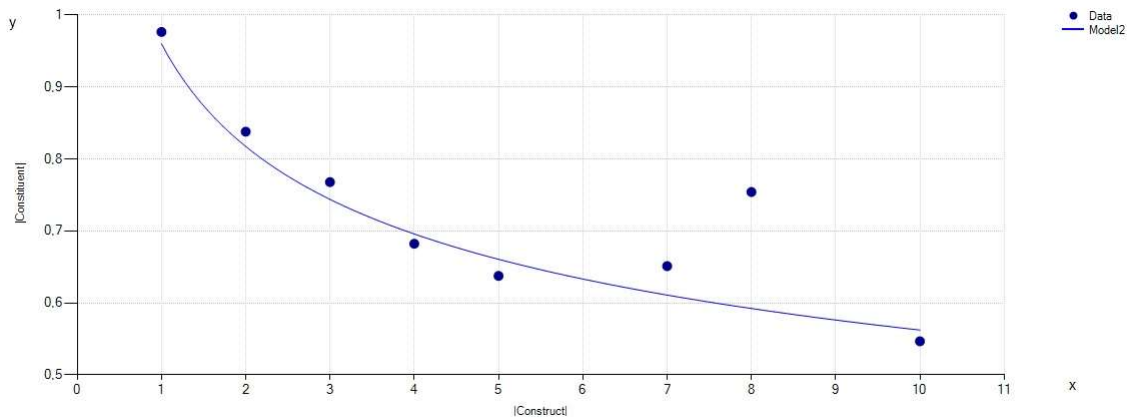


Figure 7 Graph illustrating the MAL curve calculated from the input data presented in Table 14.

The graph in Figure 7 visualizes how the tendency of the relationship is decreasing. In order to see how good the relationship is, we use statistics once again, see Table 15. The strength of the relationship is above average, namely 60.21 %.

Table 15

Statistical evaluation of the experiment; A , b are the parameters of the MAL, R^2 is the coefficient of determination, homoscedasticity and normality (explained in Section 4).

A	b	R^2	Homo.	Normal.
0.9604	0.2324	0.6021	not rejected	not rejected

7. Discussion and conclusions

From the calculations and graphs shown in Section 5, it is apparent that we have demonstrated the validity of the MAL only at the level of sentence – clause (level 1) and the level of clause – sign (level 2), and only in certain cases according to the used segmentation method:

- Level 1, Case 1 – lengths of sentences (measured in the number of clauses) and the average lengths of clauses (measured in the number of signs).
- Level 1, Case 3 – lengths of sentences (measured in the number of clauses) and the average lengths of clauses (measured in seconds).
- Level 2, Case 3 – lengths of clauses (measured in the number of signs) and the average lengths of signs (measured in seconds).

For all the cases we also verified the results statistically; in two cases with MAL confirmation (Level 1, Case 1 and Level 1, Case 3), the calculated value of R^2 (coefficient of determination) is at the edge, but the other statistical criteria (homoscedasticity and normality) confirmed the validity. In the case of Level 2, Case 3, all three statistical criteria used were very appropriate.

The calculation of the length of constructs, in the average number of constituents and the length of constituents in seconds, seems to be the most promising. For the time being, however, it is only one (though demanding) experiment. The experiments will continue in the future, for example, with an increase in the number of speeches analysed. **The fact that the results in some cases do not match the MAL assumptions can be caused, on the other hand, by an inappropriate segmentation of constructs or the omission of another intermediate level.**

After all, since our experiments were the first attempt filed in this vein of research, the results obtained must be considered as only preliminary. Nevertheless, these preliminary indicators signal that – because of closer analogies at higher levels (described in Table 1) – the verification of the MAL could be expected there. On the other hand, perhaps mainly due to the simultaneity of signs (in contrast to a linearity of spoken languages), their one-dimensional “projections” into the length, and respectively the time, dimensions, might have been too restrictive.

In any case, it would be appropriate, for example, to identify at least one more level for the intended fractal analysis, as in the spoken languages (see for example Andres 2010; Köhler 1997; Andres et al. 2019). Optimally, either the supra-level (the analogy of semantic constructs, see for example Hřebíček 2007) or another sub-level (below the level of the basic lexical unit, that is, signs) could still be detected. The first attempts at a fractal analysis of sign language have already been made (see Andres et al. 2020).


8. Concluding remarks

Due to the simultaneous character of the manual signs (as mentioned above), one of the most challenging tasks is to develop a method of counting the length of signs (the number of constituents). One of the proposed methods is a simple sum of all phonemes which are produced over its course, according to these sign parameters:

- sign location in the articulation space (TAB)
- shape of the articulating hand/hands (DEZ)
- orientation of the palm/palms (ORI1)
- orientation of fingers (ORI2)
- movements of the hand/hands (SIG)
- mutual hand arrangement in two-handed signs (HA)

Table 16

The design of the analysis and notation of the total phoneme number in the one-handed sign (CHOOSE).

	Number of phonemes^{††}		
	Parameter	Right hand	Left hand
	TAB	1	0
	DEZ	2	0
	ORI1	1	0
	ORI2	1	0
	SIG	1	0
	HA	0	
	Total	6	

A number of studies performed on ASL indicate that certain hand shapes can affect the perception of sign complexity (Brentari 2011; Brentari et al. 2017). In the case of ČZJ, this has not yet been proven. The table presented above indicates that the total number of phonemes is significantly influenced by the number of hands (which are used to create a sign) and the

^{††} The number of phonemes in the above-mentioned example indicates the number of values which can be obtained by a certain parameter in a sign. The place of articulation of the sign CHOOSE is, for example, unchanged (TAB = 1) as well as the palm orientation (ORI1 = 1) and the orientation of the fingers (ORI2 = 1) while the movement of a hand is not repeated (SIG = 1). The original shape of the hands, however, is changed (DEZ = 2).

number of phonemes in the individual parameters of signs. The question remains, however, as to whether the two-handed sign (with a higher total sum of phonemes) is actually perceived by users of sign language as significantly more difficult to learn and remember than the one-handed sign with a logically lower number of phonemes.

We consequently conducted a study aimed at detecting the dependence between the number of sign phonemes, their belonging to defined motion matrices, and the assessment of their complexity by the users (Langer & Rypka 2017). The linear dependency of the complexity of the sign on the number of its phonemes (see Figure 8 below) makes obvious, although not entirely essential (the correlation reaches the value 0.4), the distribution of the values of the average assessment of the complexity of the signs. In addition, the non-standard placement of signs S08, S09, S24, S25, S26 and S28 is apparent, where they are assessed as relatively simple, despite a rather high number of phonemes. Due to the fact that they are two-handed symmetrical signs (2H symmetrical), the question emerges if there is an unwanted overrating of the total number of phonemes among these signs, since it is basically a mirror image and motion with both hands.

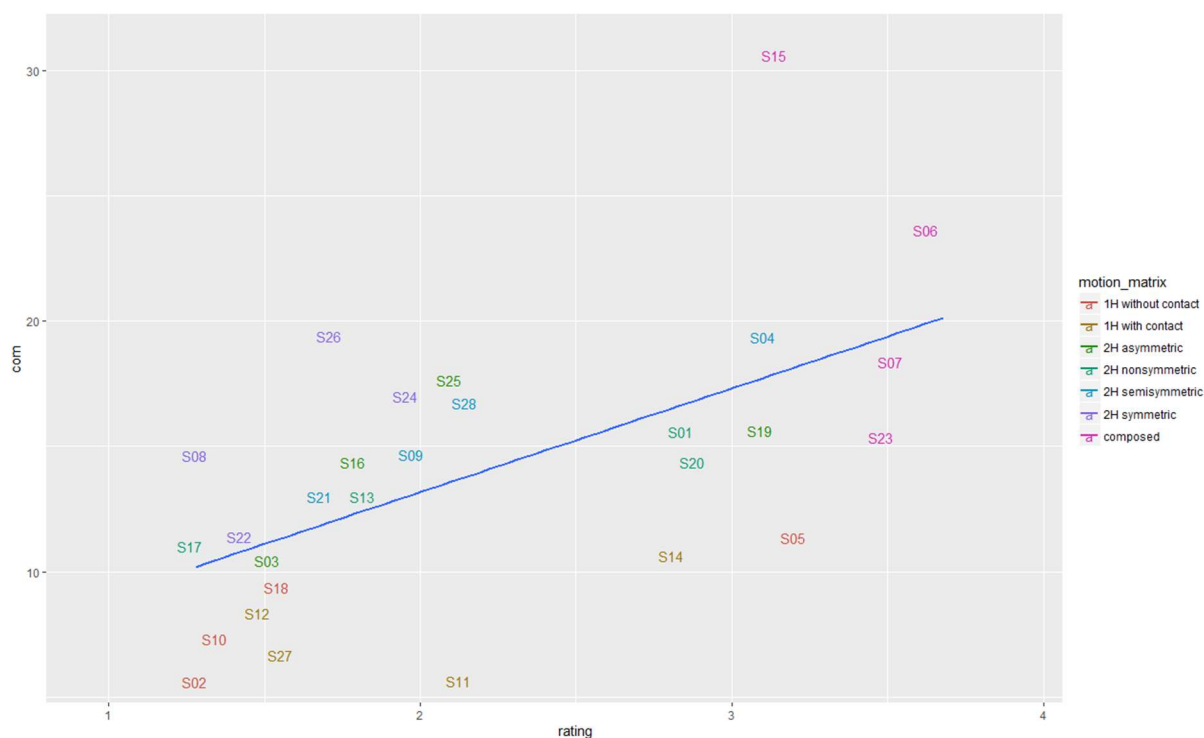


Figure 8 The dependence of the average value of the complexity of the sign on the number of phonemes.

The performed research confirmed the original theoretical assumption that the number of phonemes in ČZJ signs proportionally influences the subjective perception of the complexity of the presented signs. The number of phonemes was stated at the beginning of the research in a way that referred to the sum of phonemes occurring during the articulation of a sign within the individual parameters of a sign for both the dominant and the non-dominant hand. An important factor in perceiving the complexity of the signs may be specific hand shapes or motions, but our study does not currently address this phenomenon.

The results stated above were also essential for the execution of the quantitatively linguistic analysis of ČZJ, for which the number of phonemes of the individual signs is the lowest level of the observed hierarchical structure.

Our experiments thus far have unfortunately only shown that the relationship between the length of the constituents (counted in the number of phonemes in the signs) and the length of the constructs (counted in the number of signs) do not match the MAL. The question is what plays the role of constituents of signs. We tried to find an analogy of syllables / sounds in spoken languages. It turns out that the analogy is not there, or is, but MAL does not apply. We are aware that it is an open matter (eg including taking into account the non-manual component of signs). We also tried the linearization of simultaneously presented phonemes fo signs, but so far nothing corresponded to MAL. So far, we cannot draw a definitive conclusion. It is a challenge for our future research to detect whether or not there exists a definition of constituents of signs whose length can be effectively measured. Otherwise, it might be possible that MAL does not hold on the level of signs (in analogy to the level of syllables in spoken languages). Our next attempt will therefore be further experimental segmentation and subsequent identification of the number of constituents below the sign's level.

Acknowledgements

The study was supported by the Czech Science Foundation, project No. 17-18149S “The Theoretical Basis for Teaching Czech Sign Language Tested through Quantitative Linguistic Methods”.

References

- Altmann, G.** (1980). Prolegomena to Menzerath's Law. *Glottometrika* 2, 1-10.
- Andres, J.** (2009). On de Saussure's Principle of Linearity and Visualization of Language Structures. *Glottology* 2, 2, 1-14.
- Andres, J.** (2010). On a Conjecture about the Fractal Structure of Language. *Journal of Quantitative Linguistics* 17, 2, 101-122.
- Andres, J.** (2014). The Moran–Hutchinson Formula in Terms of Menzerath–Altmann's Law and Zipf–Mandelbrot's Law. In: Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.), *Empirical Approaches to Text and Language Analysis: 29-44*. Lüdenscheid: RAM-Verlag.
- Andres, J., Benešová, M., Chvosteková, M., Fišerová, E.** (2014). Optimization of parameters in Menzerath–Altmann law II. *Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium, Mathematica* 53, 1, 3-23.
- Andres, J., Benešová, M., Kubáček, L., Vrbková, J.** (2011). Methodological Note on the Fractal Analysis of Texts. *Journal of Quantitative Linguistics* 18, 4, 337-367.
- Andres, J., Benešová, M., Langer, J.** (2019). Towards a Fractal Analysis of the Sign Language. *Journal of Quantitative Linguistics* 26, 1-18.
<https://doi.org/10.1080/09296174.2019.1656149>.
- Andres, J., Langer, J., Matlach, V.** (2020). Fractal-based Analysis of Sign Language. *Communications in Nonlinear Science and Numerical Simulation* 84, 1-14.
<https://doi.org/10.1016/j.cnsns.2020.105214>
- Benešová, M., Faltýnek, D., Zámečník, L. H.** (2016). Menzerath–Altmann Law in Differently Segmented Text. In: Tuzzi, A., Benešová, M., Mačutek, J. (eds.), *Recent Contributions to Quantitative Linguistics: 27-40*. Berlin: De Gruyter Mouton.
- Bolshoy, A.** (2003). DNA Sequence Analysis Linguistic Tools: Contrast Vocabularies, Compositional Spectra and Linguistic Complexity. *Applied bioinformatics* 2, 103-112.
- Borneman, J. D., Malaia, E., Wilbur, R. B.** (2018). Motion Characterization Using Optical Flow and Fractal Complexity. *Journal of Electronic Imaging* 27, 5,
<https://doi.org/10.1117/1.JEI.27.5.051229>.
- Brentari, D.** (2011). Handshape in Sign Language Phonology. In: *Companion to Phonology: 195-222*. Oxford: Oxford University Press.
- Brentari, D., Coppola, M., Whan Cho, P., Senghas, A.** (2017). Handshape Complexity as a Precursor to Phonology: Variation, Emergence, and Acquisition. *Language Acquisition* 24, 4, 1-24.
- Damerau, F. J.** (1964). A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM* 7, 3, 171-176.
- ELAN** (Version 5.9) [Computer software]. (2020). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from <https://archive.mpi.nl/tla/elan>.
- Fenk, A., Fenk-Oczlon, G.** (1993). Menzerath's law and the constant flow of linguistic information. In: Köhler, R., Rieger, B. (eds.), *Contributions to quantitative linguistics. Proceedings of the first international conference on quantitative linguistics, Qualico 1001: 11-31*. Heidelberg: Springer.
- Ferrer-i-Cancho, R., Elvevåg, B.** (2010). Random Texts Do Not Exhibit the Real Zipf's Law–Like Rank Distribution. *PLoS ONE* 5. e9411.
- Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Bel-Enguix, G., Baixeries, J.** (2013). The Challenges of Statistical Patterns of Language: The Case of Menzerath's Law in Genomes. *Complexity* 18, 3, 11-17.
- Grzybek, P.** (2015). Word Length. In: Taylor, J. R. (ed.), *The Oxford Handbook of the Word*. Oxford: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199641604.013.37>

- Gustison, M. L., Semple, St., Ferrer-i-Cancho, R., Bergman, Th. J.** (2016). Gelada Vocal Sequences Follow Menzerath's Linguistic Law. *Proceedings of the National Academy of Sciences USA* 113, 19, 2750-2758.
- Handouyahia, M., Ziou, D., Wang, Sh.** (1999). Sign Language Recognition Using Moment-Based Size Functions. *Vision Interface '99, Trois-Rivières, Canada, 19–21 May: 210-216.*
- Havlin, Sh. et al.** (1995). Statistical and Linguistic Features of DNA Sequences. *Fractals* 3, 2, 269-284.
- Hřebíček, L.** (1997). *Lectures on Text Theory.* Prague: Czech Academy of Sciences.
- Hřebíček, L.** (2002). *Stories about Linguistic Experiments with Text.* Prague: Czech Academy of Sciences.
- Hřebíček, L.** (2007). *Text in Semantics.* Prague: Czech Academy of Sciences.
- Jašíčková, A., Benešová, M., Faltýnek, D.** (2013). An Application of the Menzerath-Altmann Law to a Sample Produced by an Aphasic Patient. *Czech and Slovak Linguistic Review* 3, 2, 4-27.
- Johnston, T., Schembri, A.** (2007). *Australian Sign Language (Auslan): An Introduction to Sign Language Linguistics.* Cambridge University Press.
- Köhler, R.** (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.
- Köhler, R.** (1990). Elemente der synergetischen Linguistik. *Glottometrika* 12, 179-188.
- Köhler, R.** (1997). Are There Fractal Structures in Language? Units of Measurement and Dimensions in Linguistics. *Journal of Quantitative Linguistics* 4, 1-3, 122-125.
- Köhler, R.** (2015). Linguistic Motifs. In: Mikros, G., Mačutek, J. (eds.), *Sequences in Language and Text: 89-108.* Berlin, Boston: De Gruyter Mouton.
- Köhler, R., Altmann, G., Piotrowski, R. G.** (eds.). (2008). *Quantitative Linguistik / Quantitative Linguistics: Ein internationales Handbuch / An International Handbook.* Berlin: de Gruyter.
- Köhler, R., Rieger, B. R.** (eds.) (1993). Contributions to quantitative linguistics. *Proceedings of the First International Conference on Quantitative Linguistics, Qualico 1991.* Dordrecht, Boston, London: Kluwer.
- Langer, J., Rypka, M.** (2017). Testing of Subjective Perception of Complexity of Signs of Czech Sign Language. *Journal of Exceptional People* 11, 2, 123-141.
- Li, W.** 2012. Menzerath's Law at the Gene-exon Level in the Human Genome. *Complexity* 17, 4, 49-53.
- Lorenz, W., Andres, J., Franck, G.** (2017). Fractal Aesthetics in Architecture. *Applied Mathematics & Information Sciences* 11, 4, 1-10.
- Macurová, A.** (2001). Poznáváme český znakový jazyk. (Úvodní poznámky) [Discovering Czech Sign Language. (Introductory remarks)]. *Speciální pedagogika* 11, 2, 69-75.
- Macurová, A.** (2008). *Dějiny výzkumu znakového jazyka u nás a v zahraničí [History of Sign Language Research in our Country and Abroad].* Česká komora tlumočnicků znakového jazyka.
- Macurová, A.** (2011). Směřování k typologii znakových jazyků [Towards typology of sign languages]. *Speciální pedagogika* 21, 1, 1-7.
- Macurová, A., Bímová, P.** (2001). Poznáváme český znakový jazyk II. (Slovesa a jejich typy) [Discovering Czech Sign Language II. (Verbs and their Types)]. *Speciální pedagogika* 11, 5, 285-296.
- Malaia, E.** (2017). Current and Future Methodologies for Quantitative Analysis of Information Transfer in Sign Language and Gesture Data. *Behavioral and Brain Sciences* 40, E63.
- Malaia, E., Borneman, J. D., Wilbur, R. B.** (2016). Assessment of Information Content in Visual Signal: Analysis of Optical Flow Fractal Complexity. *Visual Cognition* 24, 3, 246-251. <https://doi.org/10.1080/13506285.2016.1225142>.

- Mandelbrot, B. B.** (2000). *Les Objets Fractals: Forme, hasard et dimension*. Flammarion.
- Mantegna, R. N. et al.** (1995). Systematic Analysis of Coding and Noncoding DNA Sequences Using Methods of Statistical Linguistics. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics* 52, 2939-2950.
- Matlach, V., Faltýnek, D.** (2016). Báze nejsou písmena [The Bases Are Not the Letters]. *Studie z aplikované lingvistiky* 7, 1, 20-38.
- Menzerath, P.** (1928). Über einige phonetische probleme. *Actes du premier congrès international de linguists*.
- Menzerath, P.** (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.
- Nikolaou, Ch.** (2014). Menzerath-Altmann Law in Mammalian Exons Reflects the Dynamics of Gene Structure Evolution. *Computational Biology and Chemistry* 53, 134-143.
- Niyogi, P., Berwick, R.C.** (1995). A Note on Zipf's Law, Natural Languages, and Noncoding DNA Regions. *A. I. Memo 1530, No. 118*.
- Okrouhliková, L.** (2015). *Notace znakových jazyků [Notation of Sign Languages]*. Karolinum.
- Piantadosi, St. T.** (2014). Zipf's Law in Natural Language: A Critical Review and Future Directions. *Psychonomic Bulletin & Review* 21, 5, 1112-1130.
- Servusová, J.** (2008). *Kontrastivní lingvistika – český jazyk x český znakový jazyk [Contrastive Linguistics – Czech Language x Czech Sign Language]*. Praha: Česká komora tlumočnicků znakového jazyka.
- Shahzad, Kh., Mittenthal, J. E., Caetano-Anollés, G.** (2015). The Organization Of Domains in Proteins Obeys Menzerath-Altmann's Law of Language. *BMC Systems Biology* 9, 44, 1-13.
- Stewart, J.** (2014). A Quantitative Analysis of Sign Lengthening in American Sign Language Article. *Sign Language & Linguistics* 17, 1, 82-101.
- Stokoe, W. C.** (1960). *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf*. New York: University of Buffalo.
- Tsonis, A. A., Elsner, J. B., Panagiotis, T. A.** (1997). Is DNA a Language? *Journal of Theoretical Biology* 184, 25-29.
- Uras, C., Verri, A.** (1995). Sign Language Recognition: An Application of the Theory of Size Functions. *6th British Machine Vision Conference* 2, 711-720.
- Ziegler, A., Altmann, G.** (2007). Latent Connotative Text Structure. In: Mehler, A., Köhler, R. (eds.): *Aspects of Automatic Text Analysis: 211-229*. Berlin: Springer.

Adnominal Valency in Modern Russian

Sergey Andreev¹

Abstract

This article is devoted to the quantitative study of adnominal valency, understood as the number of attributes, or adnominals, modifying a given nominal (noun, personal pronoun, nominalized adjective, etc.). The data source of the study consists of 12 texts, written by four very popular authors of modern Russian mass literature, encompassing two female and two male writers of different professional backgrounds. The study focuses on the distribution of valencies taken separately as well as in specifically organized combinations – Köhlerian motifs. The Zipf–Alekseev function was found to fit both distributions very well. The results obtained demonstrate that all the authors follow similar implicit order in adnominal valency distribution irrespective of gender or professional background.

Keywords: adnominal valency; Zipf–Alekseev function; arithmetic mean; valency motifs; distribution.

1. Introduction

Attributive relations are a highly important feature of syntactic and semantic organization of text structure. Attributes, or adnominals, modifying nouns, personal pronouns or nominalized parts of speech play a leading role in the process of description in the fictional world of a novel. They have very tight links with the words they modify, but, at the same time, are not obligatory from the point of view of the verbocentric sentence arrangement, which means that adnominals in general are highly dependent on the individual style of an author.

This fact of supposedly arbitrary use of the number and type of adnominals by an author makes them a vivid feature of his or her individual style. The writers use adnominals consciously but choose them intuitively, without caring what form they choose or how they are distributed.

The question of how arbitrary the usage of adnominals is and if there is any order in it has been explored in a number of studies which showed certain regularities in the distribution of different types of adnominals (Altmann, 2015; Andreev, Popescu, Altmann, 2017a, 2017b; Místecký, 2019).

In this research we focus not on the adnominals themselves, but on nouns modified by adnominals. From this point of view nominals (nouns, personal pronouns, nominalized participles, adjectives, interjections and other parts of speech acquiring a nominal function) have the capacity to establish attributive syntactic and semantic links with words realizing description.

The number of such adnominals modifying a given nominal may be considered as its adnominal valency. The present study focuses on such adnominal valencies from the point of view of their distribution in fiction.

Just like with adnominals themselves, one of the main questions concerning the adnominal valency features consists in the question whether the choice of these adnominal valencies is completely optional or is governed by the underlying regularity. This study is devoted to the search of such possible regularities of using adnominal valencies in Russian mass fiction.

¹ Smolensk State University, smol.an@mail.ru.

2. Adnominal types and data source

The data source of the study includes abstracts from 12 works, written by four Russian writers (two male and two female; their names and the list of their works is given in the appendix). All these works belong to the genre of detective stories, and all the writers have been at the top of the list of the most popular authors in Russian mass literature over the past several decades. The texts chosen for the analysis are of the same length (2,000 words) and were taken from the beginning of the novels.

The authors may be grouped by gender (male and female authors) and their previous profession. According to the last criterion we also have two groups – former police officers (Marinina, Koretsky) and professional writers who worked in the sphere of other literary genres before they became known as detective fiction writers (Akunin and to some extent Dontsova).

The list of adnominals in Russian includes the following morphological types:

- A – adjective (a *red* rose);
- PT – participle (a *crying* boy);
- AC – adjectival construction (*highly valuable* resources);
- PTC – participial construction (a girl *loved by everyone*),
- RCR – subordinate clause (the book *which was missing*);
- DET – determiner (demonstrative, possessive, qualifying, indefinite, interrogative/relative pronouns – *this* book, *my* book, *other* books, *some* books, *which* books...);
- PREP – prepositional construction (a book *for children*);
- G – the genitive case (the roof *of the house*)²;
- AP – apposition (Nick, *my best friend*, ...),
- INF – infinitive (*wish to go*);
- ADV – adverb (a room *upstairs*).

These characteristics were used to annotate texts for further analysis.

As has been mentioned above, adnominal valencies (AV) are established according to the number of adnominals modifying nominals. The following examples illustrate different valency types. In the following sentences adnominal valences are given in brackets, zero shows that the noun is not modified by any adnominal, and the type of adnominal is stated in square brackets.

1. This is a letter (0).
2. This is an important [A] letter (1).
3. This is a letter (1) from John [PRR].
4. This is a letter (1) from John [PRR] (1), my [DET] best [A] friend [AP] (2)

In (1) the noun *letter* has no adnominals, in (2) *letter* is modified by one (prepositional) adnominal *from John*, so it is ascribed AV1. *John* is a noun, but since it functions only as an adnominal, it itself has no adnominal valency. In (3) *John* again functions as a prepositional adnominal of the word *letter*, but this time it is modified itself by an apposition (*friend*), which is why it in turn acquires adnominal valency which is 1. *Friend*, functioning as an apposition, at the same time is modified by two adnominals – the determiner *my* and the adjective *best*, so it has AV2.

This approach was used to establish adnominal valency structure of the texts under study. For the sake of illustration let us take 20 first sentences from the beginning of the novel *The Stolen Sleep* by A. Marinina. The adnominal structure of the sample in which different types of adnominals are found looks as follows (slashes mark the end of sentences):

N(1) N(1) G AP N(0) A N(2) PTC / DET N(1) N(0) A N(1) / N(0) DET N(1) / DET N(1) DET N(1) / A A N(3) A N(1) PREP N(1) G N(0) N(0) / N(0) / N(1) G N(1) A N(3) G G G/ N(0) / A N(1) DET A N(2) / A N(1) / DET A N(2) / N(1) G DET N(1) / A A N(2) DET PT N(2) / N(0) N(1) G A N(1) / N(0) DET N(1) / N(0) / DET (N1) / N(1) G / DET (N1) N(1) AP A N(1) DET

² It should be noted that genitive in Russian does not have any preposition.

A N(2) N(2) G DET A N(2) PREP PT N(3) A N(1) G A N(1) G / A N(1) N(1) N(0) RCR N(1) INF N(0) /

Omitting the adnominals in this sample we obtain the sequence of adnominal valencies:

1 1 0 2 / 1 0 1 / 0 1 / 1 1 / 3 1 1 0 0 / 0 / 1 1 3 / 0 / 1 2 / 1 / 2 / 1 1 / 2 2 / 0 1 1 / 0 1 / 0 / 1 / 1 / 1
1 0 2 0 2 3 1 0 / 1 1 0 1 0 /

Further on, omitting nominals with zero AV and the division into sentences, we get the following final sequence:

1 1 2 1 1 1 1 1 3 1 1 1 1 3 1 2 1 2 1 1 2 2 1 1 1 1 1 1 1 2 2 3 1 1 1 1.

This extract includes 28 nouns, modified by one adnominal (AV1), eight nouns with VA2 and three nouns with AV3. Following this kind of analysis, counts of adnominal valencies in all 12 texts were made. The results are presented in Table 1.

Table 1
Frequencies of noun adnominal valencies

Valency	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
1	163	187	247	174	152	178	150	161	212	191	178	178
2	20	53	41	43	47	49	29	33	39	56	69	60
3	2	9	3	4	10	7	5	8	8	3	14	11
4	0	4	1	1	2	2	0	2	6	2	3	2
5	0	0	1	1	1	1	2	1	0	0	1	0
6	0	0	0	0	1	0	1	0	0	0	0	1

As seen from the table, the maximum valency in our database is 6. The range of variability is rather large; the arithmetic means for the authors are presented in Table 2 and by the graph which follows (Figure 1).

Table 2
Arithmetic means

AV	Dontsova	Marinina	Koretsky	Akunin
1	199.0	168.0	182.3	174.33
2	38.0	46.3	61.7	33.67
3	4.7	7.0	9.3	7.00
4	1.7	1.7	2.3	2.67
5	0.3	1.0	0.3	1.00
6	0.0	0.3	0.3	0.33

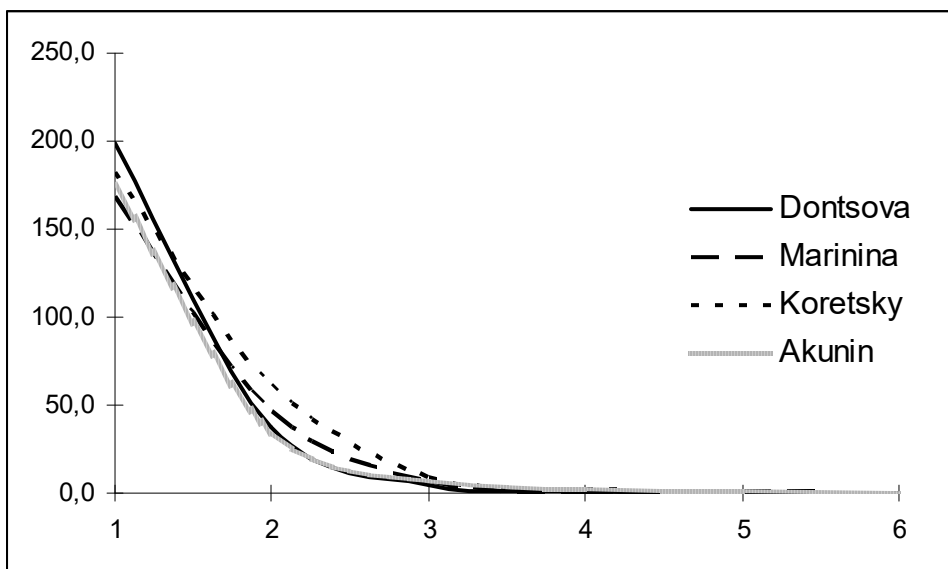


Figure 1 Means of AV in the works of the four authors

As can be seen from the graph, the distributions of AV in the works of different authors nearly coincide. From AV1 to AV3 a downward trend is observed, after AV2 the decline becoming more gradual, and after AV3 the adnominal value means are fairly static.

AV2, the point at which the decrease becomes more gradual, marks a certain opposition between AV1 on the one hand and the rest of the valencies on the other.

Valency 1 is the most frequent but at the same time the simplest type of attributive adnominal combination, admitting only two possible variants – prepositional and postpositional. AV2 has three variants (ADNOM, ADNOM)–NOUN; ADNOM–NOUN–ADNOM; NOUN–(ADNOM, ADNOM). Starting with AV3 the number of variants grows: (ADNOM, ADNOM)–NOUN–ADNOM; ADNOM–NOUN–(ADNOM, ADNOM); (ADNOM-ADNOM, ADNOM)–NOUN; NOUN–(ADNOM, ADNOM, ADNOM). With the growth of the number of syntactic links the description becomes more elaborate, which enhances perception. This makes it possible to compare the authors’ style by the strength of such opposition between AV1 and adnominal constructions of higher complexity.

To analyze such opposition, we used Busemann’s coefficient (Altmann, 2015), whose formula is:

$$(1) \quad C = \frac{A}{A + B},$$

where C is the coefficient of relationship of adnominal valencies, A is AV1 and B – all the rest types of AV (2-6).

The coefficient has the range of variation between 0 and 1. High values of this coefficient ($C > 0.5$) show that AD1 plays a more important role in the system of description; low values of the coefficient ($C < 0.5$) would indicate the predominance of adnominals with high valency.

To test the results the chi-square statistic was used (Andreev, Místeký, Altmann, 2018):

$$(2) \quad \chi^2 = \frac{(A - N)^2}{A + N}.$$

The coefficient is statistically significant with 1 degree of freedom and $p < 0.05$ if $\chi^2 > 3.4$.

Table 3 contains the values of Busemann's coefficient of AV1 against all other AVs and chi-square tests.

Table 3
AV1 vs. all other types of AV

Text	Busemann's coefficient	Chi-square
T1	0.88	111.74
T2	0.74	74.82
T3	0.84	147.35
T4	0.78	79.08
T5	0.71	55.40
T6	0.75	73.31
T7	0.80	81.79
T8	0.79	84.45
T9	0.80	119.24
T10	0.76	73.79
T11	0.67	48.10
T12	0.71	58.50

As seen from the table, all the values of C-coefficient are statistically significant. The AV1 style is predominant in all cases; the range over which this coefficient varies for different authors is between 0.67 and 0.88.

Using the data in the table one can find out the role of different types of AV in the style of the authors. Akunin has a clearly marked tendency to use more AVs with bigger scores than for example Dontsova, who on the contrary largely prefers AV1 to the other types of AV. Marinina, a retired police professional, has about the same proportion of low and big values of AV as Akunin, an author with a completely different professional background.

It should be underlined that the variation coefficient when used for every author individually showed a very low level of variation. In other words the proportion of low vs. bigger valencies for each author is rather stable, which points to a certain stability of the visualization of the fictional world by each author and raises the question of stability of an author's style in fictional description over time.

3. Male/female speech

Female authors' works at least in Russian literature are generally regarded as more emotionally vivid, containing more adnominals than those written by male authors. It should be underlined that this opinion is based mostly on impressions rather than counts.

Our material does not confirm this hypothesis. According to our counts of adnominal valencies there is no difference in the number of adnominals used by female and male authors – as a matter of fact, the latter actually demonstrate a slightly higher level of the usage of adnominals. Since the total of adnominal valencies is equal to the number of all adnominals one can count the latter using the AVs. Thus for T1 the number of adnominals, judging by the AVs, will be $N=163*AV1+20*AV2+2*AV3= 209$. In Table 4 the means of all adnominals used by female and male authors are presented.

Table 4
Mean values of the adnominals of the four authors

Text	Mean of adnominals
T1	209
T2	336
T3	347
T4	281
T5	295
T6	310
T7	293
T8	264
T9	338
T10	320
T11	375
T12	345

The following graph (Figure 2) indicates the adnominal volumes, ranked in descending order.

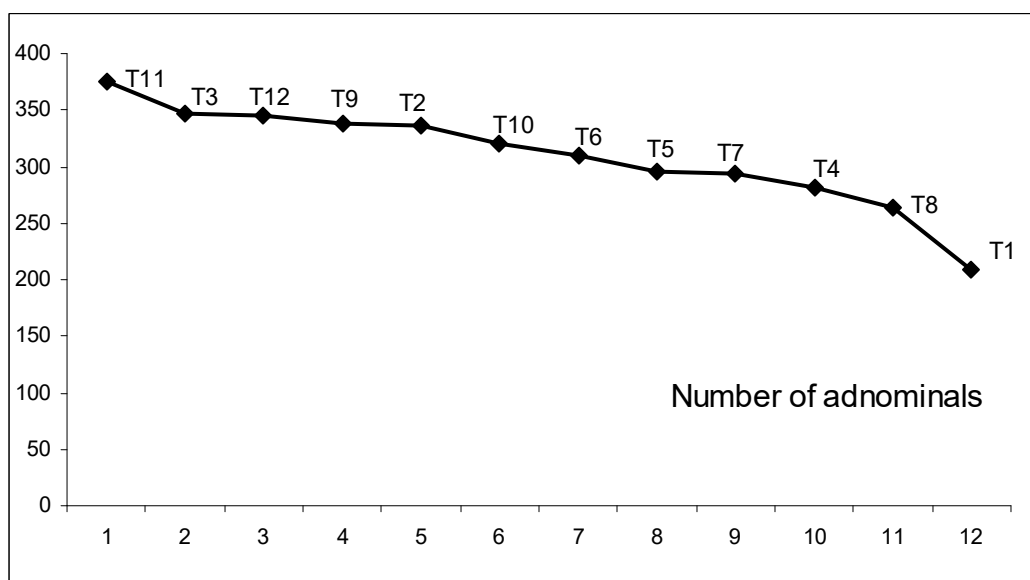


Figure 2 Number of adnominals used by the authors

As can be seen from the graph, four out of six works in the tail belong to the female authors. It is interesting to note that the former police officers are placed close to each other in the middle of the graph.

According to Mehl et al. (2007), who studied the extent of talkativeness of the speakers of different genders, one of the possible reasons for false impressions of talkativeness of females might be disproportions in the number of words in speech among women. In our case it means

that there is a possibility that in some novels female authors use many more adnominals than male writers and it is such works that create the wrong impression of high adnominal density of female written speech.

To explore this possibility we ranked the frequencies of adnominal values and obtained the following results, visualized in Figure 3.

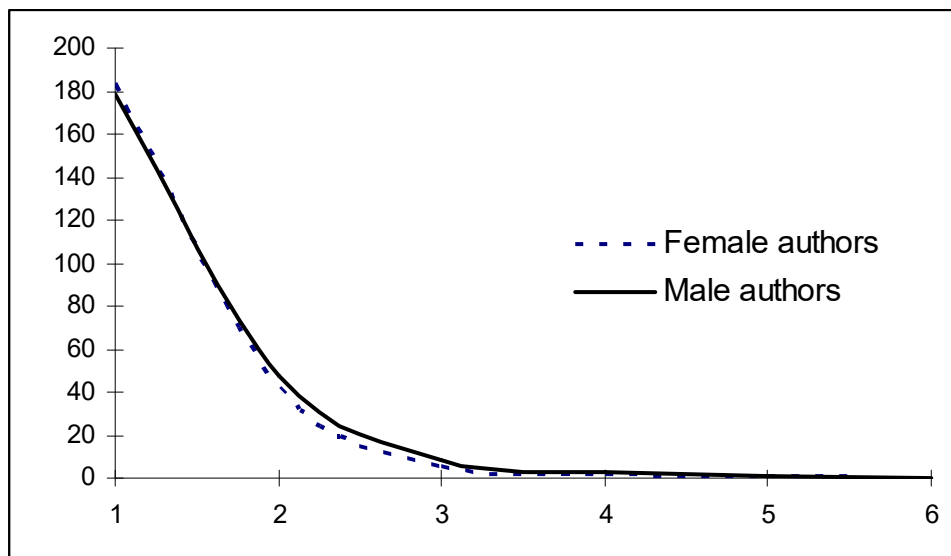


Figure 3 Ranked mean values

As seen in this figure the two distributions overlap, showing that the number of adnominals in the works with high descriptivity and low descriptivity do not depend on the gender.

One more possible explanation of the erroneous hypothesis that women use more adnominals is the following. What if the proportion of the adnominals and nouns with zero adnominal value is less in female works? If women use fewer such AV0 nouns, the general impression of descriptive intensity will be higher. In Table 5 we present the number of nouns with zero adnominal valency and Busemann's coefficient of such nouns against all nominals that are modified.

Table 5

Busemann's coefficient of modified and non-modified nominals

Text	Female		Busemann's coefficient (C)	Chi-square
	Nominals with AV1-6 (modified)	Nominals with AV0 (not modified)		
T1	185	347	0.35	49.33
T2	253	297	0.46	3.52
T3	293	293	0.50	0.00
T4	223	258	0.46	2.55
T5	213	222	0.49	0.19
T6	237	273	0.46	2.54

Male				
Text	Nominals with AV1-6 (modified)	Nominals with AV0 (not modified)	Busemann's coefficient (C)	Chi-square
T1	187	366	0.34	57.94
T2	205	318	0.39	24.41
T3	265	278	0.49	0.31
T4	252	296	0.46	3.53
T5	265	254	0.51	0.23
T6	252	291	0.46	2.80

The results lead to two main conclusions. First of all it is clear that in most works there is an approximate equality of nominals of the two above-mentioned classes. Secondly, there is no difference between genders in this respect. This is clearly demonstrated by the graph showing mean values of Busemann's coefficient (Figure 4).

A slight difference is observed only in one case for Rank 5; the other ranks demonstrate nearly complete coincidence.

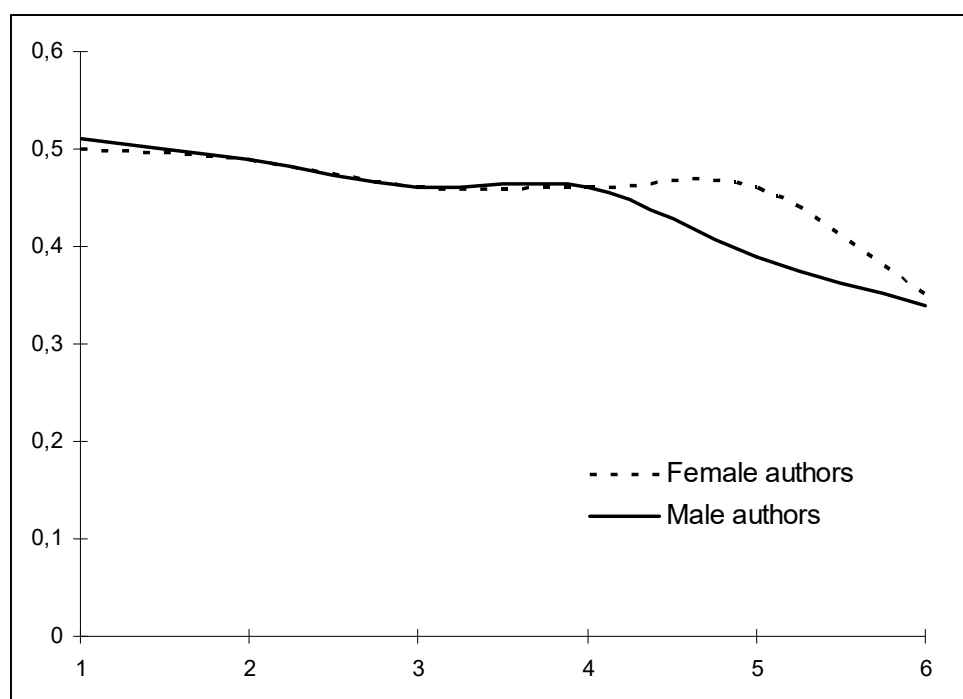


Figure 4 Ranked Busemann's coefficient values of the relationship of nominals with AV and non-modified nominals in the works by male and female authors

4. Zipf–Alekseev function

The study of AV distribution can use the frequencies of adnominal valencies directly or transfer them into a number of sequences, organized according to certain rules. In the first case we

obtain the type of distribution of the adnominal valencies on the surface level, and in the second case the distribution is studied at a deeper level, reflecting less obvious tendencies.

In this study we use both approaches and in both cases the fitting of the distribution of frequencies is done using the Zipf–Alekseev function (Hřebíček, 2002):

$$(2) \quad f_x = f_1 x^a + b \ln x \quad ,$$

where f_1 is the maximum frequency of the biggest AV score, a and b – parameters, x – the given AV type.

a. Surface level

We counted the number of adnominal valency types in all 12 texts and ranked the values in descending order. Their distribution was fitted by the Zipf–Alekseev function. The results are represented in Table 6.

Table 6
Total number of adnominal valencies in 12 texts

Valency	Observed	Theoretical
0	3493	3493.00
1	2171	2173.31
2	539	524.46
3	84	114.61
4	25	26.29
5	8	6.53
6	3	1.76
	a = 1.096 b = -2.568 R ² = 0.9999	

Judging by the determination coefficient the result of the fitting is very good. In Table 7 the distribution of adnominal valency types in 12 texts, now taken separately, is presented.

Table 7
Fitting of the Zipf–Alekseev function to the distribution of adnominal valencies in each text³

T1			T2			T3			T4		
R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.
0	347	347.00	0	297	297.00	0	293	293.00	0	258	258.00
1	163	163.00	1	187	187.34	1	247	247.06	1	174	174.33

³ R = Rank, Obs. = Observed, Th. = Theoretical.

Adnominal Valency in Modern Russian

2	20	19.97	2	53	51.08	2	41	40.42	2	43	40.91
3	2	2.14	3	9	12.81	3	3	5.20	3	4	8.47
			4	4	3.37	4	1	0.68	4	1	1.83
						5	1	0.10			
a = 1.489 b = -3.721 R ² = 0.9999			a = 0.938 b = -2.312 R ² = 0.9997			a = 2.416 b = -3.840 R ² = 0.9999			a = 1.333 b = -2.739 R ² = 0.9996		

T5			T6			T7			T8		
R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.
0	222	222.00	0	273	273.00	0	366	366.00	0	318	318.00
1	152	152.35	1	178	178.39	1	150	150.00	1	161	160.90
2	47	45.31	2	49	46.77	2	29	28.96	2	33	33.70
3	10	12.29	3	7	11.11	3	5	5.45	3	8	6.60
4	2	3.47	4	2	2.75	5	2	1.12	4	2	1.38
5	1	1.04	5	1	0.74	6	1	0.26	5	1	0.32
6	1	0.34									
a = 1.001 b = -2.228 R ² = 0.9998			a = 1.082 b = -2.447 R ² = 0.9997			a = 0.461 b = -2.521 R ² = 0.9999			a = 0.829 b = -2.615 R ² = 0.9999		

T9			T10			T11			T12		
R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.
0	278	278.00	0	296	296.00	0	254	254.00	0	291	291.00
1	212	211.86	1	191	191.90	1	178	179.56	1	178	178.97
2	39	40.01	2	56	50.94	2	69	62.84	2	60	55.62
3	8	6.24	3	3	12.30	3	14	20.42	3	11	16.34
4	6	1.00	4	2	3.10	4	3	6.88	4	2	5.04
						5	1	2.45	6	1	1.66
a = 1.954 b = -3.385 R ² = 0.9996			a = 1.044 b = -2.408 R ² = 0.9983			a = 0.818 b = -1.901 R ² = 0.9982			a = 0.6748 b = -1.9852 R ² = 0.9992		

The results again show that in all cases the fitting is very good. And again there does not seem to be any difference in the distribution of adnominal valencies due to the difference in gender of the authors or their professional background.

b. Motif distribution

The study of adnominal distribution can be further conducted by transferring the sequences of AV into new sequences, organized according to certain rules and reflecting a deeper structure of the adnominal organization of a text. In this study the formation of new sequences is done according to the principles of motif approach, worked out by R. Köhler (2008, 2015) and defined as non-decreasing sequences of numbers.

Let us take the above-mentioned sample from T5 (*The Stolen Sleep* by A. Marinina), this time without division into sentences:

1 1 0 2 1 0 1 0 1 1 1 3 1 1 0 0 0 1 1 3 0 1 2 1 2 1 1 2 2 0 1 1 0 1 0 1 1 1 1 0 2 0 2 3 1 0 1 1 0 1
0

Using the above-mentioned rule of non-decreasing values for every sequence, we break it into the following motifs which are placed in square brackets:

[1 1] [0 2] [1] [0 1] [0 1 1 1 3] [1 1] [0 0 0 1 1 3] [0 1 2] [1 2] [1 1 2 2] [0 1 1] [0 1]
[0 1 1 1 1] [0 2] [0 2 3] [1 0] [1 1] [0 1] [0].

Motifs can be analyzed from different angles, measuring number of their elements, distances between the same types in the text, their structure, etc. (Cech, Vincze, Altmann, 2016; Köhler, Naumann, 2008, 2016; Sanada, 2010; Liu, Fang, 2016; Wang, 2016). In our case motifs will be classified according to the total adnominal valency, i.e. the sum of all valencies in every motif. Using this method of typology it is possible to distinguish between the following motif types:

- AVM-1 – motifs with an adnominal valency of 1, occurs in five cases;
- AVM-2 – motifs with a valency of 2, occurs in six cases;
- AVM-3 – motifs with a valency of 3, two occurrences;
- AVM-4 – motifs with a valency of 4 – occurred once;
- AVM-5 and AVM-6 – both found twice.

There is also one case of AVM 0, but this type of motif can happen only at the end of the text and is omitted in the counts.

The results of such classification of the motifs in all the texts are presented in Table 8.

Table 8
Adnominal valency motifs

Texts	Type of AVM									
	1	2	3	4	5	6	7	8	9	13
Text 1	58	44	10	7	1					
Text 2	62	41	28	12	6	3	1			
Text 3	67	40	27	13	6	3	1		1	
Text 4	66	37	15	12	6	1	1			
Text 5	46	42	17	9	7	6				
Text 6	57	47	16	15	8	1				
Text 7	71	37	20	3	3					
Text 8	85	34	15	8	6					
Text 9	76	43	24	8	5	3	2		1	
Text 10	74	43	19	12	5	3	0	1		
Text 11	46	51	27	9	2	9	3	1	0	1
Text 12	61	55	24	13	6	0	1	1		
Total	769	514	242	121	61	29	9	3	2	1

Using means of motif types for each author we receive the data which can be graphically represented as shown in Figure 5.

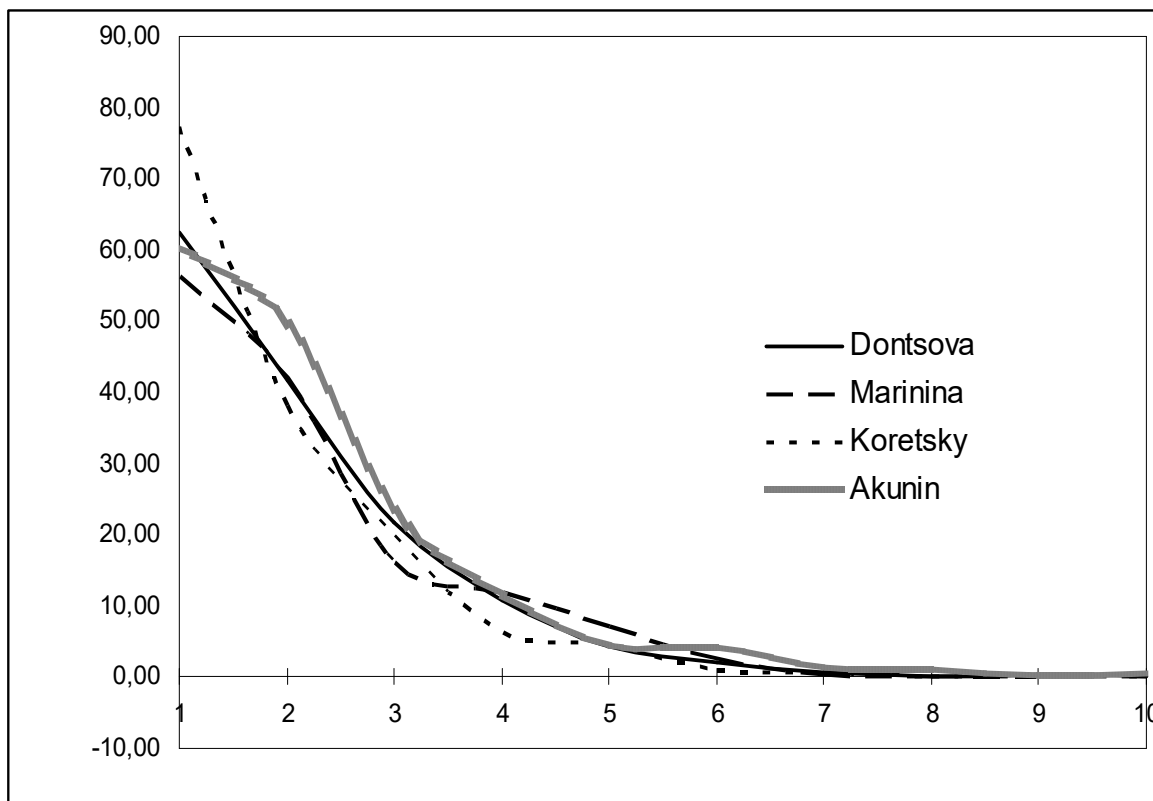


Figure 5 Motif types of each author

As is illustrated by the graph, the lines do not overlap completely, but still demonstrate similar patterns. The question arises if these distributions can be caught by the same formula.

Using the Zipf–Aleksseev function to fit the distribution of the types of motifs, the following results were obtained (Table 9).

Table 9
Fitting of the Zipf–Aleksseev function to the distribution of valency motifs

T1			T2			T3			T4		
R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.
1	58	58.00	1	62	62.00	1	67	67.00	1	66	66.00
2	44	43.60	2	41	43.17	2	40	42.26	2	37	36.29
3	10	12.18	3	28	23.59	3	27	22.91	3	15	17.72
4	7	3.00	4	12	12.89	4	13	12.73	4	12	9.04
			5	6	7.30	5	6	7.39	5	6	4.88
			6	3	4.29	6	3	4.48	6	1	2.78
						7	1	2.82	7	1	1.65

Adnominal Valency in Modern Russian

		9	1	1.83	
a = 1.313 b = -2.489 R ² = 0.9918	a = 0.088 b = -0.881 R ² = 0.9902	a = 0.132 b = -0.769 R ² = 0.9923	a = -0.292 b = -0.824 R ² = 0.9937		

T5			T6			T7			T8		
R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.
1	46	46.00	1	57	57.00	1	71	71.00	1	85	85.00
2	42	40.41	2	47	44.88	2	37	38.1	2	34	33.64
3	17	20.25	3	16	22.14	3	20	16.1	3	15	15.77
4	9	9.42	4	15	10.41	4	3	7.00	4	8	8.36
5	7	4.44	5	8	5.01	5	3	3.23	5	6	4.84
6	6	2.17	6	1	2.51						
a = 0.770 b = -1.381 R ² = 0.9791			A = 0.537 b = -1.273 R ² = 0.9708			a = -0.125 b = -1.118 R ² = 0.9900			a = -1.002 b = -0.484 R ² = 0.9995		

T9			T10			T11			T12		
R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.	R.	Obs.	Th.
1	76	76.00	1	74	74.00	2	51	51.00	1	61	61.00
2	43	43.93	2	43	42.61	1	46	46.42	2	55	54.44
3	24	21.12	3	19	20.49	3	27	25.11	3	24	25.83
4	8	10.45	4	12	10.14	4	9	12.65	4	13	11.23
5	5	5.45	5	5	5.30	6	9	6.44	5	6	4.94
6	3	2.99	6	3	2.91	7	3	3.38	7	1	2.26
7	2	1.72	8	1	1.67	5	2	1.84	8	1	1.08
9	1	1.02				8	1	1.03			
						10	1	0.60			
a = -0.151 b = -0.924 R ² = 0.9969			a = -0.159 b = -0.919 R ² = 0.9985			a = 0.734 b = -1.255 R ² = 0.9924			a = 0.893 b = -1.525 R ² = 0.9975		

As seen from the table the fitting is very good. One of the important steps in the direction from the description of regularities to their explanation is the study of possible relations of parameters (Grzybek, Kelih, Stadlober, 2009: 19). In the Zipf–Alekseev formula parameter “a” is interpreted as the feature of the language at large, whereas parameter “b” shows the changes made by the author of the text (Hřebíček, 2002). Using the values of parameter *b* in tables 8 and 10 it is possible to classify the texts, taking into account both surface-level and deep-level approaches.

Graphically this is represented in a scatterplot (Figure 6) in which the horizontal axis indicates the values of b parameter of the Zipf–Alekseev function fitting surface-level distribution and the vertical axis – the values of b parameter for adnominal valency motifs.

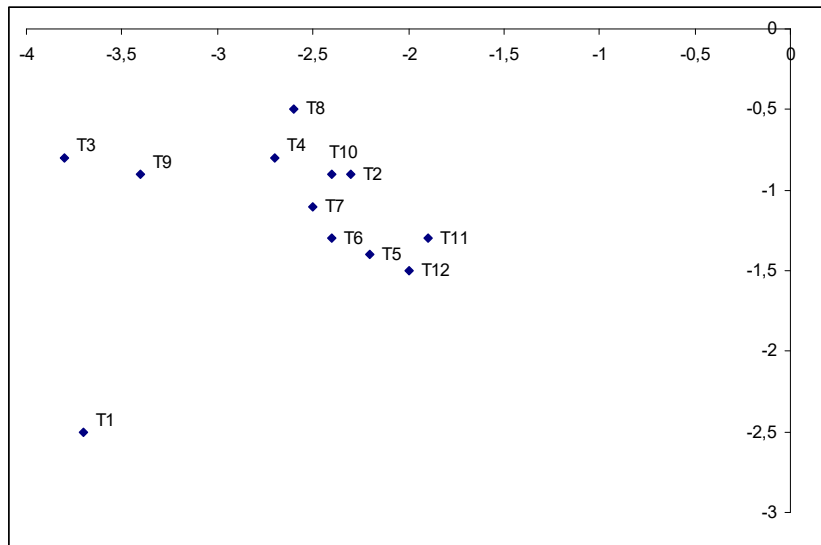


Figure 6 Scatterplot of 12 texts

As seen in the scatterplot, texts by Dontsova (T1, T2 and T3) are very dispersed, much more than the texts of the other authors. One of the possible explanations is that her texts were taken from different thematic series while the general number of books in all series written by this author is over 200. Contrary to her, the works of Marinina (T4, T5, T6), a retired police lieutenant-colonel, form a rather compact group. Another compact group (T10, T11, T12) is formed of texts of Akunun, organizing a certain “nucleus in the scatterplot.” The texts by Koresky (T7, T8, T9), also a former police professional, are dispersed, though the distances between them are smaller than those of Dontsova. In other words we again see that neither gender nor professional background influence the structure of adnominal valences in texts.

5. Conclusion

In the present article we tried to show that there is some regularity behind the optional choice of the adnominal valencies, that adnominal valencies may be guided by some general rule.

The study demonstrated that the distributions of adnominals are ordered very well by the Zipf–Alekseev function. Neither gender nor biographic background of the authors influence the distribution of adnominal valencies.

Other directions of the study of adnominal valencies may include other genres and other languages. If the relationship can also be shown for other text types and languages, it will be possible to suggest the existence of some general background law.

References

- Altmann, G.** (2015). *Problems in Quantitative Linguistics*. Vol. 5. Lüdenscheid: RAM-Verlag.
- Andreev, S., Místecký, M., Altmann, G.** (2018). *Sonnets: Quantitative Inquiries*. (Studies in Quantitative Linguistics, 29). Lüdenscheid: RAM.
- Andreev, S., Popescu, I.-I., Altmann, G.** (2017a). Some properties of adnominals in Russian texts. *Glottometrics* 38, 77-106.
- Andreev, S., Popescu, I.-I., Altmann, G.** (2017b). On Russian adnominals. *Glottometrics* 35, 64-83.
- Čech, R., Vincze, V., Altmann, G.** (2016). On motifs and verb valency. In: Liu, H., Liang, J. (eds.), *Motifs in language and text: 13-36*. Berlin: de Gruyter.
- Grzybek, P., Kelih, E., Stadlober, E.** (2009). Slavic letter frequencies. A common discrete model and regular parameter behavior. In: Köhler, R. (ed.), *Studies in quantitative linguistics. Issues in Quantitative linguistics: 17-33*. Lüdenscheid: RAM
- Köhler, R., Naumann, S.** (2008) Quantitative Text Analysis Using L-, F- and T-Segments. In: Preisach, C., Burkhardt, H., Schmidt-Thieme L., Decker R. (eds.), *Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization: 637-645*. Springer, Berlin, Heidelberg.
- Köhler, R., Naumann, S.** (2016). Syntactic text characterisation using linguistic S-Motifs. *Glottometrics* 34, 1-8.
- Liu, H., Fang, Yu.** (2016). Quantitative Aspects of Hierarchical Motifs. In: Emmerich Kelih, Róisín Knight, Ján Mačutek, Andrew Wilson (eds.), *Issues in Quantitative Linguistics. 4. Dedicated to Reinhard Köhler on the occasion of his 65th birthday: 9-26*. Lüdenscheid: RAM.
- Mehl M.R., Vazire S., Ramirez-Esparza N., Slatcher R. B., Pennebaker J.W.** (2007). Are women really more talkative than men? *Science* 317, 5834, 82.
- Místecký, M.** (2019). Five Ways of Investigating Adnominals in Czech Sonnets of the 19th and 20th Centuries. *Glottology* 9, 2, 173-200.
- Sanada, H.** (2010) Distribution of motifs in Japanese texts. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language. Structures, functions, interrelations, quantitative perspectives: 183-194*. Wien: Praesens.
- Wang, Y.** (2016). Quantitative Genre Analysis Using Linguistic Motifs. In: Liu, H. and Liang, J. (eds.), *Motifs in language and text: 165-180*. Berlin: de Gruyter.

Appendix

Data sources

- T1 D. Dontsova: *In', Jan' i vsjakaja drjan'* (*Yin-Yang and Various Stuff*)
T2 D. Dontsova: *Kleopatra s parashjutom* (*Cleopatra with a Parachute*)
T3 D. Dontsova: *Prodjuser koz'ej mordy* (*Producer of Dirty Tricks*)
T4 A. Marinina: *Kazn' bez zlogo umysla* (*Execution without Bad Intentions*)
T5 A. Marinina: *Ukradennyj son* (*The Stolen Dream*)
T6 A. Marinina: *Stechenie obstojatel'stv* (*Coincidence of Circumstances*)
T7 D. Koretsky: *Antikiller*
T8 D. Koretsky: *Antikiller-5*
T9 D. Koretsky: *Antikiller-6*
T10 B. Akunin: *Table-Talk*
T11 B. Akunin: *Pikovyj valet* (*Jack of Spades*)
T12 B. Akunin: *Turetskij gambit* (*Turkish Gambit*)

A Model of Clause Properties and the Zipf-Alekseev Function

Haruko Sanada¹, Gabriel Altmann²

Abstract

Frequency distributions of linguistic properties related to the clause are examined with respect to the Zipf-Alekseev function. Various properties of sentence and clause, i.e. (1) sentence length in terms of clauses, (2) clause length in terms of argument numbers (complements and adjuncts), (3) the position of a clause in the sentence, and (4) depth of a clause in the sentence, abide by the Zipf-Alekseev function. The determination coefficient is always greater than 0.9. It can be shown that any relation concerning length follows this function and that there are many other linguistic phenomena which abide by it.

Keywords: Zipf-Alekseev function; clause length; position of the clause; depth of the clause

1. Outline of the study

While quantifying in linguistics, one strives for attaining two aims: (1) To find models of the given properties in values of the variable, e.g. length, polysemy, inflectionality, diversification, etc. Unfortunately, this is done usually only for one language and a special type of texts; sometimes one takes a mixed sample, e.g. a corpus and hopes to find stronger confirmation. Unfortunately, a corpus is valuable only if one considers each text separately. (2) After having analyzed many texts/languages, one strives for finding a unique, common formula that expresses either several properties of the phenomenon or holds true for several (if possible all) languages.

None of these problems is simple. In the first step, one usually applies a soft-ware which finds the best model mechanically, but even here, one should not forget that the property we measured is our, human concept, exactly defined and using a quantification procedure “translated” into mathematics. There are even several prescriptions how to make something measurable and how to use the results in our formulas. The literature about measurement theory is enormous.

In the second problem which is not sufficiently developed, one strives for unification and simplification of the formulas. One tries to show (a) that the given formula holds for the same phenomenon in all languages and (b) that it holds for different phenomena in the given language. In this way, quantitative linguistics could be simplified as a whole.

For our purposes, refer to the article by Sanada (2019) in which three properties of Japanese are examined. The clause is understood as a part of a sentence containing a verb. As a grammatical level under the clause we also defined a unit called "argument" which is a com-

¹ Faculty of Economics, Risho University, Tokyo, Japan, hsanada@ris.ac.jp.

² Lüdenschied, Germany.

ponent of the clause in form of a complement, an adjunct, or a predicate. The detailed definitions and grammatical terms are shown in the next section. In principle, one should use the method which can be applied to all languages.

There are, of course, many hypotheses following from Köhler's (2005) circuit. The majority of properties are in some dependencies on other ones and the circuit is circular. Here we shall merely show the various distributions and a test can say whether they are correct. We start from the conjecture that many linguistic phenomena abide by the Zipf-Alekseev function – derived several times in linguistics – following the Weber-Fechner law from psychology. Here, the speaker/writer does not change the next frequency directly but logarithmically, i.e. we obtain the formula

$$y = cx^{(a + b * \ln(x))}. \quad (1)$$

In many cases, c can be replaced by the frequency in the $x = 1$ class, i.e. by (f_1) , because if $x = 1$ we obtain $y_1 = c$.

2. Grammatical definitions and explication of our data

The 'sentence' in Japanese is optically clear as it has a sign at the end. The clause is a topic that is still being discussed in Japanese linguistics. Minami (1974, 1993) analyzed grammatically important types of Japanese clauses³. Referring to his model we define that very clause should contain one predicate⁴, and may further consist of complements, and adjuncts. The argument is defined as the level between the clause and the morpheme. We regard the predicate and grammatical elements that are linked to the predicate as arguments in the clause. All elements in the clause are employed as arguments. We also consider the position of the clause in the sentence, counted from the beginning. In the case that the clause in the sentence is divided by an embedded clause, the beginning of the first half and the second half are taken as one clause. From the point of view of the depth of the clause, we defined an embedded clause as a lower level clause. Though there is no marker of the relative pronoun in Japanese, a sub-clause which corresponds to a clause follows a relative pronoun in English is also defined as a lower level clause. As the Japanese language is a right branching language, the absolute levels of all clauses in the sentence are determined by the predicate which appears at the end of the sentence. In other words, the levels among clauses are relative until the predicate which appears at the end of the sentence. Figure 1 shows a model of the clauses in the sentence in order to count the position of the clause and the embedded clause. The present study follows these definitions for consistency, which were employed in our former studies (Sanada 2012, 2014, 2015, 2016, 2018a, 2018b, 2019).

³ 'Minami's model' is also employed in software of the National Institute of Japanese Language and Linguistics to find the boundaries of some types of clauses. However, his model does not cover all types of Japanese clauses in the corpus because it focuses on grammatical and semantic aspects.

⁴ In two of the 692 clauses, a predicate is missing because the sentences are grammatically incorrect.

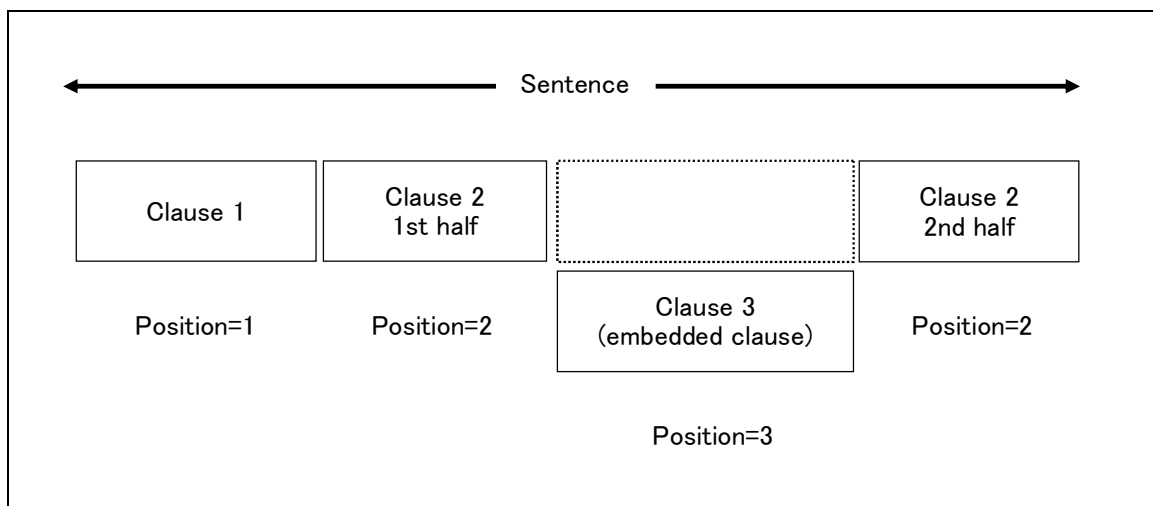


Figure 1 The position of the clause and the embedded clause.

The rules used for counting the data are shown here with examples from the Japanese valency database (Ogino et al. 2003), which were also shown in our former study. A space in the example indicates a morpheme boundary. A single slash mark (/) and a double slash mark (//) show the boundary of arguments and boundaries of the clauses, respectively. The number of clauses and arguments of the example follow its English equivalents. The numbers shown with ID after the example indicate a sentence number in the database.

Attributive elements, i.e. a noun and a postposition were treated as a part of the argument (see underlined words in Example 1). Japanese has no marker for relative pronouns in sub clauses (see the underlined beginning of the sentence in Example 2) or embedded clauses that divide the upper level clause into two parts (see the underlined words in Example 3).

Here are abbreviations and grammatical terms for the following examples: ATTR = attributive, CONNECT = connective form, CONTINUOUS = continuous aspect, COPULA, GEN = genitive, INS = instrument, LOC = location, NOM = nominalized form, OBJ = object, PAST = past tense form, PERFECT, PP = postposition, PROG = progressive form, SUBJ = subject⁵, and TOPIC.

Example 1:

Yogo no Ishikawa Keiko sensei wa,/ chugaku 3 nen no shojo no hahaoya ni/ at ta. (ID: JCO0217129)

[nurse-teacher-ATTR Mrs. Keiko Ishikawa-SUBJ/ junior high school 3rd grade-ATTR girl-GEN mother-OBJ/ meet-PAST]

"Mrs. Ishikawa Keiko of a nurse-teacher met a mother of the girl in the 3rd grade of the junior high school."

Clause = 1, Argument = 3.

Example 2:

Watashi ga/ at ta// chiji no hotondo wa,/ chiji shitsu ni/ nihon no ningyo ya okimono wo/ oi te i

⁵ The Japanese postposition *ga* is a subject marker and *wa* is a topic marker. However, *wa* also serves as a subject marker.

ta. (ID: JCO0138531)

[I-SUBJ/ meet-PAST// prefectural governors-ATTR most-TOPIC/ prefectural governor of-
fice-LOC/ Japan-ATTR dolls or ornamental objects-OBJ/ display-CONNECT- CONTINU-
OUS-PAST]

"Most of prefectural governors whom I met displayed Japanese dolls or ornaments in their office."

Clause=2, Argument = 6.

Example 3:

*Henshu cho shitsu de/ nan do ka/ at ta ga, // itsu mo/ taatorunekku no seeta ni // zakkuri shi ta //
sebiro wo/ haot te i ta.* (ID: JCO0209028)

[chief editor room-LOC/ several times/ meet-PAST-CONNECT, // always/ turtleneck-ATTR
sweater-CONNECT // roughly weaved-PAST // jacket-OBJ/ was wearing (PROG-PAST)]

"I met him for several times in the office of the chief editor, and he always wore a sweater with a turtleneck and a jacket which roughly weaved."

Clause = 3, Argument = 8.

For more detailed definitions, e.g. definitions related to the predicate or other special cases, Sanada (2016) should be referred to.

For the present investigation we employed the Japanese valency database (Ogino et al. 2003), the same one that was employed in our previous studies. 240 sentences containing the verb 'meet' were extracted from the valency database. Three of the 240 sentences have two predicates with the verb 'meet', and these extracted sentences also include many other verbs because each sentence has one or more predicates. We used the Japanese morphological analyzer *MeCab* (Graduate Schools of Informatics in Kyoto University et al. 2008) and the electronic dictionary *UniDic* (National Institute for Japanese Language and Linguistics 2008) for the extracted sentences. The software considers the boundary of the 'short unit' morpheme boundary (National Language Research Institute 1964). Errors were corrected by hand. In the 240 sentences, 691 clauses and 1,889 arguments were obtained.

3. Frequencies of properties related to the clause and the Zipf-Alekseev function

3. 1. Frequencies of length of sentences in terms of clauses

Considering first the length frequencies of Japanese sentences measured in terms of clause numbers we obtain the results presented in Table 1 and Figure 2.

Table 1

Frequencies of length of sentences in terms of clauses, which are extracted from newspaper database and contain an instance of the verb "meet"

Length	Frequency	Zipf-Alekseev function
1	30	30.00
2	74	77.26
3	66	63.30
4	43	39.23
5	19	22.33
6	8	12.44
a = 2.5358, b = -1.6896, R ² = 0.9815		

As can be seen, the Zipf-Alekseev function in which we exchanged c for $f(1)$ holds true. Figure 1 shows the image of the function. The determination coefficient, given by R^2 shows a satisfactory fitting.

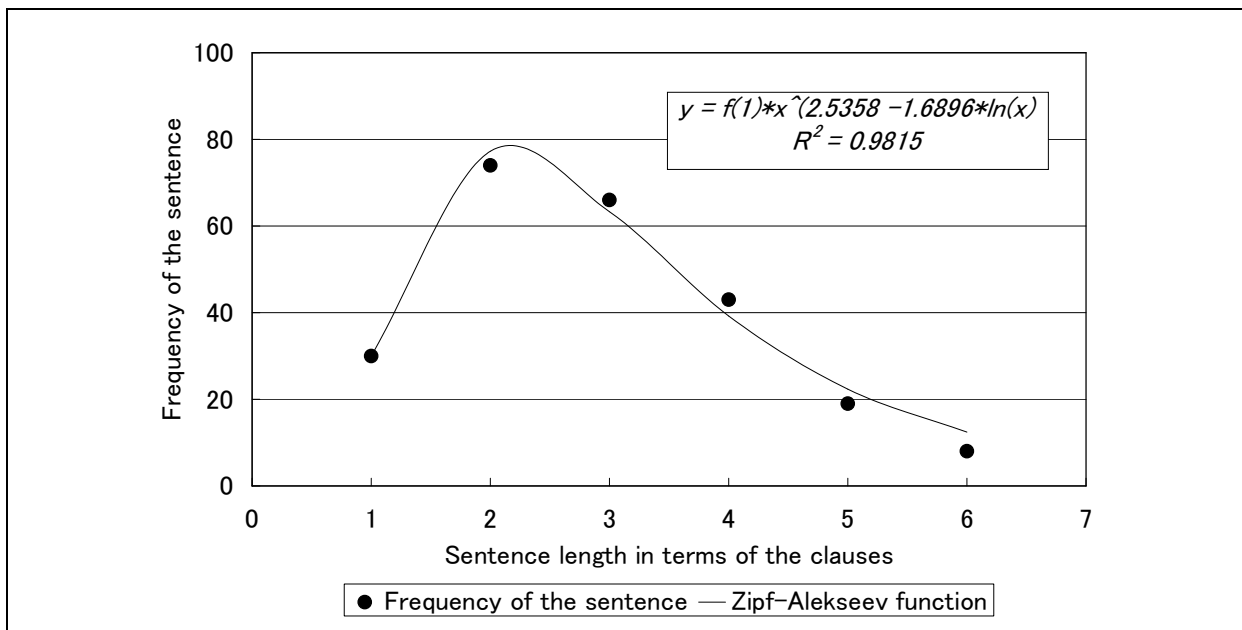


Figure 2 Distribution of the sentence length in terms of the clauses

3.2. Frequencies of clause length in terms of argument numbers

We take frequencies of clause length in terms of argument numbers, and the results are shown in Table 2 and Figure 3. The Zipf-Alekseev function fits to the data very well, and the determination coefficient, given by R^2 is very high.

Table 2

Frequencies of the clause length in terms of argument numbers

Length	Frequency	Zipf-Alekseev function
1	52	52.00
2	295	289.20
3	194	206.57
4	103	89.29
5	34	33.06
6	10	11.75
7	2	4.19
8	1	1.53
a = 4.5610, b = -3.0087, R ² = 0.9951		

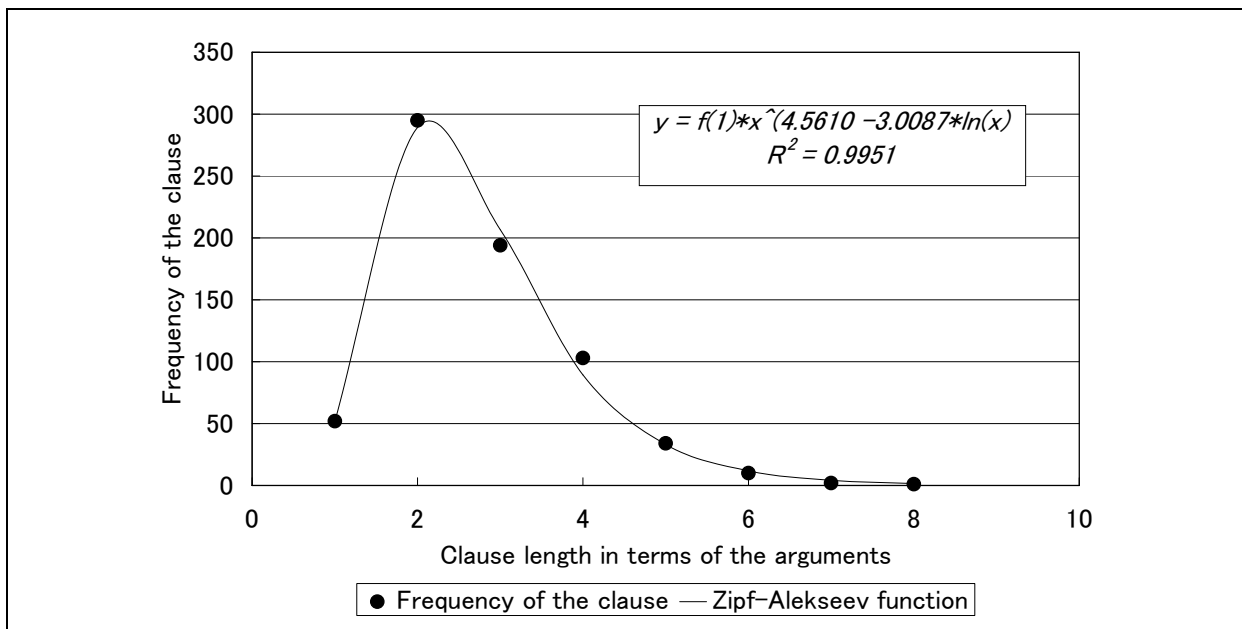


Figure 3 Frequencies of the clause length in terms of the arguments

3.3. Frequencies of the position of the clause

We take frequencies of the position of the clause in the sentence from the beginning. The model of the position of the clause is shown above in Figure 1. If there is only one clause in the sentence, it is, naturally, at position 1. But if there are several ones, they can occupy different positions. Table 3 shows the positioning of clauses in the sentence. The frequencies are shown in Table 3 and Figure 4 with the Zipf-Alekseev function, and the determination coefficient, given by R^2 is also very high. A shape of the curve is different from Figure 2 and Figure 3, and it shows a reverse S-shape.

Table 3
Frequencies of the position of the clause

Position	Frequency	Zipf-Alekseev function
1	240	240.00
2	210	217.33
3	136	123.68
4	70	66.12
5	27	35.74
6	8	19.88
a = 0.6437, b = -1.1351, R ² = 0.9905		

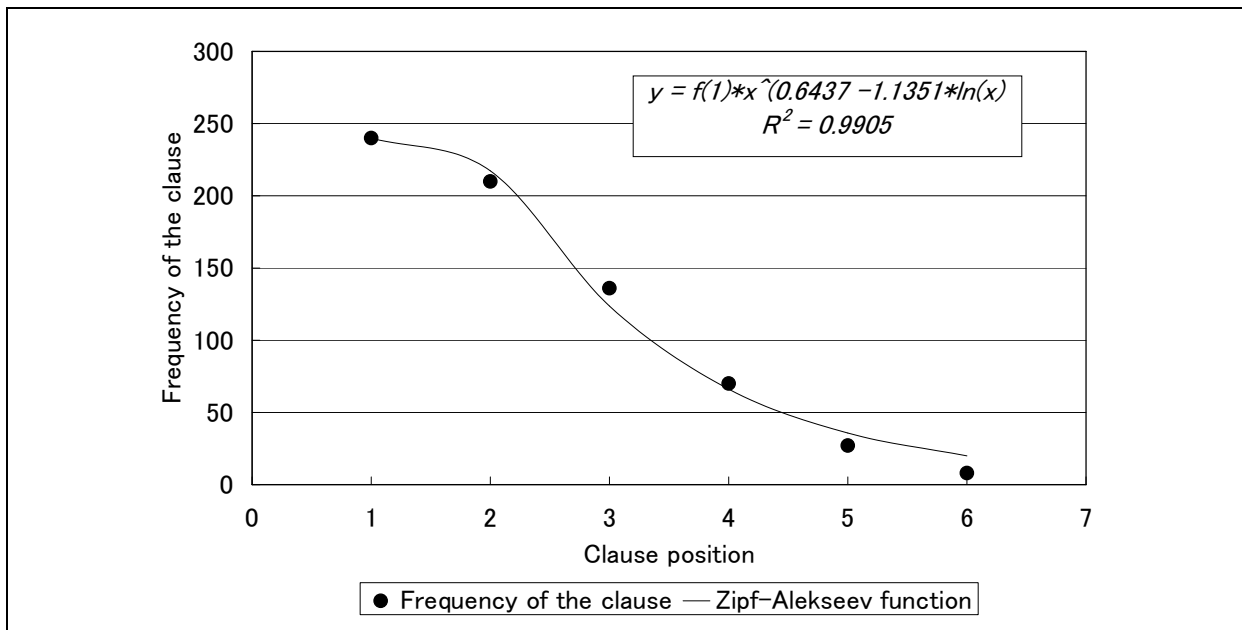


Figure 4 Frequencies of the position of the clause

3.4. Frequencies of the clause depths

We investigate the depth of the clauses, and Table 4 and Figure 5 present the depths of individual clauses. A main clause or a clause which is connected to the main clause by a conjunction has depth 1. Depth 2 or more shows the lower level clause including embedded clauses. The determination coefficient, given by R^2 is very high.

Table 4
Frequencies of the depth of the clause

Depth	Frequency	Zipf-Alekseev function
1	376	376.00
2	251	251.38
3	58	55.46
4	5	10.72
5	1	2.16
a = 1.4042, b = - 2.8639, R ² = 0.9996		

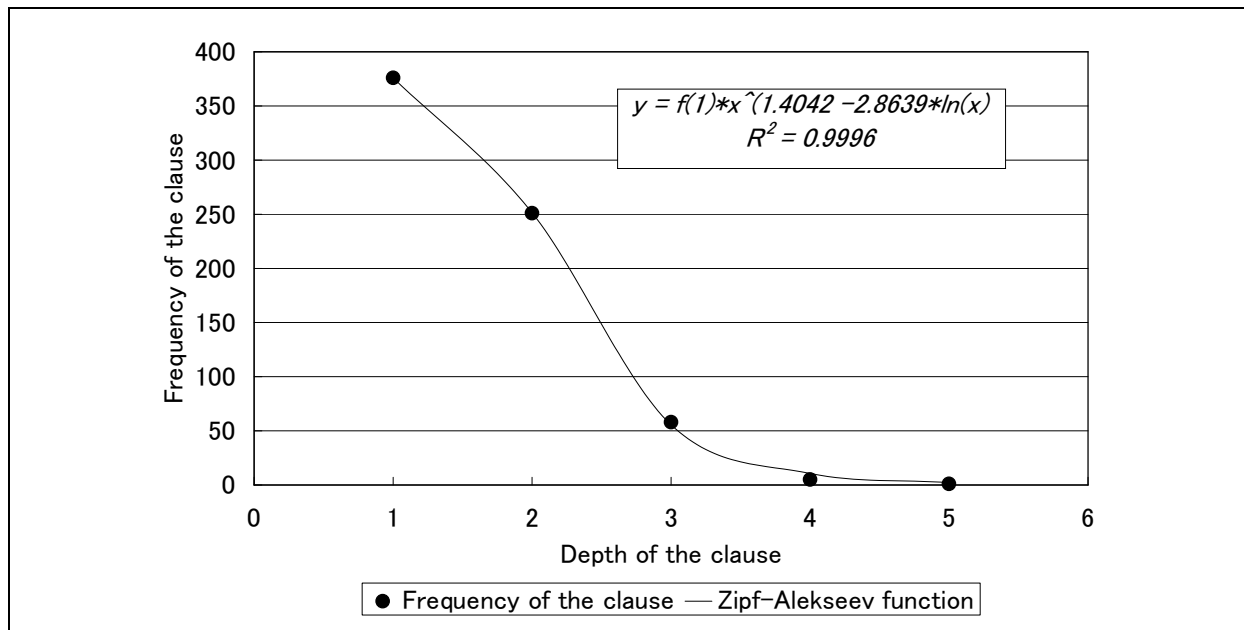


Figure 5 Frequencies of the depth of the clause

4. Discussion

As can be seen, various properties of sentence and clause abide by the Zipf-Alekseev function. The determination coefficient is always greater than 0.9. It would be very important to study these relations in other languages and in other texts in order to care for simplification and reduction of mathematics. It can be shown that any relation concerning length follows this function and that there are many other linguistic phenomena which abide by it.

References

- Köhler, R.** (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin/New York: de Gruyter.
- Minami, F.** (1974). *Gendai nihongo no kozo* (The structure of the present Japanese). Tokyo: Taishukan.
- Minami, F.** (1993). *Gendai nihongo bunpo no rinkaku* (The outline of the grammar of the present Japanese). Tokyo: Taishukan.
- National Language Research Institute** (1964). *Gendai Zasshi 90shu no Yogo Yoji: Dai3bunsatsu: Bunseki* (Vocabulary and Chinese Characters in Ninety Magazines of Today: vol. 3: Analysis of Results). Tokyo: Shuei Shuppan.
- Ogino, T., Masahiro Kobayashi, & Hitoshi Isahara.** (2003). *Nihongo Doshi no Ketsugoka* (Verb valency in Japanese). Tokyo: Sanseido.
- Sanada, H.** (2012). Joshi no Shiyo Dosu to Ketsugoka ni Kansuru Keiyoteki Bunseki Hoho no Kento (Quantitative approach to frequency data of Japanese postpositions and valency). *Rissho Daigaku Keizaigaku Kiho* (The quarterly report of economics of Rissho University) 62, 2, 1-35.
- Sanada, H.** (2014). The choice of postpositions of the subject and the ellipsis of the subject in Japanese. In: Uhlířová, L., Altmann, G., Čech, R., Mačutek, J. (eds.), *Empirical Approaches to Text and Language Analysis: 190-206*. Lüdenschied: RAM-Verlag.
- Sanada, H.** (2015). A co-occurrence and an order of valency in Japanese sentences. In: Tuzzi, A., Mačutek, J., Benešová, M. (eds.), *Recent Contributions to Quantitative Linguistics: 139-152*. Berlin: de Gruyter.
- Sanada, H.** (2016). The Menzerath-Altmann law and sentence structure. *Journal of Quantitative Linguistics* 23, 3, 256-277.
- Sanada, H.** (2018a). Negentropy of dependency types and parts of speech in the clause. In: Jiang, J., Liu, H. (eds.), *Quantitative analysis of dependency structures: 119-144*. Berlin: Mouton de Gruyter.
- Sanada, H.** (2018b). Quantitative interrelations of properties of complement and adjunct. In: Wang, L., Köhler, R. & Tuzzi, A. (eds.), *Structure, Function and Process in Texts: 78-99*. Lüdenscheid: RAM-Verlag.
- Sanada, H.** (2019). Quantitative aspects of the clause: the length, the position and the depth of the clause. *Journal of Quantitative Linguistics* 26, 4, 306-329.

Software and digital dictionaries

- Altmann-Fitter (2013). V. 3.1.3.0. Lüdenscheid: RAM-Verlag.
- Graduate Schools of Informatics in Kyoto University; NTT Communication Science Laboratories. (2008). Morphological analyzer: *MeCab*, version 0.97.
(<https://code.google.com/p/mecab/>)
- National Institute for Japanese Language and Linguistics. (2008). Digital dictionary for the natural language processing: *UniDic*, version 1.3.9.
(http://www.ninjal.ac.jp/corpus_center/unidic/)

The Flexibility of Parts-of-Speech Systems and Their Grammar Efficiency

Relja Vulcanović¹, Tayebeh Mosavi Miangah²

Abstract

We consider a recently proposed simplification of the formula for evaluating grammar efficiency. When applied to parts-of-speech systems, as defined by Hengeveld, the formula involves certain weighted quantities. We eliminate some of these weights, getting a new formula which is even simpler. We use the new formula to calculate the grammar efficiency of parts-of-speech systems found in a sample of natural languages. The results are compared to the ones obtained by Vulcanović's previous grammar-efficiency formula and a strong correlation is found. This confirms the viability of the new formula: it is simpler than the old one and produces comparable results.

Keywords: One-Meaning–One-Form Principle; bijection; propositional function; parts of speech; flexible parts-of-speech system; grammar efficiency; correlation

1. Introduction

Until relatively recently, the opinion that all languages are equally complex has been prevalent in linguistics; see the introductory surveys in (Miestamo et al. 2008) and (Newmeyer & Preston 2014). These two volumes contain papers presented at the respective meetings in 2005³ and 2012.⁴ Before that, a double issue of *Linguistic Typology* (vol. 5-2/3, 2001) was devoted to the discussion of language complexity introduced by (McWhorter 2001) as the leading article. These events indicate that there is a growing interest in the complexity of language and grammar.

McWhorter's (2001) article describes a possible metric of language complexity consisting of four components: phonemic, syntactic, semantic/pragmatic, and morphological. Vulcanović (2007) shows that his formal mathematical measure of grammar complexity agrees with the last three metric components (phonemes have never been included in Vulcanović's model). The measure of grammar complexity in (Vulanović 2007) is based on (Vulanović 2003), which is mainly concerned not with grammar complexity but with grammar efficiency. However, it is quite easy to switch from grammar efficiency to grammar complexity in Vulcanović's approach because he defines their measures as reciprocals of one another. The relation between grammar complexity and efficiency is not so simple in (Hawkins 2004). Hawkins even states that effi-

¹ Department of Mathematical Sciences, Kent State University at Stark, North Canton, Ohio, USA, rvulanov@kent.edu.

² Department of Linguistics, Payame Noor University, Tehran, Iran (mosavit@pnu.ac.ir). Work done while visiting Kent State University at Stark.

³ Approaches to Complexity in Language, held in Helsinki.

⁴ Formal Linguistics and the Measurement of Grammatical Complexity, held in Seattle, Washington.

ciency “may involve more or less complexity, depending on the syntactic and semantic representations to be assigned to a given sentence and on their required minimum of complexity” (p. 9). He proposes “three very general principles of efficiency that are suggested by the preferences of performance and grammars” (ibid.). The three principles, which are supposed to govern how the human processor works, are: Minimize Domains, Minimize Forms, and Maximize Online Processing (Hawkins 2004: Chapter 3; see also Hawkins 2014: 63). Vulanović, on the other hand, approaches grammar efficiency like a machine efficiency, viewing grammar as a machine that converts linguistic input into an output containing some linguistically relevant information. Therefore, he defines grammar efficiency as directly proportional to the quotient of the quantity measuring the output and the quantity measuring the input.

The present paper is another step in the development of the formula for measuring grammar efficiency. To specify its relation to the previous results, we list below the main stages of Vulanović’s work on grammar efficiency:

- 1) introduction of the concept in (Vulanović 1991, 1993),
- 2) further development of the grammar-efficiency formula using the abstract mathematical approach (Vulanović 2003) and a description that is adjusted to linguists (Vulanović 2007),
- 3) significant modification in Vulanović (to appear), based on the inclusion of a measure, denoted by μ , of how much a linguistic system violates the One-Meaning–One-Form Principle.⁵

The first work on the measure μ mentioned in stage 3) is (Vulanović & Ruff 2018). The original measure introduced there is modified and simplified in Vulanović (to appear) before it is incorporated in the formula for grammar efficiency of parts-of-speech (PoS) systems, taken in Hengeveld’s (1992) sense (see also Hengeveld et al. 2004). When the measure of the extent of violation of the One-Meaning–One-Form Principle is applied to PoS systems, it evaluates the degree of their flexibility. As demonstrated in Vulanović (to appear), the inclusion of the measure μ simplifies the grammar-efficiency formula to a great extent because the amount of the necessary calculations is reduced significantly.

During the period between the above stages 2) and 3), various linguistic phenomena were represented and analyzed using the grammar-efficiency formula from stage 2). Many of these papers concerned PoS systems, like (Vulanović 2008) for instance. Therefore, a natural question at stage 3) was whether the simplified formula produces results similar to those of the preceding period. This is why the correlation between the results obtained by the two formulas was considered in Vulanović (to appear). It is found there that the correlation is strong. Since the stage 3) formula is much simpler, it is more advantageous to use it than the previous one. This justifies our present interest in the stage 3) formula.

First, we here continue the simplification trend although our modification of the grammar-efficiency formula from Vulanović (to appear) is not so extensive. We only change one aspect of the measure μ . Measures of three sets are used in that part of the formula, those of the set of

⁵ The principle is mentioned for the first time under this name in (Anttila 1972: 181).

propositional functions, the set of word classes, and the set of ordered pairs showing what word classes are used in the PoS system to fulfill what propositional functions. All three sets are weighted in Vulanović (to appear). Here, we show that the sets of propositional functions and word classes do not have to be weighted. Therefore, this is a further simplification of the formula found in Vulanović (to appear). We show that the simplified formula still correlates well with the grammar-efficiency values obtained using the old formula of (Vulanović 2003, 2007). The way we analyze the correlation is different from what is done in Vulanović (to appear), where abstract PoS systems are used and their maximum grammar-efficiency values are calculated by the stage 2) and stage 3) formulas and compared. Here, we use PoS systems that are modeled after 14 natural languages, taken from the samples in (Hengeveld et al. 2004) and (Hengeveld & van Lier 2008). This is a more realistic comparison of the two formulas and it still confirms the viability of the approach that involves μ .

The rest of the paper is organized as follows. PoS systems are described in section 2 and then, in section 3, it is discussed how to measure their flexibility. Grammar efficiency is defined in section 4, the language sample is introduced in section 5, and section 6 presents the results of the grammar-efficiency calculations by the two formulas, as well as their correlation. Some concluding remarks are made in section 7.

2. Part-of-speech systems

The word classes that are relevant in Hengeveld’s approach to PoS systems (Hengeveld 1992, Hengeveld et al. 2004, Hengeveld & van Lier 2010) are defined based on what propositional functions they can assume in sentences. Four propositional functions (syntactic slots) are considered:

- the head of the predicate phrase (denoted by P),
- the modifier of the predicate phrase (p),
- the head of the referential (nominal) phrase (R), and
- the modifier of the referential phrase (r).

In this sense, there are 15 theoretically possible word classes. They are presented in Table 1. According to (Hengeveld & van Lier 2010), some word classes are unattested. They are marked in Table 1 with an asterisk. All word classes are labeled, but most of those that are unattested are left unnamed. The word classes that have exactly one propositional function are called *rigid*. They are verbs, nouns, adjectives, and manner adverbs. The remaining eleven word classes are *flexible*, which means that each has more than one propositional function.

Table 1
Word classes and the propositional functions they fulfill

Word class	P	R	r	p
V = verbs	+			
N = nouns		+		
a = adjectives			+	

m = manner adverbs				+
H = heads	+	+		
\mathcal{P} = predicatives	+			+
\mathcal{N} = nominals		+		+
M = modifiers			+	+
X_P (*)	+		+	
X_R (*)		+		+
Y_r	+	+	+	
Y_p (*)	+	+		+
Z = non-nouns (*)	+		+	+
Λ = non-verbs		+	+	+
C = contentives	+	+	+	+

A PoS system can be described by referring to its propositional functions and its word classes. There are PoS systems that lack some propositional functions. For instance, Tagalog, which is one of the languages in our sample, has such a PoS system. The propositional functions in Tagalog are P, R, and r. In general, the head functions P and R must be present in every PoS system, while the modifier functions r and p may be absent. The only exception to this is the PoS system in which P is the only propositional function. This system is considered in (Hengeveld 1992, Hengeveld et al. 2004, Hengeveld & van Lier 2010) because there are languages (e.g., Tuscarora) with PoS systems which are close to this theoretical extreme. Thus, one of the five propositional-function combinations in Table 2 is present in each PoS system.

Table 2

The possible propositional functions in PoS systems

Number of propositional functions	4	3	3	2	1
Propositional functions in the PoS system	P, R, r, p	P, R, r	P, R, p	P, R	P

A PoS system is called *rigid* if all its word classes are rigid. Otherwise, a PoS system is *flexible*, that is, it has at least one word class which is flexible. A PoS system is called *basic* (or *main*) if each of its propositional functions is fulfilled by exactly one word class (which also may have other functions). Seven basic PoS system types, see Table 3, are introduced in (Hengeveld 1992; cf. also Hengeveld et al. 2004) as points on a scale that can be used for classification purposes. Of those 7 types, only types 1, 2 and 3 are flexible.

Table 3

The 7 basic PoS system types according to (Hengeveld 1992)

PoS system type	P	R	r	p
1	C	C	C	C
2	V	Λ	Λ	Λ

3	V	N	M	M
4	V	N	a	m
5	V	N	a	—
6	V	N	—	—
7	V	—	—	—

In addition to the basic PoS systems types, the so-called *intermediate* PoS system types are considered in (Hengeveld et al. 2004, Hengeveld & Rijkhoff 2005). An intermediate PoS system type has features that fall in between two neighboring basic types. We only briefly mention the intermediate types here as they are irrelevant for the work presented here. In addition to types 1-3, we consider general flexible non-basic PoS systems. They can be described as having at least one propositional function which is fulfilled by at least two word classes. The flexible intermediate types belong to this group.

Hengeveld & van Lier (2010) discuss all theoretically possible basic PoS systems. There are 23 such systems that are flexible. They are important for our work and they are presented in Table 5 in the next section.

3. Measuring the flexibility of parts-of-speech systems

Let F and W denote the sets of all propositional functions and all word classes, respectively, in a PoS system, as described in the previous section. Whenever there are two sets, where each element of one set is paired up with one or more elements of another set, this is, mathematically speaking, a relation between the two sets. If each element one set is paired up with exactly one element of another set, and the other way around, the relation between the sets is called a *one-to-one correspondence*, or a *bijection*. Therefore, a PoS system represents a relation between F and W which is not necessarily a bijection. A measure of how far a relation between two sets is from a bijection is proposed in (Vulanović & Ruff 2018).

In general, if X and Y are two non-empty sets, a relation Φ between them is a set of ordered pairs from $X \times Y = \{(x, y) : x \in X, y \in Y\}$. We assume that for every $x \in X$ there exists an element $y \in Y$ such that $(x, y) \in \Phi$, and vice versa. Moreover, let $v_X(y)$ indicate the number of elements of X that are paired up with $y \in Y$ and let $v_Y(x)$ have an analogous meaning. Let the set B denote the set of all one-to-one ordered pairs in Φ ,

$$B = \{(x, y) \in \Phi : v_X(y) = v_Y(x) = 1\}.$$

The relation Φ is a bijection if and only if $\Phi = B$.

Let $\mu(\Phi)$ denote the quantity that measures how far Φ is from a bijection. The original formula for $\mu(\Phi)$ from (Vulanović & Ruff 2018) is simplified in Vulanović (to appear) to

$$\mu(\Phi) = \frac{|\Phi \setminus B|}{\min\{|X|, |Y|\}} + 1, \tag{1}$$

where $|\cdot|$ stands for the number of elements in the corresponding set. The formula (1) is motivated by the following considerations:

- (i) $\mu(\Phi) = 1$ if Φ is a bijection, otherwise $\mu(\Phi) > 1$.
- (ii) $\mu(\Phi)$ is greater if $|\Phi|$ is greater.
- (iii) $\mu(\Phi)$ is greater if $|B|$ is smaller.
- (iv) $\mu(\Phi)$ is greater if $|X|$ and $|Y|$ are smaller.
- (v) $\mu(\Phi) = \mu(\Phi^{-1})$, where $\Phi^{-1} = \{(y, x) : (x, y) \in \Phi\}$.

If Φ is a bijection, then $|\Phi \setminus B| = |\emptyset| = 0$. On the other hand, if Φ is not a bijection, then $\Phi \setminus B \neq \emptyset$ and $|\Phi \setminus B| > 0$. This is why property (i) is satisfied. Properties (ii) and (iii) are true because B is a subset of Φ and $|\Phi \setminus B| = |\Phi| - |B|$. Finally, properties (iv) and (v) are obvious.

It should be mentioned that (iv) means that $\mu(\Phi)$ is a relative measure. Without this requirement, the measure $\mu(\Phi) = |\Phi \setminus B| + 1$ would be appropriate because it satisfies the remaining four properties. The reason why $\mu(\Phi)$ should be a relative measure is illustrated by the following example. Consider the very simple case with $X = \{x_1, x_2\}$ and $Y = \{y\}$, and let $\Phi = \Phi_1 = \{(x_1, y), (x_2, y)\}$. Then, there is no one-to-one pair in Φ_1 and the corresponding set $B = B_1$ is empty, giving $|\Phi_1 \setminus B_1| + 1 = 3$. Now, form a new set of pairs, Φ_2 , by adding 98 one-to-one pairs to Φ_1 . The corresponding set B_2 contains 98 elements, thus, 98% of pairs in Φ_2 are one-to-one. Because of this, Φ_2 is closer to a bijection than Φ_1 , but $|\Phi_2 \setminus B_2| + 1$ is still equal to 3. On the other hand, the formula in (1) puts Φ_1 and Φ_2 in the correct relative position,

$$\mu(\Phi_1) = \frac{2}{\min\{2, 1\}} + 1 = 3, \quad \mu(\Phi_2) = \frac{2}{\min\{100, 99\}} + 1 = 1.020.$$

From the linguistic point of view, the motivation for defining the measures like μ comes from the need to quantify how much a linguistic system departs from the One-Meaning–One-Form Principle (Anttila 1972: 181). Vulcanović and Ruff (2018) discuss two linguistic applications of their formula for measuring the degree of violation of this principle. One of the applications is to the PoS systems that are described in the previous section. In this application X is the set F of the propositional functions in the PoS system, and Y is the set W of the word classes used. Thus, when applied to PoS systems, the formula (1) becomes

$$\mu(\Phi) = \frac{|\Phi \setminus B|}{\min\{|F|, |W|\}} + 1, \tag{1'}$$

with $B \subseteq \Phi \subseteq F \times W$. However, Vulcanović & Ruff (2018) show that a weighted version of their formula is more suited for applications to PoS systems. In the same way, Vulcanović (to appear) generalizes formula (1') to its weighted version (2),

$$\mu(\Phi) = \mu_w(\Phi) := \frac{\|\Phi \setminus B\|}{\min\{\|F\|, \|W\|\}} + 1, \quad (2)$$

which he uses to define and analyze the grammar efficiency of Hengeveld’s PoS systems. In this formula, $\|\cdot\|$ replaces the number of elements of the corresponding sets with the sum of their weights. A weight w is a number such that $w \geq 1$. If all weights in a set are equal to 1, then $\|\cdot\|$ is the same as $|\cdot|$, which is why (2) generalizes (1’).

Since the elements in the sets F and W may have different weights, property (v) may be lost. The other four properties are preserved since each element of the set B carries the same weight as it does in Φ , which implies that $\|\Phi \setminus B\| = \|\Phi\| - \|B\|$.

As we have already indicated, formula (2) is a simplification of the formula introduced in (Vulanović & Ruff 2018). Here, we take the simplification one step further and use (2’) instead of (2),

$$\mu(\Phi) = \mu'_w(\Phi) := \frac{\|\Phi \setminus B\|}{\min\{|F|, |W|\}} + 1. \quad (2')$$

Thus, our simplification is related to the use of weights; the set $\Phi \setminus B$ is the only one that is weighted.

The weights assigned to the word classes in Vulanović (to appear) are simpler than those defined in (Vulanović & Ruff 2018) but are still relatively complicated. This is why $\|W\|$ in (2) is replaced with $|W|$ in (2’), which is the simplest possible choice of the word-class weights. Then, it is only natural to replace $\|F\|$ with $|F|$ as well. Nevertheless, weights are assigned to the propositional functions in order to define the weights for the ordered pairs in Φ as in (3),

$$w((x, y)) = w(x), \quad (x, y) \in \Phi, \quad (3)$$

where $w(\cdot)$ indicates the weight of the corresponding element. Formula (3) agrees with Vulanović (to appear) because $w((x, y))$ is defined there as $w(x)w(y)$ and here, $w(y) = 1$ for each word class y . The weights of the propositional functions are given in Table 4. They are the same as in Vulanović (to appear) and as in one of the two weighting systems considered in (Vulanović & Ruff 2018). The weights in Table 4 are based on Table 2. We count how many times each propositional function appears in Table 2 and divide this count by 2 so that the weights of r and p are both equal to 1. With the weights defined above, we use formula (2’) throughout the rest of the paper and refer to $\mu'_w(\Phi)$ simply as μ .

Table 4
Propositional-function weights

Propositional function	P	R	r	p
Weight	2.5	2	1	1

We next evaluate μ for each theoretically possible basic PoS system (Hengeveld & van Lier 2010). The values are shown in Table 5, where only flexible basic PoS systems are presented. In a system of this kind, the number of word classes is less than the number of propositional functions. The rigid PoS systems (types 4-7 in Table 3) are not presented because they all have $\mu = 1$, Φ being a bijection. In Table 5, a PoS system is represented by the sequence of its word classes, listed in the order which corresponds to the P-R-r-p order of the propositional functions. We use \emptyset to indicate that the corresponding propositional function does not exist in the PoS system, because of which there is no word class with that function.

Table 5
The flexible basic PoS systems

$ F $	$ W $	PoS system	$\ \Phi \setminus B\ $	μ
4	3	VNMM (type 3)	2	1.667
		$V\mathcal{U}\mathcal{U}m, VX_RaX_R$	3	2.000
		$\mathcal{P}Na\mathcal{P}, X_pNX_{pm}$	3.5	2.167
		HHam	4.5	2.500
	2	V $\Lambda\Lambda\Lambda$ (type 2)	4	3.000
		ZNZZ	4.5	3.250
		$Y_rY_rY_{rm}, Y_pY_{pa}Y_p$	5.5	3.750
		$\mathcal{P}\mathcal{U}\mathcal{U}\mathcal{P}, HHMM, X_pX_RX_pX_R$	6.5	4.250
1	CCCC (type 1)	6.5	7.500	
3	2	$V\mathcal{U}\mathcal{U}\emptyset, VX_R\emptyset X_R$	3	2.500
		$\mathcal{P}N\emptyset\mathcal{P}, X_pNX_{p\emptyset}$	3.5	2.750
		HHa \emptyset , HH \emptyset m	4.5	3.250
	1	$Y_rY_rY_{r\emptyset}, Y_pY_{p\emptyset}Y_p$	5.5	6.500
2	1	HH $\emptyset\emptyset$	4.5	5.500

Note that if we left the set $\Phi \setminus B$ without weights, there would be larger groups of PoS systems in Table 5 with the same value of μ . For instance, the systems that share $|F| = 4$ and $|W| = 3$ would all have $\mu = 1.667$. With the weights, the measure of flexibility of a PoS system is increased if the word classes do not convey one or both heads unambiguously. Formula (2') still leaves groups of PoS systems with the same value of μ , but at least in some cases (HHa \emptyset and HH \emptyset m, for instance), this is quite acceptable. The formulas used in (Vulanović & Ruff 2018, Vulanović to appear; see (2) for the latter) provide finer scales of the flexibility of PoS systems, but these formulas are more complicated to evaluate because they involve $\|F\|$ and $\|W\|$. The values of μ that are calculated by the formula (2') are still quite adequate for measuring the grammar efficiency of PoS systems.

We finish the section by illustrating how μ is calculated for flexible PoS systems that are not basic. Such systems are more complicated than any of the systems in Table 5. As an example, we consider the PoS system that can be found in Santali, one of the languages in our sample. Santali has all four propositional functions and uses 3 word classes to fulfill them. The relation

Φ between the propositional functions and the word classes in Santali is shown in Table 6 (Hengeveld & van Lier 2008).

Table 6
The PoS system of Santali

Propositional functions	P	R	r	p
Rigid word classes	V	N		
Flexible word classes	C	C	C	C

We have $B = \emptyset$ and $\Phi = \{(P, V), (P, C), (R, N), (R, C), (r, C), (p, C)\}$, thus, $\|\Phi \setminus B\| = 2.5 + 2.5 + 2 + 2 + 1 + 1 = 11$. Therefore, the flexibility of the Santali PoS system is measured by

$$\mu_{\text{Santali}} = \frac{11}{3} + 1 = \frac{14}{3} = 4.667.$$

4. Grammar efficiency

We use the same approach as in Vulanović (to appear) to incorporate the measure μ of the PoS-system flexibility in the formula for calculating the grammar efficiency of the system. We only model the structure of simple intransitive sentences.

Let us continue to use the example of Santali from the previous section to illustrate what needs to be done to evaluate the grammar efficiency of a PoS system. The basic word order in Santali (Cole 1896) can be described by referring to the propositional functions as rRpP. We only consider this basic order, and, therefore, any simple intransitive sentence in Santali should be interpreted as at least one of the following strings of propositional functions:

$$\text{RP, rRP, RpP, rRpP.} \quad (4)$$

For instance, the string CNV of word classes is a formal representation of a sentence in Santali. This is a sentence because CNV can be analyzed as rRP. We assume that the analysis is carried out from left to right, one word-class symbol at a time, and without the information about the length of the string of word classes. For simplicity, we also assume that both predicate and referential phrases are continuous, that is, that the modifying propositional functions stand next to their heads. This has been an ongoing assumption since stage 2) mentioned in the Introduction.

At the same time, CNV is an unambiguous sentence because rRP is its only interpretation. On the other hand, CCC is an example of an ambiguous sentence since it can be interpreted as both rRP and RpP, whereas VC is not a sentence at all because its analysis does not give any of the strings in (4). All unambiguous sentences in Santali are listed below:

$$\text{NV, CV, NC, CC, CNV, NCV, NCC, CNC, CNCV, CCCV, CNCC, CCCC.} \quad (5)$$

Let s denote the number of unambiguous sentences permitted by the PoS system. In Santali, $s = 12$. This count is compared to all possible sentences, ambiguous or not, when the orders or propositional functions prescribed in (4) are ignored. The number of such sentences is denoted by \hat{s} . A PoS system like that of Santali can permit 25 sentences in addition to those in (5),

VN, VC, CN, VNC, VCN, NVC, CVN, CCN, VCC, CVC, CCV, CCC,
 VCNC, VCCN, CVNC, CVCN, CNVC, NCVC, NCCV, VCCC, CVCC, CCVC,
 NCCC, CCNC, CCCN.

Thus, for Santali, $\hat{s} = 37$. The quotient s/\hat{s} can be used to measure how free word order is in a PoS system. However, it is shown in Vulanović (to appear) that this quotient needs to be modified to include the number of all possible orders of propositional functions. The modification is s/m , where $m = \max\{\hat{s}, f\}$ and the values of f are given in Table 7. These values depend on how many propositional functions the PoS system has. For Santali, $m = \max\{37, 18\} = 37$.

Table 7

The values of f , the maximum number of orders of propositional functions assuming the continuity of both predicate and referential phrases

$ F $	4	3	2	1
f	18	6	2	1

Let us illustrate how the numbers in Table 7 are obtained by considering the case $|F| = 4$. Taking into account that both predicate and referential phrases are continuous, we have the following possible orders of propositional functions, thus $f = 18$:

PR, RP, PRr, PrR, RrP, rRP, PpR, pPR, RPp, RpP,
 PpRr, PprR, pPRr, pPrR, RrPp, RrpP, rRPp, and rRpP.

What words play the role of what propositional functions is determined by two grammatical devices, the word classes in the PoS system and the rules governing the permitted orders of propositional functions. Therefore, by the *grammar of a PoS system*, we mean the PoS system with its propositional functions and word classes, equipped with the permitted orders of propositional functions.

We can now define the absolute grammar efficiency, AE , of a PoS system:

$$AE = Q \frac{|F|}{|W|}, \quad (6)$$

where

$$Q = \frac{s}{m} \cdot \frac{1}{\mu}. \quad (6')$$

This definition is based on the following:

- The grammar is more efficient if its word order is less restricted. This is because word-order rules are syntactic rules and the grammar is more complex (and therefore less efficient) if its syntax has more rules to process (McWhorter 2001: 136).⁶ Accordingly, *AE* is directly proportional to the quotient s/m which measures how free word order is in the sentences permitted in the PoS system. If word order is free, then $s/m = 1$, otherwise, $s/m < 1$.
- The grammar is more efficient if it can process sentences more easily, which is the case if word classes are less flexible. Hence, *AE* is inversely proportional to μ (Vulanović to appear). This is in agreement with Miestamo (2008), who states that grammars are more complex if their degree of violation of the One-Meaning–One-Form Principle is greater (recall that μ , more generally speaking, is a measure of this degree).
- The grammar is more efficient if it conveys more information with fewer grammatical categories. In view of the discussion in the Introduction of grammar efficiency regarding the linguistic input and output, the information conveyed represents the output, whereas the grammatical categories are the input. This general principle has been used from the outset of Vulkanović's work on grammar efficiency (Vulkanović 1991, 1993, 2003, 2007). Because of this, *AE* is directly proportional to the quotient $|F|/|W|$. The propositional functions represent the information that simple intransitive sentences convey, and word classes convey this information (together with word order).

For the PoS system of Santali, we have

$$Q_{\text{Santali}} = \frac{12}{37} \cdot \frac{3}{14} = \frac{18}{259} = 0.069 \quad (7)$$

and

$$AE_{\text{Santali}} = \frac{18}{259} \cdot \frac{4}{3} = \frac{24}{259} = 0.093.$$

After defining the absolute grammar efficiency, we can define the relative grammar efficiency. We do this in the same way as in (Vulanović 2003, 2007, 2008, to appear). Consider a PoS system that has a certain number of propositional functions, $|F|$, and a certain number of word classes, $|W|$. The measure of relative grammar efficiency, *RE*, is determined within the family of grammars that also have $|F|$ propositional functions and $|W|$ word classes. Let this family of grammars be denoted by $G(|F|, |W|)$. If $|W| \leq |F|$, we find the grammar in

⁶ Gil (2001: 344) also considers free word order as an indication of a less complex grammar.

$G(|F|, |W|)$ that has the greatest value of AE , that is, the greatest value of Q , without admitting any ambiguous sentence. Such a grammar, which we denote by G^* , does not have to exist or be unique. If it does exist, we denote its absolute efficiency and the corresponding Q by AE^* and Q^* . Then, the relative grammar efficiency of the PoS system is defined as

$$RE_{\text{PoS}} = \frac{AE_{\text{PoS}}}{AE^*} = \frac{Q_{\text{PoS}}}{Q^*}, \quad (8)$$

where the subscript PoS indicates the quantities pertaining to the PoS system under consideration. According to this definition, the relative efficiency of G^* itself is $RE^* = 1$, which means that any maximally efficient grammar (with respect to AE), within the family $G(|F|, |W|)$, has relative efficiency scaled to 1. The definition in (8) reduces to $RE_{\text{PoS}} = Q_{\text{PoS}}$ if $|F| = |W|$ because $Q^* = 1$ in this case. On the other hand, if $|W| > |F|$, no grammar in this family can be proclaimed maximally efficient because $RE^* = 1$ can already be achieved with fewer word classes. In this case, we define $RE_{\text{PoS}} = AE_{\text{PoS}}$. We do the same if G^* does not exist, which happens if each grammar in $G(|F|, |W|)$ admits at least one ambiguous sentence.

Returning to the Santali example, we see that we need to find Q^* in the family of grammars with all four propositional functions and three word classes. It is the basic PoS systems that have the greatest values of Q because if a system has two or more word classes fulfilling the same propositional functions, both the s and \hat{s} counts grow, but \hat{s} grows more than s and, therefore, the value of Q becomes less. Table 10 in section 6 shows examples of this. Thus, we evaluate Q for the six basic systems in Table 5 that have $|F| = 4$ and $|W| = 3$. The results are presented in Table 8, where we repeat the values of μ for convenience. We consider all possible orders of propositional functions when both predicate and referential phrases are continuous (therefore, $f = 18$, see Table 7), and find the greatest count of unambiguous sentences that may be permitted. This count is indicated in Table 8 as $\max s$ and the corresponding value of Q is $\max Q$.

We can conclude that for the family $G(4,3)$, $Q^* = 0.533$ (more precisely, $8/15$). Using (7) and (8), we get that

$$RE_{\text{Santali}} = \frac{18}{259} \cdot \frac{15}{8} = 0.130.$$

Table 8

The greatest possible values of Q in the basic PoS systems with $|F| = 4$ and $|W| = 3$

PoS system	μ	$\max s$	$\max Q$
VNMM (type 3)	1.667	16	0.533
V \mathcal{X} \mathcal{X} m	2.000	12	0.333
VX _R aX _R	2.000	17	0.472
\mathcal{P} Na \mathcal{P}	2.167	12	0.308
X _P NX _P m	2.167	17	0.436
HHam	2.500	15	0.333

The measure of relative grammar efficiency enables a better comparison of the efficiency of grammars that belong to different families with respect to the number of propositional functions and the number of word classes.

5. Language sample

Our language sample comes from (Hengeveld et al. 2004) and (Hengeveld & van Lier 2008). Hengeveld et al. (2004) consider 50 languages, of which 12 have flexible PoS systems, whereas the (Hengeveld & van Lier 2008) sample contains 22 languages, 9 with flexible PoS systems. The two samples partly overlap. Since the previous and the new measures of grammar efficiency produce identical results for rigid PoS systems, it is interesting only to compare the measures for flexible PoS systems. There are 17 such languages in the two sources combined. Fourteen of those languages form our sample. We exclude Ngiti and Kayardild because their predicate and referential phrases are not necessarily continuous, see (Hengeveld et al. 2004) for Ngiti and (Evans 1995) for Kayardild. We also exclude Mundari from the (Hengeveld et al. 2004) sample because another language from the same Munda genus, Santali, is present in the (Hengeveld & van Lier 2008) sample, where the PoS systems are generally described with more detail.⁷ Four of the 14 languages (Samoan, Turkish, Imbabura Quechua, and Lango) belong to both samples. Whereas PoS systems of Turkish and Lango are described in the same way in (Hengeveld et al. 2004) and (Hengeveld & van Lier 2008), the descriptions of Samoan and Imbabura Quechua are again more detailed in the latter source and those descriptions we adopt here. As for the basic order of propositional functions, we use the information which is available in (Hengeveld et al. 2004). However, the order of propositional functions is not considered in (Hengeveld & van Lier 2008) and we had to consult additional sources for Kambera, Santali, Ma'di, and Abun, the 4 flexible languages which exclusively belong to the (Hengeveld & van Lier 2008) sample. The sources are identified in Table 9, where the structures of the sample languages are shown. The languages in Table 9 are ordered so that the number of word classes increases and then so that the flexibility of the PoS system decreases. Thirteen of the 14 languages have all four propositional functions. The only exception is Tagalog, which lacks p. Tagalog is also a language in which the basic order of R and r cannot be identified, and this is why both the PRr and PrR orders are considered.

Table 9

The 14-language sample. HRS stands for (Hengeveld et al. 2004) and HvL for (Hengeveld & van Lier 2008). WALs is the World Atlas of Language Structures (Dryer & Haspelmath 2013)

Language	Sources	Word classes	Basic order of propositional functions
Tagalog	HRS	Y _r	PRr, PrR
Samoan	HRS, HvL	V, C	PpRr

⁷ Since Hengeveld et al. (2004) are mainly interested in PoS system *types* as points on a classification scale, they do not consider more detailed accounts of PoS systems.

Hurrian	HRS	V, Λ	rRpP
Warao	HRS	V, Λ	rRpP
Kambera	HvL, WALS, Klamer 1998	V, C, m	pPRr
Santali	HvL, WALS, Cole 1896	V, N, C	rRpP
Turkish	HRS, HvL	V, Λ , M	rRpP
Imbabura Quechua	HRS, HvL	V, Λ , m	rRpP
Ket	HRS	V, N, M	rRpP
Miao	HRS	V, N, M	RrPp
Tidore	HRS	V, N, M	RrPp
Lango	HRS, HvL	V, N, M, m	RrPp
Ma'di	HvL, WALS, Blackings & Fabb 2003	V, \mathcal{N} , a, m, M	RrPp
Abun	HvL, WALS, Berry & Berry 1999	V, N, a, m, M	RrPp

The structures in Table 9 represent idealizations. Although we tried to make those idealizations reasonably close to reality, this is not so crucial here. The main point is that the two formulas for evaluating grammar efficiency are compared when applied to identical structures, regardless of how much those structures are realistic.

6. Results

In this section, we present the values of relative grammar efficiency for each language in our sample. Every one of them has all four propositional functions, thus $|F| = 4$, except for Tagalog which does not have the p function (therefore, for Tagalog, $|F| = 3$). The calculations are carried out like in the Santali example in section 4. The results are given in Table 10 together with the values of all quantities needed in the calculations, except for Q^* , which is discussed below.

In addition to Santali, six other languages in the sample (Kambera, Turkish, Imbabura Quechua, Ket, Miao, and Tidore) use three word classes ($|W| = 3$). Their relative grammar efficiency is calculated using (8) with $Q^* = 8/15$, just like in the case of Santali.

The grammars of the PoS systems of Samoan, Hurrian, and Warao belong to the class $G(4,2)$. The value of Q^* has to be found for this class in the same way it is found for $G(4,3)$ using Table 8 in section 4. It turns out that $V\Lambda\Lambda\Lambda$, as a basic PoS system type, has the most efficient grammar in $G(4,2)$, with $Q^* = 1/6$, when nine possible unambiguous sentences are permitted.⁸

As for Tagalog, the corresponding class of grammars is $G(3,1)$. The grammars of this

⁸ The sentences are $V\Lambda$, ΛV , $V\Lambda\Lambda$, $\Lambda V\Lambda$, $\Lambda\Lambda V$, $V\Lambda\Lambda\Lambda$, $\Lambda V\Lambda\Lambda$, $\Lambda\Lambda V\Lambda$, and $\Lambda\Lambda\Lambda V$.

class can only vary with respect to the permitted order of propositional functions. If the order is fixed, say, to PRr, the greatest value of Q is obtained with both $Y_r Y_r$ and $Y_r Y_r Y_r$ unambiguous sentences, $Q^* = 2/39$. The sentence $Y_r Y_r Y_r$ is ambiguous in Tagalog because of the two permitted orders of propositional functions. This reduces the value of Q in Tagalog to $1/39$.

For all other languages of the sample, there is no need to do any additional work regarding Q^* . Lango uses four word classes and, therefore, the relative grammar efficiency of its PoS system is the same as the corresponding value of Q . Ma'di and Abun have five word classes and in this case, $RE = AE$.

Table 10 also lists the values of relative grammar efficiency RE_o obtained using the previous formula from (Vulanović 2003, 2007). In general, the subscript o refers to the quantities related to the old grammar-efficiency formula. The previous absolute grammar efficiency is defined as in (6) but with

$$Q = Q_o = \frac{s}{a}.$$

Like in (6'), s is the number of unambiguous sentences admitted by the grammar, but a is a quantity which is much more complicated to calculate than m and μ in (6'). It stands for the total number of parsing attempts of all permutations of each sentence in the PoS system. We will not explain this here any further. We refer the reader to (Vulanović 2003, 2007) and to the more recent (Vulanović to appear), where the Turkish PoS system is used to compare the calculations needed by the stage 2) and 3) approaches. That example shows convincingly that the new formula is much simpler. It is even simpler when it comes to the evaluation of relative grammar efficiency. In the old approach, RE_o is defined analogously to RE in section 4, but the search for maximally efficient grammars involves much more tedious computations. Nevertheless, they have been done in (Vulanović 2008) for basic PoS systems and the results from there are used to find the values of RE_o in Table 10 (where the values of a are also presented).

Table 10
The relative grammar efficiency and of the quantities
needed to evaluate it for the 14-language sample

Language	μ	s	\hat{s}	m (if $m \neq \hat{s}$)	a	RE	RE_o
Tagalog	6.5	1	2	6	10	0.500	0.500
Samoan	5.5	4	12	18	85	0.242	0.129
Hurrian	3	2	9	18	34	0.222	0.162
Warao	3	2	9	18	34	0.222	0.162
Kambera	4.333	10	27		158	0.160	0.101
Santali	4.667	12	37		188	0.130	0.102
Turkish	3	7	32		100	0.137	0.112
Imbabura Quechua	2.667	4	17	18	56	0.156	0.114

Ket	1.667	4	16	18	28	0.250	0.229
Miao	1.667	4	16	18	28	0.250	0.229
Tidore	1.667	4	16	18	28	0.250	0.229
Lango	1.75	6	28		44	0.122	0.136
Ma'di	2.75	12	58		119	0.060	0.081
Abun	2	9	48		68	0.075	0.106

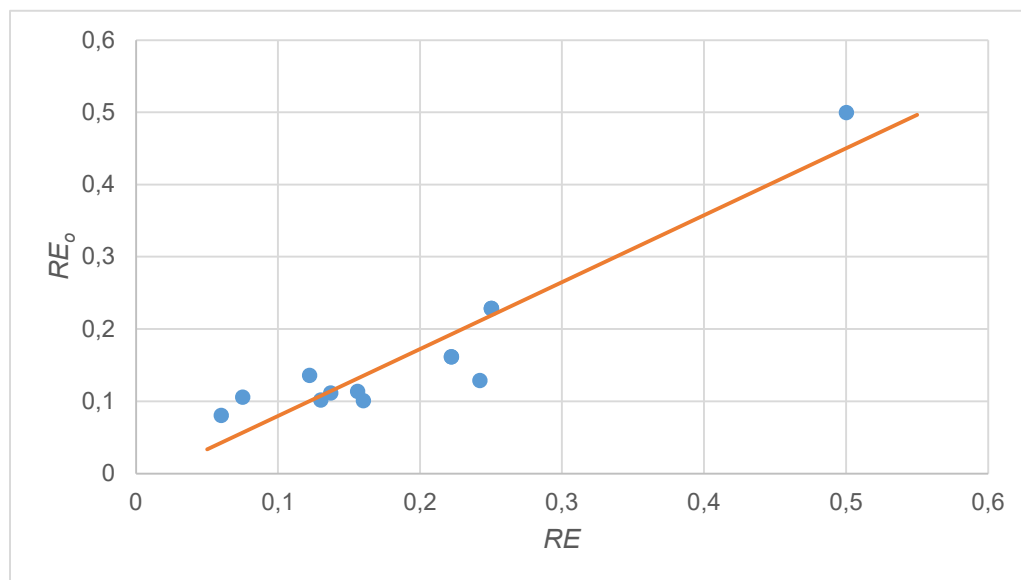


Figure 1 The correlation of RE and RE_o

We end the section by discussing the correlation of the RE and RE_o values. Figure 1 shows the scatter plot with the regression line for the results in the last two columns of Table 10. The correlation between RE and RE_o is strong; the coefficient of correlation is $r = 0.938$.

7. Conclusion

The paper can be summarized as follows. We considered the formula from Vulanović (to appear) for calculating the grammar efficiency of parts-of-speech systems which are described following Hengeveld's (1992) approach. We modified the formula by eliminating some weights from it. Although this modification is not substantial, the new formula is somewhat simpler than the one in Vulanović (to appear) and significantly simpler than the previous grammar-efficiency formula from (Vulanović 2003, 2007). This gives the new formula an advantage over the original one. To fully establish the viability of the new formula, we showed that it produces results that are comparable to those obtained by the old formula. We did this by finding a strong correlation between the grammar-efficiency values calculated by the two formulas when applied to parts-of-speech systems found in a sample containing 14 natural languages.

In conclusion, the replacement of the grammar-efficiency formula from (Vulanović 2003, 2007) with the newly developed one can be justified. The simplicity of the new formula will enable calculations of the grammar efficiency for linguistic structures that are more complicated. For instance, only simple intransitive sentences and continuous propositional and referential phrases have been modeled so far. We can now move onto transitive sentences and discontinuous phrases

References

- Anttila, R.** (1972). *An introduction to historical and comparative linguistics*. New York: Macmillan.
- Blackings, M. & Fabb, N.** (2003). *A grammar of Ma'di*. Berlin/New York: de Gruyter.
- Berry, K. & Berry, Ch.** (1999). *A description of Abun*. Canberra: Research School of Pacific and Asian Studies, The Australian National University.
- Cole, F.T.** (1896). *Santali primer*. Pokhuria, Manbhum: The Santali Mission Press.
- Dryer, M.S. & Haspelmath, M.** (eds.) (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info>, accessed on 2020-01-26).
- Evans, N.D.** (1995). *A grammar of Kyardild*. Berlin/New York: Mouton de Gruyter.
- Gil, D.** (2001). Creoles, complexity, and Riau Indonesian. *Linguistic Typology* 5, 325-371.
- Hawkins, J.A.** (2004). *Efficiency and complexity in grammars*. Oxford/New York: Oxford University Press.
- Hawkins, J.A.** (2014). *Cross-linguistic variation and efficiency*. Oxford/New York: Oxford University Press.
- Hengeveld, K.** (1992). Parts of speech. In: Fortescue, M., Harder, P. & Kristoffersen L. (eds.), *Layered structure and reference in functional perspective: 29-55*. Amsterdam/Philadelphia: John Benjamins.
- Hengeveld, K., Rijkhoff, J. & Siewierska, A.** (2004). Parts-of-speech systems and word order. *Journal of Linguistics* 40, 527-570.
- Hengeveld, K. & Rijkhoff, J.** (2005). Mundari as a flexible language. *Linguistic Typology* 9, 406-431.
- Hengeveld, K. & van Lier, E.** (2008). Parts of speech and dependent clauses in Functional Discourse Grammar. *Studies in Language* 32, 753-785.
- Hengeveld, K. & van Lier, E.** (2010). An implicational map of parts of speech. *Linguistic Discovery* 8, 129-156.
- Hengeveld, K. & Leufkens, St.** (2018). Transparent and non-transparent languages. *Folia Linguistica* 52, 139-175.
- Klamer, M.** (1998). *A grammar of Kambara*. Berlin, New York: Mouton de Gruyter.
- McWhorter, J.H.** (2001). The world's simplest grammars are creole grammars. *Linguistic Typology* 5, 125-166.
- Miestamo, M., Sinnemäki, K. & Karlsson, F.** (eds.) (2008). *Language complexity: Typology, contact, change*. Amsterdam: Benjamins.
- Miestamo, M.** (2008). Grammatical complexity in a cross-linguistic perspective. In: Miestamo, M., Sinnemäki, K. & Karlsson, F. (eds.), *Language complexity: Typology, contact, change: 23-41*. Amsterdam: Benjamins.
- Newmeyer, F.J. & Preston, L.B.** (eds.) (2014). *Measuring grammatical complexity*. Oxford/New York: Oxford University Press.
- Vulanović, R.** (1991). On measuring grammar efficiency and redundancy. *Linguistic Analysis* 21, 201-211.
- Vulanović, R.** (1993). Word order and grammar efficiency. *Theoretical Linguistics* 19, 201-222.

- Vulanović, R.** (2003). Grammar efficiency and complexity. *Grammars* 6, 127-144.
- Vulanović, R.** (2007). On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics* 20, 399-427.
- Vulanović, R.** (2008). A mathematical analysis of parts-of-speech systems. *Glottometrics* 17, 51-65.
- Vulanović, R. & Ruff, O.** (2018). Measuring the degree of violation of the One-Meaning–One-Form Principle. In: Wang, L., Köhler, R. & Tuzzi, R. (eds.), *Structure, function and process in texts*: 67-77. Lüdenscheid: RAM.
- Vulanović, R.** (to appear). Grammar efficiency and the One-Meaning–One-Form Principle.

Intrinsic Intentionality and Linguistic Meaning: An Historical Outline

Hermann Moisl¹

Abstract

A long-standing problem in linguistics and cognitive science more generally is how natural language expressions come to possess, and how artificially intelligent systems can be endowed with, intrinsic meaning. There is a long tradition in Western thought whereby the meanings of linguistic expressions are their significations of mental concepts, and concepts are representations of the mind-external environment causally generated by the cognitive agent's interaction with that environment. This paper outlines this tradition with the aim of providing the intellectual context in which cognitive models with intrinsic meaning can be constructed.

Keywords: Cognition; intentionality; concepts; linguistic meaning; Chinese Room

1. Introduction

A long-standing problem in linguistics and cognitive science more generally is how natural language expressions come to possess, and how artificially intelligent systems can be endowed with, intrinsic meaning. There is a long tradition in Western thought whereby the meanings of linguistic expressions are their significations of mental concepts, and concepts are representations of the mind-external environment causally generated by the cognitive agent's interaction with that environment. This paper outlines this tradition with the aim of providing the intellectual context in which cognitive models with intrinsic meaning can be constructed.

The discussion is in two main parts: the first motivates engagement with this topic, and the second presents the historical outline.

2. Motivation

In the second half of the twentieth century, cognitive science in general and linguistics in particular were dominated by the Computational Theory of Mind (CTM), whereby the mind is seen as a Turing Machine whose program is cognition (Rescorla 2020). Its dominance in recent decades has been challenged by neural (Churchland 2012) and dynamical systems (Metzger 2017) approaches to cognitive theory, but with respect to meaning in particular the most fundamental challenge has come from a philosophical thought-experiment formulated by the philosopher John Searle (1980) and subsequently developed by him (Searle 1984, 1989, 1990, 2002, 2010). Searle's initial aim was to counter the claim by 'strong' artificial intelligence that it is possible to construct machines with human-level intelligence by basing their design on CTM, but he subsequently extended the implications of the experiment to the philosophical foundations of CTM itself, arguing that it cannot in principle offer a complete theory of cognition. This attack on *de facto* standard cognitive science and its application in AI precipitated an extensive controversy which continues to the present day, but to date there is no

¹ hermann.moisl49@gmail.com.

consensus either on the validity of his position or, assuming that it is valid, whether or how AI and CTM might adapt to it (Cole 2020).

Searle's argument (1980) is based on two propositions:

- (1) 'Intentionality in human beings (and animals) is a product of causal features of the brain'.
- (2) 'Instantiating a computer program is never by itself a sufficient condition of intentionality'.

He takes (1) as 'an empirical fact about the actual causal relations between mental processes and brains. It says simply that certain brain processes are sufficient for intentionality', and (2) as something to be established by argument, which he undertakes. The conclusions are that 'the explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program', which is 'a strict logical consequence' of (1) and (2), and that 'any mechanism capable of producing intentionality must have causal powers equal to those of the brain. This is meant to be a trivial consequence of 1. Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain'.

These conclusions assume the validity of (2), the argument for which is based on the Chinese Room thought experiment. There is a closed room containing Searle and a list of rules in English for manipulating Chinese orthographic symbols. Chinese speakers outside the room put sequences of these symbols into the room and, using the rules available to him, Searle assembles and outputs sequences of Chinese symbols in response. The people outside interpret the input sequences as sentences in Chinese whose meaning they understand and the output sequences as reasonable responses to them, and on the basis of the room's conceptually coherent input-output behaviour conclude that it understands Chinese. Searle himself, however, knows that the room does not understand Chinese because he, the interpreter and constructor of the sequences, does not understand Chinese, but is only following instructions without knowing what the input and output strings mean.

Key to interpretation of the experiment is the philosophical concept of intentionality (Jacob 2019). Etymologically it is related to Latin *intendere*, 'to point at, to direct', and was used in medieval European philosophy to refer to the mind's ability to direct its attention to specific mental concepts as well as to things and states of affairs in the mind-external world. In present-day philosophy of mind *intentionality* is used to denote the 'aboutness' of mental states, '*the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs*' (Jacob 2019). Searle distinguishes two types of intentionality, original and derived, where the locus of original intentionality is the human head, and derived intentionality is that which we attribute to physical mechanisms which we have good reason to believe do not have original intentionality, such as a thermostat, whose operation is routinely interpreted by humans as 'wanting' to maintain an even temperature.

The Chinese Room is, of course, a computer. Searle is the CPU, the list of English instructions is a program, and the input / output sequences are symbol strings; by concluding that the room understands Chinese, its observers have confirmed the Turing Test (Oppy & Dowe 2016), which says that any device which can by its observable behaviour convince human observers that it has human-level cognition, that is, intentionality, must be considered to possess it. Searle knows, however, that the room's intentionality is derived, not original, the implication being that physical computer implementations of CTM models in AI systems, like thermostats, can only ever have derived intentionality. The intentionality of the symbols and symbol structures manipulated by the algorithm of a CTM model are in the heads and only in the heads of their human designers. When physically instantiated, for example by compilation of a

program onto a physical computer, this intentionality is lost: the symbols and symbol structures cease to be symbolic and become physical tokens which drive the physical causal dynamics of the machine, but intentionality is not a factor in that causal structure. If well designed, the behaviour of the machine with respect to whatever aspect of cognition the CTM algorithm was intended to model can be interpreted as intentional, just as the behaviour of the Chinese room can be so interpreted, but once again the semantics is derived because the only locus of intentionality is in the heads of observers. Put simply, a physical computer does not in any sense understand what it is doing any more than a vending machine does; it simply pushes physical tokens around, and humans interpret that activity as intentional.

Given the lack of consensus on Searle's position or prospects of one, an alternative approach to *a priori* philosophical discussion of it is empirical: assume that Searle is right about the original / derived intentionality distinction and about the view that a physical system can only have original intentionality if intentionality is a causal factor in its operation, and construct models based on those assumptions to see if any useful scientific insights ensue. The present discussion takes a first step in that direction by reviewing the history underlying these assumptions.

3. Historical outline

The above-mentioned tradition has a two and a half millennium-long history that begins with the pre-Socratic philosophers of ancient Greece and finds its current expression in the various academic disciplines that comprise cognitive science. A thorough review of this body of thought would require engagement with truly vast associated literatures, and would amount to a history of a large part of Western thought. Even assuming the requisite authorial competence, the sheer size of such a review would clearly rule it out here. What follows is therefore a sketch, the aim being to provide a conceptual context for model construction. To forestall a bibliographical deluge, the policy is to cite recent work containing further references.

3.1. Antiquity (c.600 BC - c.500 AD)

Philosophical interest in the nature of linguistic meaning is documented in the Ancient Greek cultural world from the 6th century BC onwards (Allan 2009). The framework for discussion was predominantly psychological in that language was analyzed as an aspect of mind. The most important figures were Aristotle (384–322 BCE) , Augustine (354–430 CE), and Boethius (c.480–524 CE).

Aristotle's views on meaning in language have been hugely influential in Western philosophy (Cohen 2016; Modrak 2001; Shields 2012, 2014, 2015, 2016). They can only be fully understood in the context of complex ontology, epistemology, and philosophy of mind developed across the corpus of his work, but the essence for present purposes is this. The mind has two main components: the perceptual faculty which gives it access to the external world via the senses, and the noetic faculty which synthesizes and interprets the output of the perceptual one. The perceptual faculty is innate and includes both sensation and imagination, where the former, as its name indicates, deals with the senses and their role in providing input from the mind-external world, and the latter with synthesis of sensory input. His epistemology is based on the assumption that the structure of independently-existing mind-external physical reality is accessible to the perceptual faculty, and his account of perception is framed in terms of his concept ofhylomorphism, 'matter-formism', whereby a natural substance comprises matter and form such that the form is what differentiates the matter of any one thing from any other. The form but not the matter of a mind-external object enters the perceptual faculty via the senses, where it is represented as an image; Aristotle's analogy is a signet ring pressed into hot wax,

which leaves the form of the design on the the ring's face in the wax, but not the gold from which the ring is made. The perceptual faculty not only represents mind-external objects in this way, but also abstracts over formally-similar representations using the 'active intellect' innate in all humans, so that, for some collection of physically similar forms in external reality, it generates a representation of the essential form, or universal, underlying the variant physical instantiations; such a representation is an *eikon*, an image that represents not particulars of an object in reality, but an abstraction of a class of similar objects. And, finally, Aristotle is clear that words signify the mental representations generated by the perceptual faculty and that this signification is conventional in the sense that it is based on societal consensus.

Augustine (Tornau 2019; Meier-Oeser 2011; Klima 2017) proposed a theory of signification in which a sign is “*something that shows itself to the senses and something other than itself to the mind*”, and can be one of two kinds, conventional and natural, the latter of which '*apart from any intention or desire of using them as signs, do yet lead to the knowledge of something else*', that is, things which by their nature signify without requiring conventional agreement, such as footprints signifying the passing of an animal (quotations from Meier-Oeser 2011). With reference to language, spoken words in any given language are conventional signs of the speaker's mental concepts which the speaker uses to convey those concepts to other minds. Concepts are generated by a combination of sensory experience of the mind-external world and divine illumination, and, following Aristotle, are natural signs of things and events in the world - there is a causal, nonconventional connection between states of the world and how they are represented in the mind, and consequently a resemblance between the structure of mind-external reality and mind-internal conceptual structure. Words in the various spoken languages conventionally signify their counterparts in the mental language which, also following Aristotle, is common to all humans in the sense that humans all experience the world in closely similar ways because we share the same kinds of mind and sensory apparatus as part of human nature.

Boethius (Klima 2017; Marenbon 2016; Meier-Oeser 2011) was the main conduit of Aristotelian philosophy to medieval Europe. His main personal contribution lies in his attempt to reconcile the empirical Aristotelian view of universals with that of Augustine, whose view was strongly influenced by the neoplatonic tradition of abstract Forms or Ideas existing independently of human mentality. The problem had to do with the semantics of so-called common names. Proper names and indexicals can refer directly to particular entities in the physical world (Ashworth 2015), but what are the referents of common terms that refer to things which have no physical existence in the world - not only words for obvious abstractions like 'truth', but also ones which pervade natural language and denote types of real-world object such as 'man'? Boethius took an Aristotelian empiricist view, whereby universals are mental concepts derived by abstraction from sensory experience, and cites with approval the view of Alexander of Aphrodisias (*floruit* c.200 CE) that spoken words do not signify things in the world but our mental concepts of such things, and that universals have no independent existence in the mind-external world but are mental abstractions of sensory experience generated by the Aristotelian active intellect: spoken words conventionally signify concepts, and concepts nonconventionally signify states of the world.

3.2. The European Middle Ages (c.500 - c.1500 AD)

There is nothing of importance for present purposes in the period from Boethius until after 1000 AD. This changed thereafter for two main reasons. One was the growth in academic activity in newly-established European universities. The other was increasing availability of the works of Aristotle via contact with the Islamic world (Shields 2015; Spade 2016). In combination, these developments revolutionized philosophical discourse and led to a substantial amount of work on the theory of mind and language (Allan 2009; King 2007; Meier-Oeser 2011; Spade 2016).

Anselm of Canterbury (1033-1109) combined Aristotelian and Augustinian views on concepts and linguistic meaning in mental words, which have natural signification and are '*resemblances and images of things*' (Meier-Oeser 2011). It was, however, his younger contemporary Peter Abelard (1079–1142 CE) who pioneered subsequent developments in medieval philosophy of mind and language (Allan 2009; King 2007; King & Arlig 2010). His ideas were based on a sparse ontology that was strongly Aristotelian and therefore physical monist and reductive, whereby the only real things in the mind-external world are physical particulars. Following Aristotle, thought is based on concepts which are derived from sensory experience of the world and which represent that experience in the mind. These concepts are nonconventionally related to the world in that there is a causal connection between states of the world and how they are represented in the mind, and they are common to all humans in that, relative to some given state, all humans represent that state in similar ways because we share the same kind of mind and sensory apparatus as part of human nature. Experientially-derived concepts are the basis for the mind's creation of abstract concepts which have no existence in mind-external reality. In arguing for this view of concepts Abelard had to resolve a problem transmitted to him by Aristotle via Boethius and deriving from his ontology: given that, to have meaning, words must refer ultimately to particulars in the world, what is the reference of a so-called common name, that is, of universal words like 'man'? There are particular men in the world, but no universal man. Abelard's solution was twofold. Firstly, he proposed that common names do not signify universals but rather are universals themselves - universals are words, not things that words signify (King 2007; King & Arlig 2010; Klima 2017; Rodriguez-Pereyra 2014). And secondly, anticipating Frege, he proposed a binary decomposition of Augustinian signification into sense (*significatio*) and reference (*nominatio*). The *nominatio* of a so-called proper name like 'Aristotle' is a reference to a particular man in the world, and of a common name like 'man' is the particulars to which the word applies by distributive reference, or, in Aristotle's formulation, the word is 'predicable of many'; the *significatio* of a proper name is a sense-derived concept, and of a common name an abstract concept. The meaning of a word is thereby a combination of its *nominatio* and its *significatio*. Underlying all this, finally, is the traditional principle that both the sense and the reference of a word in spoken language are conventional.

Abelard initiated an extensive discussion of semiotic and specifically linguistic semantic topics based on a complex interplay of Aristotelian and Augustinian thought which lasted until the end of the Middle Ages and beyond (Allan 2009; Meier-Oeser 2011; Shields 2015; Spade 2007, 2016). There are too many important figures in that discussion to deal with individually, so topics of particular relevance to present purposes are outlined instead.

- Universals

The ontological status of universals fell into three broad categories derived from Antiquity: that they are *universalia ante rem*, 'before the thing', that is, as independently existing Platonic objects, or *universalia in re*, 'in the thing', as somehow implicit in particulars, or *universalia post rem*, 'after the thing', as concepts without existence independent of the mind's abstraction over particulars (Klima 2017). The main disagreement was *in re* versus *post rem*, or in present-day terms between realists and nominalists respectively; the *universalia post rem* position had become dominant by the 14th century with the resurgence of the nominalism pioneered by Abelard and associated primarily with William of Ockham (c. 1287-1347; Meier-Oeser 2011; Spade 2016), whereby the only universals are words that signify concepts, and concepts are mental representations that signify particulars in the physical world.

- Signification

Signification was fundamental in Abelard's philosophy of language and remained so in medieval semiotics and linguistic semantics (Meier-Oeser 2011; Spade 2007), as did the

distinction, derived from Aristotle and Augustine, of natural and conventional signs, where natural signs are causally generated in the mind by perception of states of the world. There is a three-level hierarchy of signification for words (Spade 2007): a written word signifies the corresponding spoken word, a spoken word signifies the corresponding mental concept, and a mental concept signifies a mind-external object or state of affairs in the natural world.

- Mental representation

Aristotle's view of mental representation was that representation of the world external to the cognitive agent was essentially pictorial. It was not until the 13th century that the idea of mental imagery was questioned by, for example, Aquinas (1225-1274), who used the metaphor of a blueprint for a building to argue for a more abstract notion of mental encoding rather than direct resemblance (Lagerlund 2017; Spade & Panaccio 2019).

3.3. The modern era (c.1500 - present)

The medieval 'scholastic' tradition was gradually supplanted by the rise of empirical science in the course of the 16th, 17th, and 18th centuries (Allan 2009; Gendler Szabo 1998; Klima 2017; Meier-Oeser 2011), in which thinking on language was strongly influenced by this new scientific spirit. An early example was the *Port Royal Grammar* of 1660 (Buroker 2014; Gendler Szabo 1998), whose fundamentals were that (i) thought is prior to language, (ii) proper names signify ideas of particulars, and common names signify ideas of universals, and (iii) The relationship between linguistic expressions and ideas is conventional, but between ideas and what they represent in the mind-external world is not. An approximate contemporary was John Locke (1632 - 1704; Allan 2009; Gendler Szabo 1998; Uzgalis 2018), who famously maintained that the mind is a blank slate at birth on which experience writes knowledge of the mind-external world. Experience is of two kinds, sensation and reflection, where the first is information about the world provided by the senses, and reflection is a mental process that builds new, more complex ideas from existing ones, both sensory and other products of reflection; the mental process is a set of innate 'faculties' whereby ideas are structured. Proper names signify ideas based on perception, and common names signify universal ideas constructed by the mind's 'faculties'.

The period from the end of the eighteenth century onward has seen rapid development of empirically based science and technology, and developments in the study of language have mirrored that empirical orientation. John Stuart Mill, for example (1806-73; Macleod 2016), regarded the mind as part of the causal order of physical nature amenable to study by the empirical methods of the natural sciences rather than as a metaphysical entity. Understanding of the world comes solely from inductive inference from sensory experience: sense impressions frequently experienced simultaneously or in immediate succession are associated to become ideas and causal relations among ideas respectively. This process is iterative, so that as experience of the world grows over time, the ideational and causal structure of the mind increasingly comes to resemble the corresponding structure inherent in mind-external reality. From these fundamentals the mind constructs complex ideas and relations all of which, however abstract, are ultimately grounded in experience; language expresses these complex ideas and relations. Another example is Charles Peirce (1839-1914; Atkin 2010), who developed a theory of signification, representation, reference, and meaning which became the foundation of the present-day discipline of semiotics. The fundamental idea is that a sign has three components: a signifier, that which is signified, and an interpreter who connects the two. Signs can be of three kinds: (i) an icon is a sign whose signifier bears a resemblance to what it signifies, like a painting of a person, (ii) an index is a sign whose signifier is causally connected to what is signified, like smoke signifying fire, and (iii) a symbol is a signifier whose connection with what it signifies is conventional.

The 19th century saw the first attempts to understand the interrelationship of mind, brain

and of sensory systems from a scientific rather than purely philosophical point of view, as seen in the work of, for example, E. H. Weber (1795–1878), Gustav Fechner (1801–1887), Hermann von Helmholtz (1821–1894), Ernst Mach (1838–1916), and Wilhelm Wundt (1832–1920). Of these, von Helmholtz and Mach are of particular relevance to present concerns on account of their views on the genesis of perceptually-based mental representations.

- Von Helmholtz (Patton 2018) was a physicist with a strong interest in perception as the basis for scientific epistemology, and was a pioneer of the view that epistemology is crucially dependent on the dynamic interaction of body and environment. The essence of his view was that percepts were 'signs' which symbolize sensory stimuli, but that, unlike in earlier so-called picture theories of perception, there is no necessary direct resemblance between the form of the stimuli and their perceptual representation. Physical stimuli occur in temporal sequence, and the mind makes 'unconscious inferences' from the sequence of percepts to construct a temporal ordering and thereby a coherent mental representation of experience whose structure reflects regularities in the physical stimuli on which it is ultimately based. Perception of space and sequence is determined by constraints that the structure of the perceiver's body places on interaction with the physical environment: '*perceptual space is a mental generalization of our orientation with respect to objects in space*' (Patton 2018).
- Ernst Mach (Pojman 2019) was a physicist and mathematician with a conviction of the need to take the body into account when studying perception. On his account, sensory systems are dynamical systems tending to equilibrium with the environment by a continual process of adaptation to physical stimuli over time; what the senses thereby make available to the brain are not direct representations of reality but generalizations over experiential history, with the result that there is no necessary isomorphism between percept and reality, or, in other words, human access to knowledge of mind-external reality is indirect. This tendency to equilibrium, moreover, is teleological in the Darwinian sense of fitness for an environment leading to optimization of survival chances. Mach's empiricism was not of the Lockean 'blank slate' variety, but had the apparently-paradoxical effect of allowing him to admit a priori structure to his account of perception in that, like evolutionary processes in nature more generally, the human sensory apparatus had evolved and continues to evolve towards equilibrium with a structured mind-external reality.

Since the mid-20th century the study of intentionality and of linguistic meaning specifically have been addressed by a range of disciplines, chief among them linguistics and philosophy of language, philosophy of mind, cognitive psychology, and cognitive neuroscience.

3.3.1. Linguistics

Linguistics in the second half of the 20th century was, at least in the USA and the UK, dominated by the generative framework initiated and thereafter guided by Chomsky (Newmeyer 1995). It has been and continues to be strongly orientated towards study of syntax; semantics consists essentially of application of ideas from the philosophy of language (Lycan 2018), which has in turn developed the tradition of truth-conditional semantics initiated by Frege. This approach regards language, and with it meaning in language, as an abstract object to be understood using mathematical logic independently of its instantiation in human minds; natural language is regarded as one instance of a universe of possible languages to the understanding of which general semantic principles such as reference and compositionality can be applied. Because of its apsychological orientation it is of peripheral interest for present concerns. There are exceptions to the dominance of truth conditional semantics in linguistics, however.

- Generative semantics (McCawley 1976, 1995) was an alternative to the standard generative linguistics view of the relationship between syntax and semantics developed in the late 1960s to the mid-1970s. It proposed that, contrary to the standard generativist position, deep structures were semantic rather than syntactic, which transformational rules converted into surface strings and associated syntactic structures. The semantic subcomponent thereby made meaning primary in sentence generation.
- Conceptual Semantics (Jackendoff 2002, 2015) emphasizes the relationship between language and other aspects of cognition, which among other things involves a demotion of syntax from the centrality accorded it in generative linguistics to a subsidiary role. In Jackendoff's words, '*the central hypothesis of Conceptual Semantics is that there is a level of mental representation, Conceptual Structure, which instantiates sentence meanings and serves as the formal basis for inference and for connection with world knowledge and perception*'. Conceptual structure interfaces with two subsystems: a perception / action subsystem that generates representations of the mind-external world, and a linguistic input / output one that generates phonological and syntactic representations of acoustic input and of acoustic output that a listener with the same cognitive system can interpret as meaningful. Reference is a relation between a mental representation in the linguistic input / output subsystem and some part of conceptual structure: linguistic expressions do not refer directly to things in the world but indirectly via mental representations in conceptual structure wholly or partly generated by interaction of conceptual structure with the perception / action subsystem.
- Cognitive linguistics (Croft & Cruse 2004; Geeraerts & Cuyckens 2010) grew out of the generative semantics of the 1970s, and is now an umbrella term for a wide range of approaches. It resembles Conceptual Semantics in locating meaning in general cognition rather than in the linguistic system specifically, and in stressing the mapping from cognitive meaning to linguistic structure as the primary focus of linguistic theory. The currently dominant variety of this approach to language is Cognitive Grammar (CG) (Croft & Cruse 2004; Langacker 2008; Talmy 2000). Geeraerts & Cuyckens (2007) identify three fundamental tenets of CG: (i) meaning is primary in linguistic analysis, (ii) meaning draws selectively on the totality of the individual speaker's world knowledge, and (iii) linguistic utterances communicate not objective truths about the world, but rather the speaker's perspective on it.

3.3.2. Philosophy of mind and cognitive psychology

Study of the psychological mechanisms of intentionality has largely been the preserve of philosophy of mind and cognitive psychology. Topics of particular relevance here are perception, representation, and cognitive architecture.

3.3.2.1. Perception

Perception is the cognitive process by which humans acquire knowledge of the mind-external world (Crane & French 2015; Lyons 2016; Siegel 2016). A central problem has been skepticism (Comesana 2019; McKinsey 2018), which is the view that we cannot have knowledge, that is, true justified belief, of anything about the mind-external world on account of the unreliability of perception. This remains a problem, but, setting it aside, proposals for how we acquire knowledge of the world have been on a continuum between the rationalistic view that such knowledge is innate, where the knowledge intrinsic to the mind determines our perception of the world and thereby creates our reality, and the empiricist view that knowledge of the world

is derived solely from sensory experience of a mind-independent physical reality (Adams & Aizawa 2017; Margolis & Laurence 2019; Markie 2017; Pitt 2012; Samet 2019; Samet & Zaitchik 2017). Few philosophers and psychologists have been pure innatists or pure empiricists (Griffiths 2009; Samet & Zaitchik 2017). Instead, most have taken intermediate positions: on the one hand, innatists have found it necessary to accommodate the self-evident fact that we use our senses to perceive the world, and have proposed sensory perception as a trigger which brings the innate knowledge to consciousness, and, on the other, empiricists have found it necessary to posit an innate mental capacity for induction of knowledge from sensory experience. Most recent and contemporary cognitive scientists subscribe to some combination of innate and environmentally-derived empirical factors as the bases for the acquisition of world knowledge (for example Barsalou 2016a; Griffiths 2009; Lau & Deutsch 2014; Margolis & Laurence 2019; Petersson & Hagoort 2012; Samet & Zaitchik 2017; Shea 2018; Wilson & Foglia 2015).

The nature of this interaction has in recent decades been developed by a variety of research programmes including naturalistic (Rysiew 2016), evolutionary (Bradie & Harms 2016), and teleological epistemology (Neander 2012, 2017; Shea 2018), evolutionary psychology (Buss 2007; Downes 2014), and embodied / situated / grounded cognition (Barsalou 2008, 2010, 2016a; Clark 2008; Lakoff & Johnson 1999; Matheson & Barsalou 2018; Wilson & Foglia 2015; Yeh & Barsalou 2006;). Though differing in focus, the core ideas of these programmes overlap substantially. The guiding principle is that philosophical discussion of knowledge acquisition should work closely with relevant work in the natural sciences rather than relying exclusively on the traditional *a priori* philosophical method. In particular, they have adopted the adaptationist position in the biological sciences (Orzack & Forber 2010), whereby natural selection is an important and probably the primary causal factor in the adaptation of living organisms to fit their specific environmental niches: human cognition is explained in terms of interaction between evolutionarily-generated cognitive mechanisms, physical environment, and individual learning of culturally-transmitted behaviour and knowledge (Barsalou 2016a; Shea 2018). The traditional opposition of innate versus empirical as factors in the acquisition of world knowledge is replaced by a dynamical process which unifies them whereby, by adaptation over evolutionary time, the physical and cultural environment shapes cognitive structure, which in turn interprets and acts on the environment.

3.3.2.2. Representation

Like Descartes, my individual consciousness tells me that I have a mind, that my mind is constituted by thoughts, and that most of these thoughts are about my relationship with the mind-external world. Intuitively, and in the history of philosophy of mind specifically, various terms have been used to designate the components of thought, such as 'phantasms', 'ideas', 'impressions', 'notions', and 'concepts'. In everyday usage the definitions of these terms and their interrelationships are vague, and in the academic literature writers typically use them in different ways at different times and places, the result of which is terminological confusion. The present discussion follows Margolis & Laurence (2019) in standardizing on 'concepts' as *'the building blocks of thoughts'*.

The historical view of concepts as pictures in the mind is long-obsolete, and the dominant view in present-day cognitive science, articulated in the Representational Theory of Mind (RTM; Pitt 2012), is that concepts are mental representations. Because of their central importance in cognitive science there is an extensive literature on the nature of these representations, reviewed for example in (Adams & Aizawa 2019; Carey 2009; Margolis & Laurence 2015, 2019; Pinker 2007; Pitt 2012), which precludes a comprehensive summary here. Instead, what follows identifies some mainstream features and then goes on to describe one particular class of theories: the causal view of mental representation.

In semiotic terms, the move from mental pictures to representations is from icon to symbol. Representations are symbols, and as such there is no necessary formal resemblance between a representation and what it represents. Representations are physically individuated tokens in the heads of cognitive agents, and each token has a semantic interpretation, where the referent or 'content' is a state of the world or another head-internal representation. In this way, representations connect the agent with the external world, and this connection is the foundation on which intentionality is constructed in the agent's mind. Representations are typically regarded as structured mental entities of arbitrary complexity, compositionally constructed from primitives which are either innate or empirically inferred via perception or some combination of the two. An influential view, proposed at various times over the millennia and reintroduced most recently by Fodor (1975), is that the mind's representational system is a language - a language of thought - in which primitive representations are analogous to words in natural language, whose forms are independent of the forms of what they represent and which can be compositionally assembled into complex expressions.

Of particular relevance to the main theme of the present discussion are causal theories of mental content, which *'attempt to explain how thoughts can be about things...These theories begin with the idea that there are mental representations and that thoughts are meaningful in virtue of a causal connection between a mental representation and some part of the world that is represented. In other words, the point of departure for these theories is that thoughts of dogs are about dogs because dogs cause the mental representations of dogs'* ; the general principle is that any given symbol 'X' means X because 'X's are caused by Xs (Adams & Aizawa 2017). Proposed causal factors include cognitive development via the individual cognitive agent's interaction with a structured physical environment under so-called normal conditions, and the development of cognitive functions by natural selection in the human genome consequent on such experience. Stampe (1977) suggested that this causal connectedness implies a homomorphism between mental representational and mind-external physical environmental structure, and subsequent work (or example Adams & Aizawa 2017; Piccinini 2018; Piccinini & Scarantino 2011; Rupert 2008; Shagrir 2012; Thomson & Piccinini 2018) has extended this idea by distinguishing natural and non-natural information and arguing that it enables non-derived or, in Searle's terms, original intentionality in physical systems, including the brain.

3.3.2.3. Cognitive architecture

The dominant view since the second half of the twentieth century has been that the mind is a computer. This is an ontological claim, not just a metaphor: the mind actually is a computer, where 'computer' is understood as a Turing Machine, a mechanism for algorithmic string transformation. The essence of this view of mental architecture, known as the Computational Theory of Mind (CTM) (Boone & Piccinini 2016; Piccinini 2007, 2016; Rescorla 2020), is this:

- The brain implements a Turing computational architecture.
- The mind is an algorithm running on that architecture.
- Concepts are data structures of arbitrary degrees of complexity manipulated by the mental algorithm.
- The primitive data types from which the conceptual data structures are built are representations of mind-external reality.

In its contemporary default form, CTM is closely associated with the Representational Theory of Mind outlined above. When combined with RTM, CTM sees the mind is a computer that transforms strings of primitive representations: the mental program interprets such strings as having compositional syntactic structure of arbitrary complexity, and the nature of the

transformation in any particular case is sensitive to this structure. Fodor famously referred to CTM as 'the only game in town' (Fodor 1975), and it's is easy to see why. CTM is a physicalist theory, and hence one compatible with mainstream science, that explains how a physical system such as a computer or, in humans, the brain, can generate semantically coherent cognitive behaviour from the causal structure of a physical symbol system; for the classic defense of CTM see (Fodor & Pylyshyn 1988).

Since the mid-1980s the dominance of CTM has been challenged by two alternative approaches to modelling the mind: cognitive neuroscience and dynamical systems theory. The first of these is discussed in the next subsection. The second is based on the observation that nonlinear response to input and output latency in individual neurons together with pervasive feedback connectivity among biological neurons and neural assemblies make the brain a physical nonlinear dynamical system capable of behaviours characteristic of such systems, ranging from fixed-point through periodic and fractal to chaotic. The mathematical theory of dynamical and complex systems has consequently been proposed as an alternative or at least an adjunct to CTM (Churchland 2012; Metzger 2017; Port & van Gelder 1995; Ward 2001).

3.3.3. Cognitive neuroscience

The rapid development of cognitive neuroscience and neural modelling in recent decades has been providing ever more detailed insight into the mechanisms of the brain's implementation of traditional ideas about natural meaning. Cognitive neuroscience attempts to identify correlations and, ideally, causal relationships between the various aspects of cognition and the anatomy and dynamics of the brain in the cognitive agent's interaction with the environment (Boone & Piccinini 2016; Gazzaniga et al 2019; Piccinini & Bahar 2013; Piccinini & Shagrir 2014). Empirical results come, on the one hand, from anatomical study and from observation of brain activity correlated with cognitive activity using the various monitoring technologies, and on the other from neural modelling (Buckner & Garson 2019; Schmidhuber 2015) using simplified artificial neural network (ANN) models of neural structure and dynamics to study the behaviour of different architectures relative to cognitive function. These approaches are complementary and often combined.

Because CTM explains cognition in terms of computation over representations, cognitive neuroscience has been particularly interested in implementation of representations in the brain (Barsalou 2016b; Boone & Piccinini 2016; Wilson-Mendenhall et al 2013). An emerging view is that representations are dynamic neural activation patterns distributed over disparate areas of the brain which are proximately or ultimately based on structures and processes in the brain's sensimotor areas generated by interaction with the environment (Barsalou 2017; Conway & Pisoni 2008; Harris et al 2001; Prinz 2002), and that a hierarchy of association areas or 'convergence zones' integrates sensimotor activations from the various modalities and other association areas to generate increasingly abstract representations (Anderson 2010; Barsalou 2016a,b, 2017; Binder 2016; Binder et al 2005, 2009, 2016; Wilson-Mendenhall et al 2013). There is, moreover, evidence for homomorphism between the structural relations among neural states and the states of the external environment they represent, which arises from a causal interaction between cognitive agent and environment and which constitutes a model of the environment (Matheson & Barsalou 2018; Neander 2017; Piccinini & Bahar 2013; Piccinini 2009, 2015, 2018). Such association areas provide a plausible implementation mechanism for the age-old problem of how sensory input is integrated in the mind so as to generate abstract concepts or, in medieval terms, universals. Barsalou & Dutriaux (2018), for example, propose the Situated Conceptualization Framework to explain how such itegration works in the head. Representations grounded in temporally co-occurring sensory and enactive experiences of the world are associated in the convergence zones, thereby generating concepts which can in turn be associated to generate increasingly abstract concepts; the conceptual structures so generated

are stored in memory and interact with the cognitive agent's subsequent experience of the world, thereby dynamically constructing the agent's conceptualization of the world throughout life.

Neurolinguistics, or alternatively biolinguistics (Boeckx & Grohmann 2013; Bookheimer 2002; Friederici 2011; Kemmerer 2015), is a subfield of cognitive neuroscience which attempts to relate brain structures and processes to the comprehension, production, and acquisition of natural language. Lenneberg's influential *Biological Foundations of Language* (1967) identified biological issues relevant to the study of language that complemented Chomsky's abstract computational approach. Since then the development of electrophysiological and brain imaging technologies has generated a huge volume of empirical results which have allowed the brain to be mapped with respect to correlations between language tasks and neural activity. Such studies have established that the traditional view, which locates language exclusively in Broca's and Wernicke's areas, is only part of the answer. In common with the general multifunctional organization of brain regions mentioned above, the language network also includes the left-lateralized areas of the inferior frontal and the temporal lobes in which they are embedded as well as more distant parts of the frontal, parietal, and occipital lobes (Bookheimer 2002; Friederici 2011; Kemmerer 2015); Broca's area in particular is involved in a variety of visual and motor functions (Anderson 2010). It has also been possible to distinguish the areas specific to semantic processing within the more general language areas, and also from areas specific to sensory input modality processing (Binder et al 2009).

With respect to word meaning (Binder et al 2009; Kemmerer 2015, chs. 10-12), the closely related Grounded Cognition and Hub-and-Spoke models are gaining increasing acceptance. In the first of these the referents of linguistic expressions are mental representations of mind-external reality, and mental representations are based ultimately on perceptions of that reality as mediated by the various motor and perceptual modalities; the neural implementation of representations is held to be based on the physical activations of these modalities in response to external stimulation and motor interaction with the environment. The Hub-and-Spoke model goes on to claim that the activations of modality-specific regions are physically connected to and integrated in representations located in the temporal cortices, and that these synthetic representations are the referents of words, that is, word meanings; the representations are the hub and the modality-specific representations are the spokes; the hub integrates cortically distributed sensory and motor features of the mental representations to which words refer.

Influential examples of neural cognitive modelling are Ryder's SINBAD (Ryder 2004), which combines ANN architecture with teleological functionality to learn representations of environmental regularities via sensory input the structure of which becomes homomorphic with the environment over time, and State Space Semantics, which Paul and Patricia Churchland have developed over several decades and which is comprehensively stated most recently in Churchland (2012); for critiques see (Fodor & Lepore 1996, 1999) and for defenses (Laakso & Cottrell 2000; Shea 2018). With the explicit aim of supplanting CTM as the preferred explanatory paradigm for cognition, the Churchlands propose a dynamical systems account of how what is known of brain physiology and temporal processing generates cognition, where that system comprises a collection of interacting artificial neural networks. The essential features are as follows.

- Assuming a structured environment and beginning with the neonatal brain, a stimulus from the environment to a sensory modality generates a pattern of neural activation in the brain. With repeated presentation of any given stimulus A, dendritic and synaptic growth maps A to activation in a specific brain location. Cognitively this is learning; from a dynamical systems viewpoint the location is an attractor in the physical brain activation space. Via the same mechanism, stimuli B, C, D... are mapped to brain locations whose distance from one another is homomorphic with the similarity structure of the stimuli. As learning accumulates for a large number of stimuli, an activation

structure emerges in which similar stimuli generate activation in closely adjacent brain locations, and dissimilar stimuli activate relatively more distant locations. Each resulting concentration of activation locations is, in dynamical systems terms, a basin of attraction to which any future stimuli of a similar type are attracted. Over time, for each sensory modality, there emerges a structure which is a map of basins of attraction. Such maps are modality-specific representations of the external environment.

- Connections exist not only from sensory input modalities to brain maps, but also between maps and to association areas, the last-mentioned of which are activation areas which learn attractors from the co-activation of the sensory areas to which they are connected. These association areas implement cognitive concepts.
- The configuration of numerous interconnected sensory and association maps is formed in early life, and their representation of the mind-external environment is the basis for subsequent cognitive activity and learning. Interaction with the environment via sensory and motor functions generates trajectories whose relative similarity through the attractor maps is homomorphic with the structure of the perceived and enacted environmental interactions, with gradual modification of map structure and connectivity via dendritic and synaptic growth and atrophy. These trajectories represent the temporal structure of the environment.
- Churchland maintains that the vectors which constitute the attractor maps and trajectories through them constitute the brain's semantic system, which he refers to as 'state space semantics'. (Churchland 2012, 82).

Laakso & Cottrell (2000) proposed a modification of the Churchlands' position whereby the theoretical representational primitive should be partitioning of the high-dimensional neural state space into activation subspaces, replacing absolute position in the space: the number of units in the neural model determines the dimensionality of the state space and representations are vectors in the space, but because similar inputs cause similar vectors, activation - clusters form, and these clusters constitute the partition; the motivation was to counter criticism of state space semantics by Fodor & Lepore. This idea has subsequently been refined by Shea (2012, 2014, 2018), who argues that clusters in state space are the brain's 'vehicles of content'.

4. Conclusion

The foregoing historical outline confirms that many serious thinkers over a very long time span have subscribed to the ideas with which this discussion began: that the meanings of linguistic expressions are their significations of mental concepts, and that concepts are representations of the mind-external environment generated by the cognitive agent's interaction with that environment. What the outline has added to this general position is:

- The distinction between natural and conventional signification, which aligns well with Searle's original / derived intentionality one.
- The identification of the neural mechanisms of environmentally-caused conceptual representation.

These ideas provide a plausible context in which models of intrinsic linguistic meaning can be constructed.

References

- Adams, F. & Aizawa, K.** (2017). Causal theories of mental content. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Summer 2017 Edition). <https://plato.stanford.edu/archives/sum2017/entries/content-causal/>
- Allan, K.** (2009). *The Western Classical Tradition in Linguistics. 2nd ed.* London: Equinox Publishing.
- Anderson, M.L.** (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences* 33, 245-266.
- Atkin, A.** (2010). Peirce's theory of signs. Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Summer 2013 Edition). <https://plato.stanford.edu/archives/sum2013/entries/peirce-semiotics/>
- Barsalou, L.** (2008). Grounded Cognition. *Annual Review of Psychology* 59, 617-645.
- Barsalou L.** (2010). Grounded cognition: past, present, and future. *Topics in Cognitive Science* 2, 716-724.
- Barsalou, L.** (2016a). On Staying Grounded and Avoiding Quixotic Dead Ends. *Psychonomic Bulletin & Review* 23, 1122-1142.
- Barsalou L.** (2016b). Situated conceptualization: Theory and applications. In: *Foundations of embodied cognition: Volume 1 Perceptual and emotional embodiment*: 11-37. East Sussex: Psychology Press.
- Barsalou, L. W.** (2017). What does semantic tiling of the cortex tell us about semantics? *Neuropsychologia* 105, 18-38.
- Barsalou, L., Dutriaux, L. & Scheepers, C.** (2018). Moving beyond the distinction between concrete and abstract concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences* 373. 20170144, DOI: <http://dx.doi.org/10.1098/rstb.2017.0144>.
- Binder, J., Westbury, C., McKiernan, K., Possing, E. & Medler, D.** (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience* 17, 905-917.
- Binder, J., Desai, R., Graves, W. & Conant, L.** (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex* 19, 2767-2796.
- Binder J.** (2016). In defense of abstract conceptual representations. *Psychonomic Bulletin & Review* 23, 1096-1108.
- Binder, J., Conant, L., Humphries, C., Fernandino, L., Simons, S., Aguilar, M. & Desai, R.** (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology* 33, 3-4, 130-174.
- Boeckx, C. & Grohmann, K.** (eds.) (2013). *The Cambridge handbook of biolinguistics*. Cambridge, UK: Cambridge University Press.
- Bookheimer, S.** (2002). Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience* 25, 151-188.
- Boone, W. & Piccinini, G.** (2016). The cognitive neuroscience revolution. *Synthese* 193, 1509-1534.
- Bradie, M. & Harms, W.** (2016). Evolutionary epistemology. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2017 Edition). <https://plato.stanford.edu/archives/spr2017/entries/epistemology-evolutionary/>
- Buckner, C. & Garson, J.** (2019). Connectionism. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Fall 2019 Edition). <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>
- Buroker, J.** (2014). Port Royal logic. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*,

- (Spring 2017 Edition). <https://plato.stanford.edu/archives/spr2017/entries/port-royal-logic/>
- Buss, D.** (2007). *Evolutionary Psychology: The New Science of the Mind*. Boston: Allyn and Bacon.
- Carey, S.** (2009). *The Origin of Concepts*. Oxford: Oxford University Press.
- Churchland, P.** (2012). *Plato's camera. How the physical brain captures a landscape of abstract universals*. Cambridge: MIT Press.
- Clark, A.** (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford University Press.
- Cohen, S.** (2016). Aristotle's metaphysics. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Winter 2016 Edition). <https://plato.stanford.edu/archives/win2016/entries/aristotle-metaphysics/>
- Cole, D.** (2020). The Chinese Room Argument. In: Zalta, E. (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). <https://plato.stanford.edu/archives/spr2020/entries/chinese-room/>
- Comesana, J. & Klein, P.** (2019). Skepticism. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Winter 2019 Edition). <https://plato.stanford.edu/archives/win2019/entries/skepticism/>
- Conway C. & Pisoni D.** (2008). Neurocognitive basis of implicit learning of sequential structure and its relation to language processing. *Annals of the New York Academy of Sciences* 1145, 113-31.
- Croft, W. & Cruse, D.** (2004). *Cognitive linguistics*. Cambridge: Cambridge University Press.
- Crane, T. & French, C.** (2015). The Problem of Perception. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2017 Edition). <https://plato.stanford.edu/archives/spr2017/entries/perception-problem/>
- Downes, S.** (2018). Evolutionary psychology. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Fall 2018 Edition). <https://plato.stanford.edu/archives/fall2018/entries/evolutionary-psychology/>
- Fodor, J.** (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.
- Fodor, J. & Lepore, E.** (1996). Churchland on state space semantics, In: R. McCauley (ed.), *The Churchlands and Their Critics: 145-148*. Oxford: Blackwell.
- Fodor, J. & Lepore, E.** (1999). All at Sea in Semantic Space: Churchland on Meaning Similarity. *Journal of Philosophy* 96, 381-403.
- Friederici, A.** (2011). The brain basis of language processing: from structure to function. *Physiological Review* 91, 1357-1392.
- Gazzaniga, M., Ivry, R. & Mangun, G.** (2019). *Cognitive neuroscience: The biology of the mind*. New York: Norton.
- Geeraerts, D. & Cuyckens, H.** (eds.) (2010). *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press.
- Gendler Szabo, Z.** (1998). Language, early modern philosophy of. In: *Routledge Encyclopedia of Philosophy. Philosophy of Language: 128-136*. London: Routledge.
- Griffiths, P.** (2009). The Distinction Between Innate and Acquired Characteristics. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2017 Edition). <https://plato.stanford.edu/archives/spr2017/entries/innate-acquired/>
- Harris J., Petersen, R. & Diamond M.** (2001). The cortical distribution of sensory memories. *Neuron* 30, 315-318.
- Jackendoff, R.** (2002). *Foundations of language: brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
- Jackendoff, R.** (2015). *A user's guide to thought and meaning*. Oxford: Oxford University Press.

- Jacob, P.** (2019). Intentionality. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2019 Edition). <https://plato.stanford.edu/archives/spr2019/entries/intentionality>
- Kemmerer, D.** (2015). *Cognitive Neuroscience of Language*. Psychology Press.
- King, P.** (2007). Abelard on mental language. *American Catholic Philosophical Quarterly* 81, 169-187.
- King, P. & Arlig, A.** (2010). Peter Abelard. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Fall 2018 Edition). <https://plato.stanford.edu/archives/fall2018/entries/abelard>
- Klima, G.** (2017). The medieval problem of universals. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Winter 2017 Edition). <https://plato.stanford.edu/archives/win2017/entries/universals-medieval>
- Laakso, A. & Cottrell, G.,** (2000). Content and Cluster Analysis: Assessing Representational Similarity in Neural Systems. *Philosophical Psychology* 13, 47-76.
- Lagerlund, H.** (2017). Mental representations in medieval philosophy. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Fall 2017 Edition). <https://plato.stanford.edu/archives/fall2017/entries/representation-medieval>
- Lakoff, G. & Johnson, M.** (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York: Basic Books.
- Langacker, R.** (2008). *Cognitive Grammar: A Basic Introduction*. New York: Oxford University Press.
- Lau, J. & Deutsch, M.** (2014). Externalism about mental content. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Fall 2019 Edition). <https://plato.stanford.edu/archives/fall2019/entries/content-externalism>
- Lenneberg, E.** (1967). *Biological foundations of language*. New York: John Wiley and Sons.
- Levin, J.** (2018). Functionalism, *Stanford Encyclopedia of Philosophy*. In: Zalta, E. (ed.), (Fall 2018 Edition). <https://plato.stanford.edu/archives/fall2018/entries/functionalism>
- Lycan, W.** (2018). *Philosophy of language: a contemporary perspective*. 3rd ed. London: Routledge.
- Lyons, J.** (2016). Epistemological Problems of Perception. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2017 Edition). <https://plato.stanford.edu/archives/spr2017/entries/perception-episprob>
- Macleod, C.** (2016). John Stuart Mill. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Fall 2018 Edition). <https://plato.stanford.edu/archives/fall2018/entries/mill>
- Marenbon, J.** (2016). Anicius Manlius Severinus Boethius. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Winter 2016 Edition). <https://plato.stanford.edu/archives/win2016/entries/boethius>
- Margolis, E. & Laurence, S.** (2015). *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, MA: MIT Press.
- Margolis, E. & Laurence, S.** (2019). Concepts. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Summer 2019 Edition). <https://plato.stanford.edu/archives/sum2019/entries/concepts>
- Markie, P.** (2017). Rationalism vs. Empiricism. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Fall 2017 Edition). <https://plato.stanford.edu/archives/fall2017/entries/rationalism-empiricism>
- Matheson, H. & Barsalou, L.** (2018). Embodiment and grounding in cognitive neuroscience. In: J. Wixted, (ed.) *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, vol. 3. 4th ed.: 1-32*. Hoboken, NJ: Wiley.
- McCawley, J.** (1976). *Grammar and meaning*. New York: Academic Press.
- McCawley, J.** (1995). Generative semantics, In: Verschuren, J. et al. (eds.), *Handbook of Pragmatics: 311-319*. Amsterdam: Benjamins.

- McKinsey, M.** (2018). Skepticism and Content Externalism. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Summer 2018 Edition). <https://plato.stanford.edu/archives/sum2018/entries/skepticism-content-externalism>
- Meier-Oeser, S.** (2011). Medieval semiotics. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Summer 2011 Edition). <https://plato.stanford.edu/archives/sum2011/entries/semiotics-medieval>
- Metzger M.** (2017). Dynamical models of cognition. In: L. Magnani, T. Bertolotti (ed.) *Springer Handbook of Model-Based Science*: 639-655. New York: Springer.
- Modrak, D.** (2001). *Aristotle's theory of language and meaning*. Cambridge: Cambridge University Press.
- Neander, K.** (2012). Teleological Theories of Mental Content. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2018 Edition). <https://plato.stanford.edu/archives/spr2018/entries/content-teleological>
- Neander, K.** (2017). *A mark of the mental: In defense of informational teleosemantics*. Cambridge: MIT Press.
- Newmeyer, F.** (1995). *Generative linguistics: an historical perspective*. London: Routledge.
- Oppy, G. & Dowe, D.** (2016). The Turing test. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). <https://plato.stanford.edu/archives/spr2019/entries/turing-test>
- Orzack, S. & Forber, P.** (2010). Adaptationism. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2017 Edition). <https://plato.stanford.edu/archives/spr2017/entries/adaptationism>
- Patton, L.** (2018). Hermann von Helmholtz. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Winter 2018 Edition). <https://plato.stanford.edu/archives/win2018/entries/hermann-helmholtz>
- Petersson, K. & Hagoort, P.** (2012). The neurobiology of syntax: beyond string sets. *Philosophical Transactions of the Royal Society B* 367, 1971-1983.
- Piccinini, G.** (2007). Computationalism, the Church-Turing thesis, and the Church-Turing fallacy. *Synthese* 154, 97-120.
- Piccinini, G.** (2009). Computationalism in the Philosophy of Mind. *Philosophy Compass* 4, 515-532.
- Piccinini, G. & Scarantino, A.** (2011). Information processing, computation, and cognition. *Journal of Biological Physics* 37, 1-38.
- Piccinini, G. & Bahar, S.** (2013). Neural computation and the computational theory of cognition. *Cognitive Science* 34, 453-488.
- Piccinini, G. & Shagrir, O.** (2014). Foundations of computational neuroscience. *Current Opinion in Neurobiology* 25, 25-30.
- Piccinini, G.** (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.
- Piccinini, G.** (2016). The computational theory of cognition. In: Müller, V. (ed.), *Fundamental issues of artificial intelligence*: 203-221. New York: Springer.
- Piccinini, G.** (2018). Computation and Representation in Cognitive Neuroscience. *Minds and Machines* 28, 1-6.
- Piccinini, G. & Bahar, S.** (2013). Neural computation and the computational theory of cognition. *Cognitive Science* 34, 453-488.
- Pinker, S.** (2007). *The Stuff of Thought: Language as a Window into Human Nature*. London: Penguin.
- Pitt, D.** (2012). Mental representation. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Winter 2018 Edition). <https://plato.stanford.edu/archives/win2018/entries/mental-representation>

- Pojman, P.** (2019). Ernst Mach. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2019 Edition). <https://plato.stanford.edu/archives/spr2019/entries/ernst-mach>
- Port, R. & van Gelder, T.** (1995). *Mind as motion: dynamics, behavior, and cognition*. Cambridge: MIT Press.
- Prinz, J.** (2002). *Furnishing the Mind*. Cambridge: MIT Press.
- Rescorla, M.** (2020). The computational theory of mind. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). <https://plato.stanford.edu/archives/spr2020/entries/computational-mind>
- Rodriguez-Pereyra, G.** (2014). Nominalism in metaphysics. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Summer 2019 Edition). <https://plato.stanford.edu/archives/sum2019/entries/nominalism-metaphysics>
- Rupert, R.** (2008). Causal Theories of Mental Content. *Philosophy Compass* 3, 353-80.
- Ryder, D.** (2004). SINBAD Neurosemantics: A Theory of Mental Representation. *Mind & Language* 19, 211-240.
- Rysiew, P.** (2016). Naturalism in epistemology. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2017 Edition). <https://plato.stanford.edu/archives/spr2017/entries/epistemology-naturalized>
- Samet, J.** (2019). The Historical Controversies Surrounding Innateness. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Summer 2019 Edition). <https://plato.stanford.edu/archives/sum2019/entries/innateness-history>
- Samet, J. & Zaitchik, D.** (2017). Innateness and Contemporary Theories of Cognition. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy* (Fall 2017 Edition). <https://plato.stanford.edu/archives/fall2017/entries/innateness-cognition>
- Schmidhuber, J.** (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks* 61, 85-117.
- Searle, J.** (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences* 3, 417-57.
- Searle, J.** (1984). *Minds, Brains and Science*. Cambridge: Harvard University Press.
- Searle, J.** (1989). 'Artificial Intelligence and the Chinese Room: An Exchange', *New York Review of Books* 36, 2, February 16.
- Searle, J.** (1990). Is the Brain's Mind a Computer Program? *Scientific American* 262, 1, 26-31.
- Searle, J.** (2002). Twenty-one Years in the Chinese Room. In: Preston, J. & Bishop, M. (eds.). *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence: 51-69*. Oxford: Oxford University Press.
- Searle, J.** (2010). Why Dualism (and Materialism) Fail to Account for Consciousness. In: R. Lee (ed.) *Questioning Nineteenth Century Assumptions about Knowledge III: Dualism*. New York: SUNY Press.
- Shea, N.** (2012). New thinking, innateness and inherited representation. *Philosophical Transactions of the Royal Society B. Biological Sciences* 367, 1599, 2234-2244.
- Shea, N.** (2014). Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society* 114, 123-144.
- Shea, N.** (2018). *Representation in cognitive science*. Oxford: Oxford University Press.
- Shields, C.** (2012). *The Oxford Handbook on Aristotle*. Oxford: Oxford University Press.
- Shields, C.** (2014). *Aristotle. 2nd ed.* London: Routledge.
- Shields, C.** (2015). Aristotle. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Winter 2016 Edition). <https://plato.stanford.edu/archives/win2016/entries/aristotle>
- Shields, C.** (2016). Aristotle's psychology. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Winter 2016 Edition). <https://plato.stanford.edu/archives/win2016/entries/aristotle-psychology>

- Siegel, S.** (2016). The Contents of Perception. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Winter 2016 Edition). <https://plato.stanford.edu/archives/win2016/entries/perception-contents>
- Spade, P.** (2007). *Thoughts, Words and Things: An Introduction to Late Mediaeval Logic and Semantic Theory*, Version 1.2. <https://www.scribd.com/document/30475288/Spade-Thoughts-Words-and-Things>
- Spade, P.** (2016). Medieval philosophy. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Summer 2018 Edition). <https://plato.stanford.edu/archives/sum2018/entries/medieval-philosophy>
- Spade, V. & Panaccio, C.** (2019). William of Ockham. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2019 Edition). <https://plato.stanford.edu/archives/spr2019/entries/ockham>
- Stampe, D.** (1977). Toward a Causal Theory of Linguistic Representation. In: French, P., Wettstein, H. & Uehling, T. (ed.), *Midwest Studies in Philosophy: 42-63*. Minneapolis: University of Minnesota Press.
- Talmy, L.** (2000). *Toward a cognitive semantics*. Cambridge: MIT Press.
- Thomson, E. & Piccinini, G.** (2018). Neural representations observed. *Minds and Machines* 28, 191-235.
- Tornau, C.** (2019). Saint Augustine. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Winter 2019 Edition). <https://plato.stanford.edu/archives/win2019/entries/augustine>
- Uzgalis, W.** (2018). John Locke. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2019 Edition). <https://plato.stanford.edu/archives/spr2019/entries/locke>
- Ward, L.** (2002). *Dynamical cognitive science*. Cambridge: MIT Press.
- Wilson, R. & Foglia, L.** (2015). Embodied Cognition. In: Zalta, E. (ed.), *Stanford Encyclopedia of Philosophy*, (Spring 2017 Edition). <https://plato.stanford.edu/archives/spr2017/entries/embodied-cognition>
- Wilson-Mendenhall C., Simmons W., Martin A. & Barsalou L.** (2013). Contextual processing of abstract concepts reveals neural representations of nonlinguistic semantic content. *Journal of Cognitive Neuroscience* 25, 920-935.
- Yeh, W. & Barsalou, L.** (2006). The situated nature of concepts. *American Journal of Psychology* 119, 349-384.

Fitting the Menzerath-Altmann Law: How Much Data Do You Need?

Andrei Beliankou¹, Reinhard Köhler²

Abstract

The present article shows by means of an example in connection with the Menzerath-Altmann law (MAL), in which way a goodness-of-fit test can correlate with the available amount of data. The experiment was carried out on three levels of the MAL and shows partly unexpected behaviour of this dependency.

Keywords: Menzerath-Altmann Law; Arens' Law; goodness-of-fit; data size

1. Introduction

Probably almost everyone who is working scientifically actively in the field of quantitative linguistics, whether as a researcher, lecturer or student, is constantly confronted with the question of whether a sufficient amount of data is available to reliably fit a mathematical model to one's data. A slightly different aspect of the same problem is addressed by the question of how much data is needed to test a corresponding hypothesis. In some cases, e.g. when working with Gauss or Binomial-distributed data it is possible to calculate the probability with which a test result will be realised or, vice versa, to determine the minimum amount of data with which a desired probability will be achieved. In very many cases, however, no methods are known with which such questions can be answered. Here we have to rely on empirical values, if such are available. However, we are usually dependent on specific experiments.

The present article shows by means of an example in connection with the Menzerath-Altmann law (MAL) (Altmann 1980) in which way a goodness-of-fit test may fail to correlate with the available amount of data. The experiment was carried out on three levels of the MAL and shows partly unexpected behaviour of this dependency.

Our study was first designed to contribute evidence on data from a corpus (cf. below) analysing the relations Sentence-Cause-Word, Clause-Word-Morpheme, and Word-Morpheme-Grapheme with respect to the lengths of the constructs in terms of the numbers of constitutes, as usual in MAL research. In view of the relatively large number of syntactically and morphologically annotated data we were able to acquire, the idea arose to systematically track the quality of fit as the amount of data increases. The fitting was made using the NLREG (2005) program, whose calculation of the coefficient of determination R^2 was adopted.

2. Data structure and acquisition

Availability of a high-quality annotated source of linguistic evidence is crucial for every data-based study. Claiming scientifically valid investigation on a growing dataset we rely on years

¹ Universität Trier 54286 Trier, Germany, Andrej.Belenkow@gmail.com.

² Universität Trier 54286 Trier, Germany, koehler@uni-trier.de.

of work by linguists annotating texts on different levels. It is crucial to have such a base with comparable annotations for texts, sentences, clauses, words, and even morphemes.

We have chosen to use the *TüBa-D/Z treebank* (Telljohann et al., 2004) in version 9. This treebank was one of the first large scale annotation experiments on German newspaper texts using a hybrid syntactic approach. The newest version 11 is larger but conceptually the same and does not comprise any additional structural information for the same texts.

The corpus is segmented into sentences on the highest level of annotation. An exhaustive syntactic layer as well as morphological annotations on the word level are present. This resource reflects some aspects of derivative morphology regarding splittable verbal prefixes and nominal composita.

In the corpus, an explicit text segmentation is not provided. Nevertheless the text margins are recognizable from the raw corpus exports and can be easily reconstructed in comparison with the underlying newspaper articles from the renown German *TAZ* newspaper. All texts date from the 1980s and 1990s.

Having text, sentence, clause, and word boundaries in the corpus we had to extend the data set by morphological and graphemical levels of annotation. For these tasks we used an excellent morphological analyser *SMOR* (Schmid et al., 2004). This analyser provides both inflectional and derivational analyses.

We extracted all tokens from provided texts and normalized the spelling in regard to writing of the numbers (we spelled them out: “10” became “ten”) and different spellings of the same words (e.g. “Strasse” became systematically “Straße”). This normalization step was crucial for consistent length determination of morphemes and reliable morphological segmentation.

The morphological analyser *SMOR* was used in the not disambiguated mode and provided all possible analyses, not only the most probable. We filtered segmentation examples according to referent POS tags from the corpus (i.e. we retained the correct form analysis) and systematically removed shorter analyses since the deeper segmentation often revealed correct handling of compound nouns.

Sentence and word boundaries and the corresponding counts could be obtained from the raw corpus data. The clause annotation has been done semi-automatically: due to differences in text genre, a substantial part of sentences in some articles lacked any verbal forms and contained only nominal phrases. We had to process elliptical sentences as well. Manual annotation in such cases was more efficient since an automated parser accounting all these cases was too error-prone.

The overall analysed material amounts to 700 sentences with 12795 words in total.

We prepared for every analysis scheme a dataset in the following form:

- sentence length in clauses, sentence count, average clause length in words
- clause length in words, clauses count, average word length in morphemes
- word length in morphemes, words count, average morpheme length in letters

Every dataset was split into 20 parts ranging from 5% to 100% in 5% steps concerning the sentence count in the overall data set. We did not employ any shuffling or sampling, the parts are extracted from annotated data in the linear order.

3. The Word-Morpheme-Grapheme level

The lowest level for which we had reliable data was that of graphemes, the number of which could be regarded as a measure of the length of the units constituting them, the morphemes. As usual, we measured the lengths of the words in terms of the number of morphemes or syllables, etc. In this case, we had morphologically annotated data containing morpheme boundaries.

If you have a program with an algorithm that allows multiple values for the x-variable, the numerical values of the lowest level can be used directly, otherwise you have to use averages.

In the first case, you will find a greater variance than in the second, which pretends a lower variance through implicit smoothing. Therefore, a higher tolerance threshold must be applied with the first method than with the second (An R^2 value of 0.5 should be considered as a good result in the first case whereas an R^2 above 0.9 can be assessed as good in the latter case). The most frequently applied method for MAL studies is the second one, probably because Gabriel Altmann used it when he introduced the law and illustrated it in this way (Altmann 1980).

Our question was now how much data would be needed to obtain a stable fit of the MAL function and also stable parameter estimations. Table 1 shows the Coefficient of determination and its adjusted version of multiple determination in the course of data incrementation. The entire corpus contains 700 sentences, the individual increments are formed by portions of 5% each. Figure 1 shows the course of the coefficient.

Table 1
Data Size and Determination Coefficient

Data size	R^2	R^2 ad-justed	Data size	R^2	R^2 ad-justed
5	0.9241	0.8938	55	0.9379	0.9202
10	0.9352	0.9092	60	0.9386	0.9211
15	0.9361	0.9105	65	0.9395	0.9222
20	0.9352	0.9092	70	0.9362	0.9179
25	0.9611	0.9482	75	0.9358	0.9175
30	0.9626	0.9501	80	0.9355	0.9170
35	0.9226	0.9004	85	0.9341	0.9153
40	0.9163	0.8924	90	0.9344	0.9156
45	0.9175	0.8939	95	0.9335	0.9145
50	0.9166	0.8928	100	0.9323	0.9130

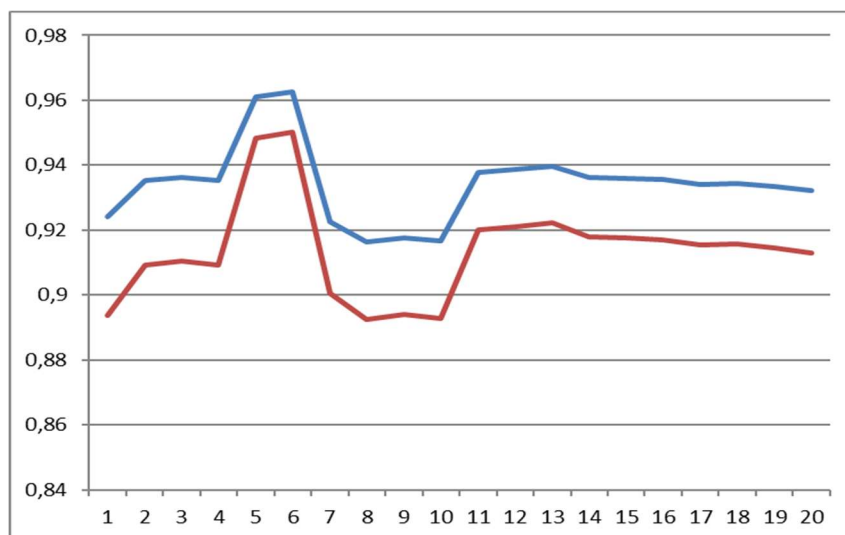


Figure 1 The change of the R^2 value in dependence of the data size

The first steps of enlargement show large changes in the R^2 values, which is not surprising. Also the rapid decrease of the coefficients from step 6 on and the subsequent increase are to be

regarded as random, which we did not test, since there was no hypothesis to test the contrary. The further course of the values is then quite even, albeit with decreasing tendency. So we see roughly what was to be expected. If we use even more data, we could perhaps obtain again increasing R^2 values. At this point it is interesting to observe the changes in the estimated parameter values. For this purpose we will look at Table 2 and Figure 2.

Table 2

The three parameters of the MAL on the Word-Morpheme-Grapheme level in the course of increasing data

a	b	c	a	b	c
5.1926	-0.0648	0.0856	5.2378	-0.0043	0.0864
5.2756	-0.0490	0.0942	5.2100	-0.0086	0.0850
5.3282	-0.0219	0.1019	5.1908	-0.0027	0.0858
5.2756	-0.0490	0.0942	5.1835	-0.0052	0.0851
5.4118	0.0398	0.1123	5.1656	-0.0035	0.0848
5.4579	0.0624	0.1189	5.1556	-0.0030	0.0847
5.3063	-0.0244	0.0835	5.1469	-0.0027	0.0847
5.2694	-0.0237	0.0803	5.1564	0.0055	0.0869
5.2531	-0.0234	0.0796	5.1727	0.0080	0.0881
5.2292	-0.0225	0.0793	5.1867	0.0103	0.0890

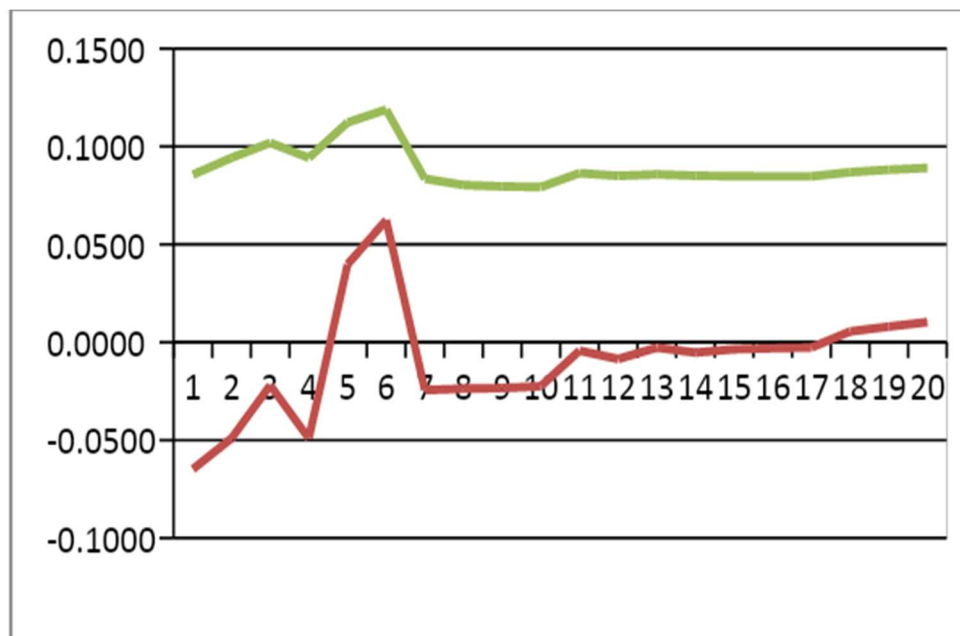


Figure 2 The parameters *b* and *c*

Of the three parameters, b is particularly interesting because it changes, after 17 steps of data enlargement, its sign. Apparently, even characteristic properties of the parameters such as the shape of the function may change depending on data size.

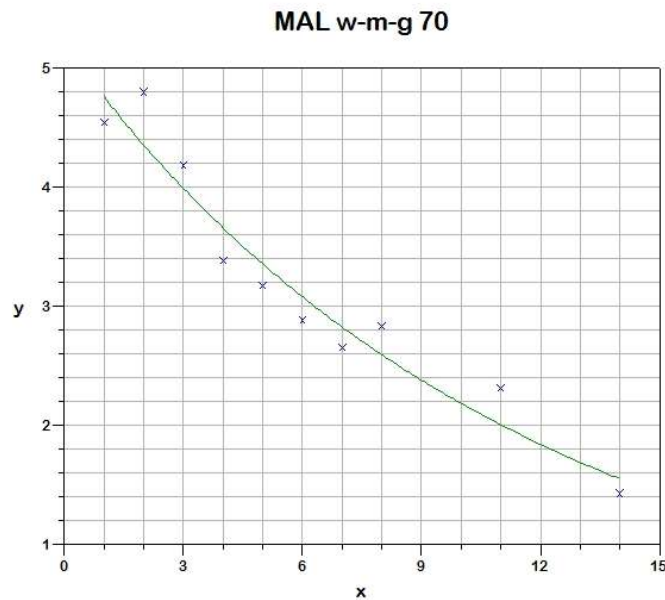


Figure 3 The shape of the MAL function on the word-morpheme-grapheme level on 70% of the data, i.e. at step 14 (cf. Figure 2)

4. The Clause-Word-Morpheme level

The values of R^2 in Table 3, which describes the fitting of the MAL on the next higher level to the data are rather disappointing although we use the same data source. The corresponding curves look, at first sight, similar to those of the lower level but the poor R^2 values speak a different language.

Table 3
Data size and coefficients of determination

data size	R^2	R^2_{adjusted}	data size	R^2	R^2_{adjusted}
5	0.0573	-0.0606	55	0.2665	0.2054
10	0.5137	0.4596	60	0.2752	0.2148
15	0.5754	0.5307	65	0.2804	0.2205
20	0.5267	0.4816	70	0.2902	0.2310
25	0.2382	0.1720	75	0.3163	0.2593
30	0.2098	0.1411	80	0.3234	0.2670
35	0.2165	0.1483	85	0.3199	0.2632
40	0.1474	0.0764	90	0.2553	0.1957

45	0.1791	0.1107	95	0.2791	0.2214
50	0.2216	0.1567	100	0.1159	0.0422

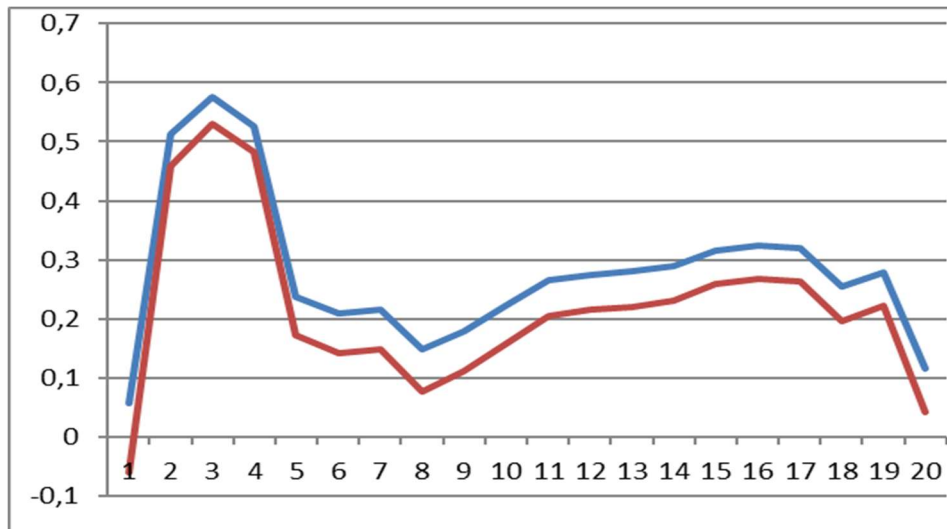


Figure 4 The R² values on the clause-word-morpheme level

Table 4
The estimated values of the three parameters

a	b	c	a	b	c
1.3421	-0.0030	0.0091	1.2174	0.0082	0.0045
1.2609	-0.0374	-0.0170	1.2181	0.0064	0.2148
1.2120	-0.0242	0.0159	1.2091	0.0090	-0.0045
1.1757	0.0590	-0.0032	1.1900	0.0149	-0.0044
1.2432	0.0237	-0.0033	1.1655	0.0243	-0.0040
1.3201	-0.0133	-0.0052	1.1631	0.0233	-0.0042
1.3186	-0.0121	-0.0052	1.1721	0.0170	-0.0046
1.3092	-0.0119	-0.0045	1.2149	-0.0088	-0.0057
1.2745	0.0000	-0.0041	1.2258	-0.0132	-0.0057
1.2576	-0.0039	-0.0048	1.1581	0.0637	0.0034

A look at the parameters b and c in Table 4 and Figure 5 shows erratic values. The estimates are not stable even on the material of 700 sentences. The development of R² is apparently still unpredictable at this point.

On all levels of the MAL a decreasing tendency is predicted whereas a clear increasing course is obtained. This phenomenon cannot be explained by the hypothesis that a level was skipped because this would yield a concave shape of the curve, known as Arens' Law (Arens 1965). Another, yet unknown problem must be responsible for the observed violation of our expectations.

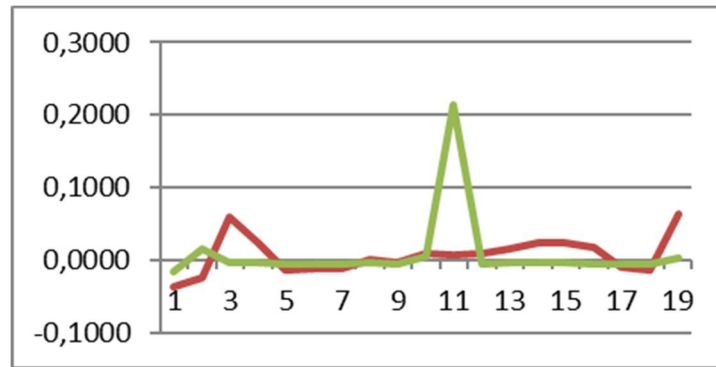


Figure 5 Parameters b and c do not show any trend

Figure 6 On the Clause-Word-Morpheme level, an unexpected trend can be observed

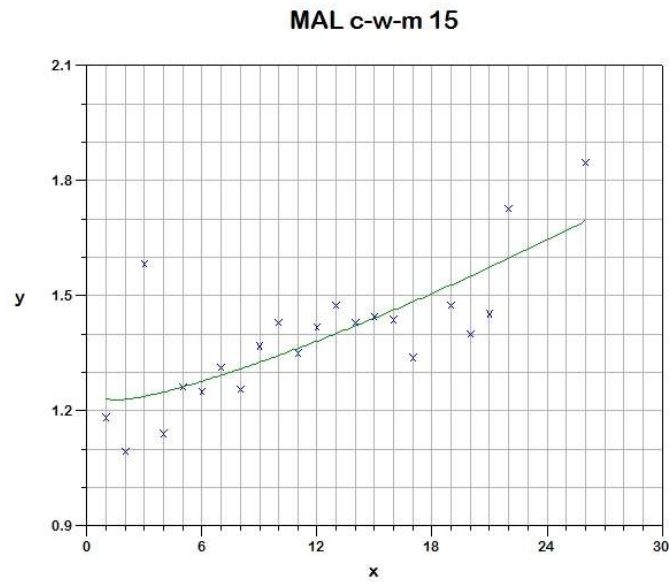


Figure 6.1 On the Clause-Word-Morpheme level, an unexpected trend can be observed

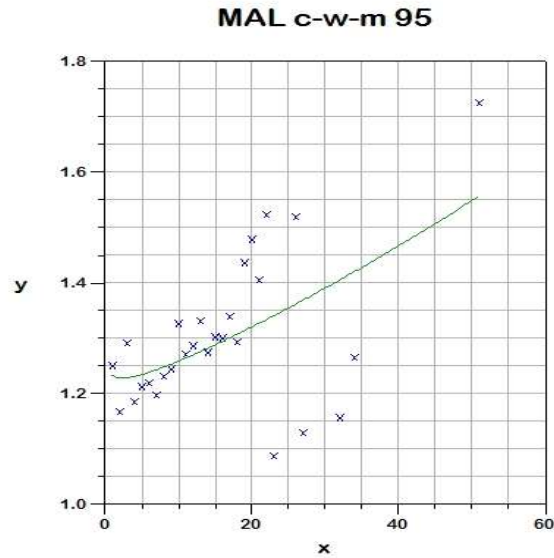


Figure 6.2 On the Clause-Word-Morpheme level, an unexpected trend can be observed

5. The Sentence-Clause-Word level

Finally, we studied the situation on the level of sentences and their constituents. Table 5 and Figure 7 show the development of the R^2 values with increasing data size.

Table 5
Goodness-of-fit on the sentence level

data size	R^2	R^2 adjusted	data size	R^2	R^2 adjusted
5	0.612	0.2241	55	0.8446	0.7825
10	0.1655	-0.669	60	0.8549	0.7969
15	0.322	-0.356	65	0.8706	0.8189
20	0.8423	0.6847	70	0.8709	0.8192
25	0.8193	0.6387	75	0.8673	0.8231
30	0.8172	0.6345	80	0.8642	0.8189
35	0.7578	0.5157	85	0.8689	0.8252
40	0.2623	-0.4755	90	0.8709	0.8278
45	0.7236	0.4472	95	0.8849	0.8465
50	0.8781	0.7968	100	0.8895	0.8527

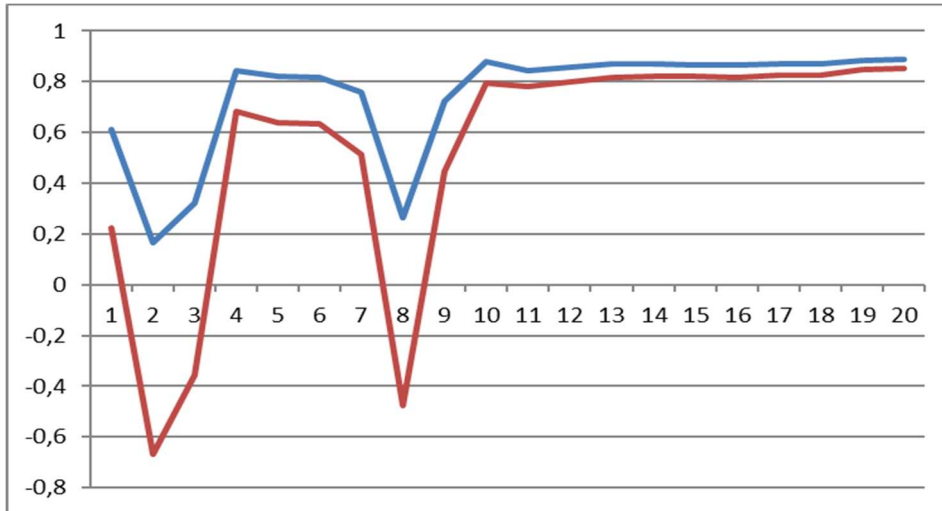


Figure 7 Development of the R^2 values with increasing data size

In the first half of the incremental process, the values of the determination coefficients drop dramatically and return to higher numbers, reaching stable good fits in the second part. Interestingly, the second downwards peak of the R^2 curve (Figure 7) corresponds exactly to the peak of both a and b parameters (Figure 8) while the first valley of R^2 has no counterpart in the parameter values.

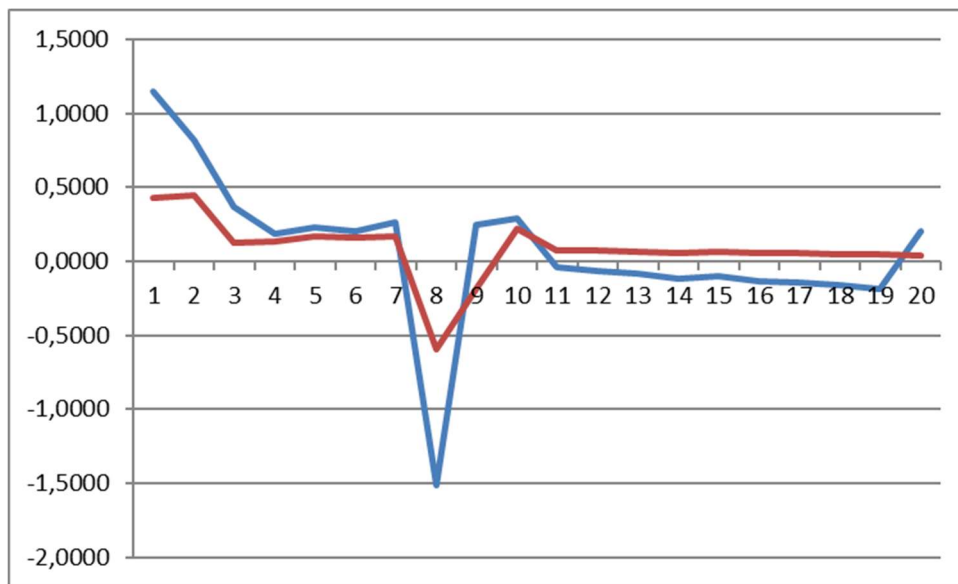


Figure 8 Parameters b and c in the course of increasing data size

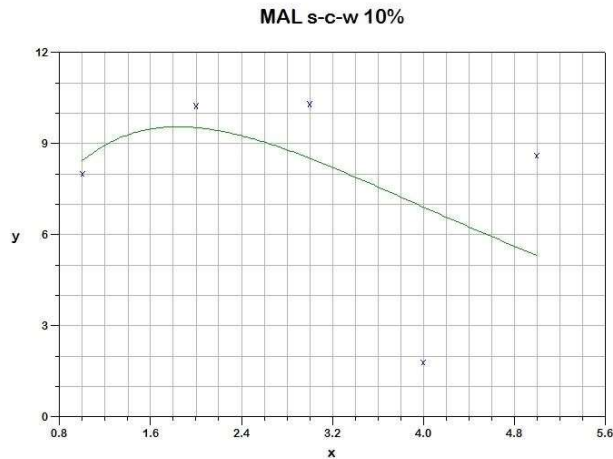


Figure 9.1 Not enough data for a correct shape of the function at 10% of the data (left); best result is obtained at 100% (right)

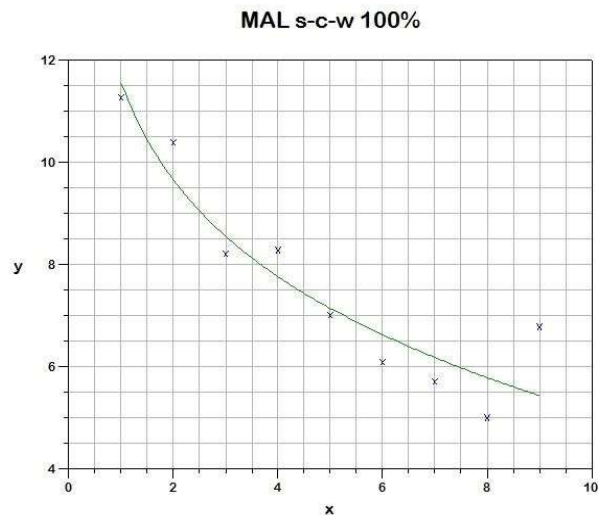


Figure 9.2 Enough data for a correct shape of the function at 100% of the data; best result is obtained at 100%

On this level, the fit of the function to the data is as expected.

6. Conclusion

Our experiments indicate an estimation of the amount of data which is needed for a successful fitting of the MAL to empirical data is not possible in a straightforward manner. The result of a fitting procedure depends on the level on which the individual study is performed and may also depend on a number of yet unknown factors. We could see that, in our case, the fitting of data on the clause level yielded the worst result. Researchers could assume that the concept of "clause" was ill-defined but, on the other hand, fitting of the function on the two other levels which involve the clause level (with the same data source) resulted in acceptable R^2 values and correct function shapes. Follow-up studies are needed to clarify the reasons for our preliminary perplexity.

References

- Altmann, G.** (1980). Prolegomena to Menzerath's law. In: Grotjahn, R. (ed.), *Glottometrika 2: 1-10*. Bochum: Brockmeyer.
- Altmann, G.** (1983). "H. Arens' «Verborgene Ordnung» und das Menzerathsche Gesetz". In: Faust, M., Harweg, R., Lehfeldt, W. & Wienold, G. (eds.), *Allgemeine Sprachwissenschaft. Sprachtypologie und Textlinguistik: 31-39*. Tübingen: Narr.
- Arens, H.** (1965). *Verborgene Ordnung*. Düsseldorf: Schwann.
- Grzybek, P. & Stadlober, E.** (2007). Do we have problems with Arens' law? A new look at the sentence-word relation. In: Grzybek, P. & Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 203-218*. Berlin: de Gruyter.
- Heike, T., Hinrichs, E. & Kübler, S.** (2004). *The TüBa-D/Z treebank: Annotating German with a context-free backbone*. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004): 2229-2235*.
- Schmid, H., Fitschen, A. & Heid, U.** (2004). SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004): 1263-1266*.
- Sherrod, P.H.** (2005). *NLREG. Nonlinear Regression Analysis (Software)*.

Theoretical Thoughts and Practical Advice on the Length of Shots by Early Soviet Film Directors

Veronika Schmidt¹

Abstract

This paper focuses on the criteria determining shot length according to the early Soviet avant-garde film directors and film theorists Timošenko, Pudovkin, Vertov, Ėjzenštejn and Kulešov. They used shot length to optimise the attentiveness and comprehension of their audience, and to influence their emotions with the rhythm created by it. Processing speed of sensory perception as a determining factor is discussed. When the dimension of sound was introduced to the realm of film, the rhythmic function of shot length was transferred to movements within the shot and to the soundtrack. Forces influencing the filmmakers are intuition, creative ecstasy and censorship.

Keywords: quantitative film studies; shot length; film length; feature film; documentary; montage; Russian; Soviet; Sergej M. Ėjzenštejn; Sergej M. Eisenstein; Lev Kulešov; Lev Kuleshov; Vsevolod Pudovkin; Dziga Vertov; Semen Timošenko; Semen Timoshenko

1. Introduction

Some aspects of Peter's academic life are touched upon here: Russian studies and his work on the new branch of quantitative film studies. He was probably one of the first to publish a paper on quantitative film studies (Grzybek & Koch 2012), suggesting that one observable surface structure of film, i.e. successive single shots with their specific lengths, can be summarized into intervals depending of their length, thus forming a shot length distribution. Such a shot length distribution can be fitted to a theoretical distribution – the Zipf-Alekseev distribution – and this theoretical distribution can be interpreted in terms of underlying processes which might have generated it. In other words, he used a method from quantitative linguistics and successfully transferred the approach to the realm of film. P. Grzybek (2012-2013) also oversaw the first project on quantitative film studies, searching for regularities of shot length distributions on the basis of 25 Yugoslav partisan films of the 1960s.² Finally, within the context of film studies, he was the supervisor of a PhD thesis in which shot length data of 70 Soviet feature films were successfully fitted to the Zipf-Alekseev distribution and the Menzerath-Altmann law (cf. Schmidt 2019). This paper here highlights his converging research interest in quantitative film studies, quantitative linguistics and Russian studies. The idea is to review theoretical ideas on shot length in order to better understand underlying processes and (potential) boundary conditions³ resulting in particular distributional patterns of shot lengths.

¹ Schmidt_Vero@web.de.

² The results of this project have not been published. Therefore, I take the liberty to reproduce some of the results in this memorial volume. The following distributions were tested and yielded good or very good R²-values, when fitted to the data of the summarized shot lengths: mixed negative binomial distribution, negative binomial distribution, hyper-Pascal distribution, double exponential distribution and the Zipf-Alekseev distribution. There was recently no time to analyse and discuss these results in terms of underlying regularities and boundary conditions.

³ If the reader is unfamiliar with the concept of boundary conditions and the logic of explanation within the context of synergetic linguistics, see Köhler (2005) for an introduction.

A specific feature of the Soviet cinema is that filmmaking and progress in film theory went closely hand in hand (cf. Albersmeier 1979: 9), at least in its early days during the 1920s and early 1930s. The question of how long a shot should be was discussed mainly by directors who were considered to belong to the avant-garde⁴ during the 1920s, and who continued to produce films and writings during the subsequent time of socialist realism. The articles and books selected in this paper were written during the first three decades of the Soviet Union. At first sight, this period might seem somewhat short, but taking the history of Soviet cinema into account⁵, this selection should suffice to provide an overview.

To capture the spectrum of the main theoretical concepts concerning shot length, texts by Sergej Ėjzenštejn, Lev Kulešov, Vsevolod Pudovkin, Semen Timošenko, and Dziga Vertov will be discussed in this article. Those texts cover the experimental phase during the 1920s, which is the time with the largest theoretical and practical diversity in the history of the Soviet cinema as well as the time of socialist realism. The following article will not be a chronological account of how thoughts on shot lengths developed, since such a detailed treatment is not possible within the given limited space. What is presented here is an outline, including main ideas determining shot length: comprehension by the audience, the intended emotional impact upon the viewer, the content and rhythm of the film, as well as the differentiation between descriptive and narrative shots.

Timošenko is discussed at the beginning. He is the first Russian theoretic, who calculated the average shot length of a scene to support his arguments and who addressed minimum shot length. The next chapter covers remarks by Pudovkin. When dealing with shot length, his emphasis was on rhythm and the distinction between descriptive and narrative shots. Pudovkin is also included here, because he wrote the first book on direction, a book that soon became the “bible” for filmmakers in many different countries. The following director and theoretic to be dealt with is Vertov. He has been included for his deviating concepts on mainstream narrative film, including his free rhythmic experiments with various shot lengths. His work influenced the style of generations of documentary directors. The fifth chapter looks at some concepts elaborated upon by Ėjzenštejn. With regard to the importance of his writings and the sophistication his ideas he should have occupied the first place. The reason for putting him in this position is to put emphasis on how much his ideas contrast with those of his predecessors so that the originality of Ėjzenštejn’s thoughts can be even more appreciated. Last but not least we will mention Kulešov. From all of his writings, the textbook from 1941 has been selected, because it reflects the officially recognized guideline for choosing the right shot length during the time of the socialist realism.

To properly assess the importance of Ėjzenštejn, Kulešov and Pudovkin for Soviet cinema, it is also necessary to mention that they all taught at the All-Russian State Institute of Cinema (VGIK, *Vserossijskij Gosudarstvennyj Institut Kinematografii*) or its predecessor the State College of Cinema (GTK, *Gosudarstvennyj tehnikum kinematografii*), which has been the first state film school in the world. (cf. Rollberg 2009: 735-736) Thus, chances are high that their attitudes have had an effect on generations of Soviet film students.

The following chapters start with a short introduction of the five directors highlighting some of their achievements in the realm of filmmaking and film theory. This is followed by a more detailed account of their writings on shot length. The issue of film length is taken up in

⁴ A prominent position in the theoretical discussion of film was also held by the Russian Formalists, whose articles were published in the book *Poëtika kino*, edited by Ėjchenbaum in 1927. In line with their novel approach to literature, the Russian Formalists proposed a new methodological approach to film studies as well. Among other things, one focus was on various formal features and how they are governed by regularities. Their approach was later on elaborated by Structuralists, decades after the publication from 1927. Since hardly any attention was paid by the Formalists to the issue of shot length, there is no need to pay closer attention to their writings at this place.

⁵ See Schmidt (2019: chap. 7) for a summary of Russian and Soviet film history as regards shot length in feature films.

some cases to complete the picture. Finally, instances are mentioned when experiments went wrong in connection with the length of shots and/or the lack of synchronization between the sound track and the visual track in order to preserve a fast and rhythmic editing style. Those failures are included, because they seem to provide an especially fertile ground for future investigations in quantitative film studies.

2. Semen Alekseevič Timošenko

Timošenko (1899-1958) was a director from St. Petersburg, who wrote also screenplays and acted in films. He is known for his sound comedies like *Tri tovarišča* (*Three Comrades*, 1935) and so-called film concerts like *Koncert na ekrane* (*Concert on the Screen*, 1935). Timošenko was also a film theorist with a concern for both practical and theoretical issues of filmmaking. (cf. Rollberg 2009: 693-694) His two major theoretical contributions are *Iskusstvo kino: montaž fil'ma* (*Art of the Cinema: Montage of a Film*, 1928) and *Čto dolžen znat' kino-režisser* (*What a Film Director Needs to Know*, 1929). Both books are the basis of the following summary on his ideas on shot length. The main points made by Timošenko relate to the audience, their comprehension time, potential weariness, and emotions. Furthermore, he includes formal statistics in his arguments: he uses the average shot length of a scene for illustration as well as the minimum shot length and he measures film length in the number of shots⁶. From the technical point of producing a film, he advocates a very clear and precise shooting script, including the length of every single shot.

Shot lengths are not an arbitrary issue to him. They are limited by the period of time, which is necessary for the audience to understand the content of the shot. This is his first criterion. Timošenko warns the director against making a shot longer than this minimum time, because if the shot is longer, the audience could notice dispensable trifles. (cf. Timoschenko 1928: 162)

The second point concerning the audience as well is given in the context of an exemplary film by Dziga Vertov. In his example Timošenko provides the shot length not in the unit 'metre' but in single frames⁷ – just like Vertov did himself. In the discussion of a certain scene, the number of shots are counted and Timošenko calculates their whole length as well as the average shot length. Timošenko recommends using this kind of fast editing pace only during important scenes. If a longer succession of many short shots appears repeatedly during a film, then the viewer becomes tired very quickly⁸. (cf. Timoschenko 1928: 174) Even though very short shots are comprehensible and even though very fast scenes are also comprehensible, it is not advisable to have too many of them in a film, because it would expect too much of the audience. Therefore, the second criterion is the ability of the audience to stay attentive.

The third and last criterion is related to the emotions of the audience whose arousal is the aim of every film. These emotions can be triggered if the shots are in a certain consecutive order and of a certain length. (cf. Timošenko 1929: 8) Unfortunately, Timošenko does not provide any more details on this issue.

⁶ In the context of illustrating one type of editing – analytical montage – Timošenko uses parts of a shooting script by a director named Piotrovskij. In the chosen fighting scene, the description of 30 shots from the shooting script are cited. First the scale of the shot is mentioned, which is followed by a short summary of what is shown in the shot and then the length of the shot is stated (cf. Timoschenko 1928, 170f.). The interesting part starts with the discussion: „The whole scene of the fight, which lasts for nine lines in the [...] narrative script, covers in the shooting script 94 shots, with an average of 5/8 metres per shot that is approx. 60 metres”. (Timoschenko 1928: 172; transl.: VS) To the knowledge of the author, this is the first calculation of average shot length. The two other statistics, minimum shot length and film length, measured in shots, are used in another context.

⁷ Timošenko points out that a single frame is 1/52 metres long (cf. 1928: 174).

⁸ His comments on rhythm and climaxes (cf. Timoschenko 1928: 197-200) might be an interesting historical source for the studies of CineMetrics graphs.

In his manual from 1929 “*Čto dolžen znat' kino-režisser*” (“*What a Film Director Needs to Know*”), Timošenko advocates using a »steel-like shooting script« (*stal'nyj scenarij*)⁹ that is a shooting script in which the most important details of every single shot are laid down¹⁰. Apart from the number of the shot, the camera angle, the content of the shot, the length of each shot is always included. Here, the shot length is measured in metres and not in seconds. (cf. 1929: 14f.) The idea behind the steel-like shooting script is that the planning and shooting of the film can be carried out as economically as possible. When talking about the shooting script, Timošenko emphasises that even the shortest shots should be included in this script, i.e. ones as short as 3 to 4 frames that is 0.19 to 0.25 seconds long¹¹. The shooting script of an ordinary film at that time included between 1,000 and 1,500 shots. (cf. Timošenko 1929: 31) These figures show that he was very well aware of the minimum shot length and of the average film length, for which he provides as a guideline a range of total shots.

To summarize the main points made by Timošenko. First, he is very much concerned about audience: A shot should last as long as the audience needs to understand what is depicted and not longer. However, since very fast editing puts a strain on the audience, it should not be employed too often and only at important scenes. Second, Timošenko is aware of the difficult economic situation and the importance of making good use of expensive film stock. This is one reason, why he is a strong advocate of planning a future film as thoroughly as possible, including the length of all single shots. Besides his theoretical thoughts on shot length, his writings show that he was also aware of the minimum shot length, the average shot length of scenes as well as the length of average films measured in the number of shots. As far as it is known, Timošenko is the first who has calculated the average shot length and used this data to support his arguments. No attention was spent on a possible classification of shots according to their content. The following theoretic was concerned with this issue.

3. Vsevolod Illarionovič Pudovkin

Pudovkin (1893-1953) was a Russian director, film theorist, screenwriter and actor. The two silent films he is most famous for are *Mat' (The Mother, 1926c)* and *Konec Sankt Peterburga (The End of St. Petersburg, 1927)*. He was the first to publish two books on film making: *Kino-scenarij (The Film Script, 1926b)* and *Kinorežisser i kinomaterial (Film Director and Film Material, 1926a)* (cf. Bulgakowa 1996: 239). This last treatise was translated in to numerous languages, including German in 1928 (cf. Rollberg 2009: 552-554).

Concerning the choice of shot length dependent on the content, Pudovkin differentiates between shots, which are bound to a narrative, and shots depicting a theme, which is not bound to some narrative element, i.e. descriptive or non-narrative shots. Generally, a film tells a story and single shots are embedded in the logical framework of time and space. Pudovkin uses as an example a short encounter of one person passing by another. All decisions as to the shot length for this scene are for him easily determinable: He starts with the glance of one person, ..., the passer-by raises his hat, ..., and ends with the leave of the passer-by. The tricky part starts with shots, which are not directly linked to a narrative that are shots, which depict some kind of theme. To illustrate his point, Pudovkin uses the images of a ‘sleeping German’ and a ‘factory’. In such cases, the length depends on the rhythm created by the length of the surrounding shots and this rhythm depends on the director. The ‘sleeping German’ for instance could be depicted for 6 seconds or for 12 seconds. The choice might seem somewhat arbitrary, however, a suitable

⁹ The „steel-like“ shooting script was introduced by Thomas Ince, an American director (cf. Beilenhoff 2005).

¹⁰ Sergej M. Ėjzenštejn was an opponent of the „steel-like“ script; he was in favour of an „emotional“ shooting script (cf. Eisenstein 2010b cited in Beilenhoff 2005).

¹¹ The conversion from frames to seconds was done on the basis of 16 frames per second.

duration can be found, if the shot is integrated into the whole artistic concept. (cf. Pudovkin 1983c: 68)

Without differentiating between narrative and non-narrative shots, Pudovkin emphasized the importance of rhythm numerous times. The function of rhythm is to evoke certain emotions within the spectator. In his book on film technique (1958a), he remarks that the succession of a large number of short shots – no clear statement of how long a “short” shot is and when a long shot starts is given in the text – leads to an excitement of the audience, whereas longer and calmer shots are more soothing or calming. (cf. 1958a, for instance 54, 105, 114, 131, 168). Finding a suitable rhythm that conveys the meaning is the essence of the director’s art (cf. 1958a: 169):

“Good editing will be achieved when for it is found the correct rhythm, and this rhythm is dependent on the relative lengths of the pieces, while the lengths of the pieces are in organic dependence on the content of each separate one. Therefore the director must enclose every shot he takes into a harsh, severely limited, temporal frame.”

So when photographing the various takes, the director has to determine beforehand how long the respective shot should be (cf. 1958a: 114) and make the actor adhere to very strict temporal limits, in which the given meaning has to be conveyed clearly. For Pudovkin, there is no room for chance: „The exactitude of work in space and in time is an indispensable condition, by fulfilment of which the film technician can attain a clearly and vividly impressive filmic representation” (1958a: 160). Thus, it appears that rhythm is the guiding principle when it comes to the determination of the suitable shot lengths.

Pudovkin is a clear advocate of fast editing. He advises the director and the scriptwriter to grab the attention of the viewer as strongly as possible. The viewer should not have the time to ponder over what he sees and there should be no time for critique or doubts. (cf. 1926b: 10) Just like Timošenko, Pudovkin emphasizes that fast editing of the whole film might exhaust the audience. Therefore, he also gives the advice of combining sequences with varying speeds (cf. Pudovkin 1983a: 311). In other words, cognitive abilities by the audience are clearly a limiting factor to be taken into account. As far as silent films are concerned, the above-mentioned issues cover Pudovkin’s ideas on the matter of shot length. The introduction of sound in the early 1930s made the theoretician ponder and reconsider.

The newly introduced sound track had been a challenge for most avant-gardist directors in the Soviet Union (cf. Rollberg 2009: 555). Whereas in silent films the director could use numerous short shots to convey the desired meaning, the freedom of a sound film director was more limited by fixed sequences (cf. Pudovkin 1983b: 80). Pudovkin laments the loss of the dynamic rhythm and richness of visual form due to the introduction of sound. Most of the sound films are “characterised by exceedingly slow development of subject and dialogue full of interminable pauses” and “explanatory words [are used] for matters that should be conceived visually” (Pudovkin 1958b: 194). For him this development is re-introducing the style of theatre back into the realm of film. The proponents of the new editing style argue that the ear reacts much slower to stimuli as the eye, which can handle a much faster alternation of visual stimuli. Therefore, the pictures have to adapt to the slower rhythm of the sound for the viewers to have enough time to comprehend. (cf. Pudovkin 1958b: 194f.) Accordingly, the technical innovation of a sound track meant that another sensory was predominantly being addressed now and its limitations, i.e. slower reaction time of the ear, set the tone from then on.

The last point is ‘intuition’. Just like any other artist, the director should be guided by his intuition when working on the shooting script, which breaks down the film to every single shot, its content and length (cf. Pudovkin 1958a: 114). This concept is somewhat contradictory to some of his other remarks regarding the planning of the whole film in other parts of the book.

His advice spans from the appropriate length of the whole film in dependence of the genre, to the suitable number of reels and the average number of shots within the single reels (cf. 1958a: 36).

Moreover, Pudovkin gives some practical advice for the planning of a film: The average length of a shot should be 6 to 10 feet or 6 to 9 seconds. On a single reel – which last for less than 15 minutes – should be between 100 to 150 shots. The whole film itself should be made of 6 to 8 reels. So the total number of shots in a feature film should range from 600 to 1200¹². These figures are a rough guideline for the scenarist to decide upon what is included in the film and how it fits into the whole scenario. (cf. 1958a: 73) From these unambiguous figures, it becomes clear that Pudovkin gave rather precise guidelines on how long a film and how long single reels ought to be and how many shots should be contained within the single reels.

Pudovkin's main points are: First, shot length depends on the kind of the shot: descriptive vs. narrative. The length of a descriptive shot is limited only by the rhythmic requirements of the respective scene. Things become a bit more complex when dealing with narrative content within the shot. In this case, both content, i.e. the movements of the actors, and rhythm determine the length. Second, cognitive limits by the audience come into play, when planning the whole film: Too many fast sequences tire the audience, so the director should be aware of the rhythmic development of the whole film and include fast and slow sequences. To keep the audience alert, their emotions can be aroused by the choice of suitable rhythms, which are connected, to certain shot lengths. The shorter the shots, the more excited the audience gets, the longer the shots are on average, the calmer the audience becomes. The third criterion relates to sound films. Since it takes more time to comprehend acoustic information than to comprehend visual information, shots with synchronous sound are inevitably longer. Taken all factors together: Shot length is determined by the rhythm, the content, the cognitive abilities by the audience and the differing processing speeds concerning visual and aural information.

4. Dziga Vertov

Vertov (1896-1954) is one of the founders of documentary film, for which he lay not only a theoretical but also a practical foundation. He worked as a director of newsreels and documentaries as well as a screenwriter. (cf. Rollberg 2009: 731-735) Along with the other Soviet film avant-gardists Ėjzenštejn, Kulešov and Pudovkin he established a new view on cinema with his contribution to the theory of documentaries. The three foundations of Vertov's ideas on cinema are his ideas of a) *kinoglaz (cine-eye)*, which aimed at the education of the people using facts, b) his demands to depict social processes with the help of cinema and c) his denunciation of narrative films. His methods of shooting and montage have also had a strong influence on the way documentaries were made outside the Soviet Union. To enhance the emotional impact upon the viewer, Vertov used for instance rhythm to structure montage or unusual camera angles. The film he is best known for is *Čelovek s kinoapparatom (Man with the Movie Camera, 1929)*, in which his ideas are put into practice rather playfully. (cf. Rollberg 2009: 733) When it comes to sound technology, his ideas are in line with other Soviet theoreticians like Ėjzenštejn or Pudovkin, i.e. a simple agreement between sight and sound was not desirable. The sound ocumentary *Ėntuziazm: Sinfonija Donbassa (Symphony of Donbass*

¹² These numbers are also in line with the data gathered from the silent films in the corpus of a study by Schmidt (2019). From the 13 silent films, 9 are well within this range of 600-1200 shots, 3 are slightly higher or lower, and one outlier exists with over 2000 shots.

also known as *Enthusiasm*, 1931) went down the same road as Pudovkin's first sound film: it alienated the spectators¹³. (cf. Rollberg 2009: 733)

As far as it is known, he did not give detailed advice as to a proper choice of shot lengths. In Vertov's articles, the author of this publication found only two instances in which this issue is addressed in more or less detail. Nonetheless, the metric table of a scene from one of his films shows the details of his editing style. The first instance of mentioning shot length is when he writes about the history of 'his' artist group *Kinoki*. Vertov recalls an early experiment, which was not hailed unanimously. The film *Boj pod Caricynom* (*The Battle before Caricyn*, 1920) was cut far too fast for the art council and the management of the studio, whereas some artists were pleased with it. (cf. Vertov 1973a: 82)

A variation of this problem should crop up again, years later. Even though Vertov tried to avoid the accusation of editing too fast, by conducting an experiment; this experiment was obviously not comprehensive enough. He discusses this instance in his article on instructions to the *Kinoglaz* community (cf. Vertov 1973b: 49f.). This is the second instance in which the issue of shot length is mentioned. The experiment was conducted with workers in the course of the documentary *Kino-glaz – 1-ja serija cicla »Žizn' vrasploch«* (*Cinema Eye – Part 1 »Life Caught Unawares«* 1924). During this test, workers saw a scene with a length of 17 metres that contained 53 shots, with some of the shots as short as $\frac{1}{4}$ seconds. The result of this experiment was that the spectators could comprehend the scene and did not get tired because of its fastness. (cf. Vertov 1973b: 49) Whereas the shortness of the shots did not seem to be a problem, the length of the whole film was. Vertov admits that the almost 80 minutes – which seemed to be the average length for narrative feature films at that time – were too tiring for the general audience. To avoid such problems in the future, he recommends letting the audience get used to it by showing shorter films first and then slowly enhancing their length as time goes on. (cf. Vertov 1973b: 49f.) The alternative of using sequences with longer shots and sequences with shorter shots, as is recommended by Pudovkin (cf. 1983a: 311), is not discussed in his articles.

Even though no direct reference as to Vertov's criteria of determining the lengths of the shots could be found in his writings, table 1 suggests that he might not have planned the shot length in advance, but when it came to the editing, he knew exactly what he was doing. The second shot of the above mentioned film *Kino-glaz* says that the film is made without a script. Table 1 (<http://www.cinematics.lv/vertov1.php>, last accessed October 02, 2012) depicts Vertov's plans for shots 28-40 of one scenes from his documentary *Kino-glaz* (1924). The content of the shots is summarized in the first column, the numbers of the individual shots of this scene are indicated in the first row and the numbers in the cells represent the length of the respective shots measured in frames¹⁴. So the shot depicting "Leader's head" is 10 frames long and it is the 35th shot of the scene.

¹³ This might make it a suitable object for further quantitative studies, because it crossed the border of easy comprehension.

¹⁴ If the projection speed is set at 20 frames per second then 0.5 seconds are needed to show these 10 frames.

Table 1

Metric table for a scene in the documentary *Kino-glaz* (1924) for shots 28-40 (reproduced from the table at <http://www.cinematics.lv/vertov1.php>, last accessed April 04, 2020).

numbers	28	29	30	31	32	33	34	35	36	37	38	39	40
1. Leader's head								10					
2. Girl Pioneer at the mast													
3. Mast and flag	10		10		9				10				
4. Intertitle: "Raise the flag!"													
5. Trumpeters							10						
6. Face no. 1											10		
7. Face no. 2												10	
8. Face no. 3													10
9. Face no. 4		10											
10. Hand shielding the sun													
11. Peasant children's hand raised				28									
12. Peasant woman with her hands raised						13							
13. Face no. 5										10			
14. Unit aligned													
15. Flag's shadow													
16. Feet of the girl Pioneer on duty													

The pattern from table 1 suggests that rhythm seems to be an important driving force behind the succession of the shots. Of all 13 shots in this metric table, ten are ten frames long, two are longer, and one is slightly shorter.

Even though his theoretical contributions to the question of shot length might be somewhat limited, his early films and the way they are edited provide a good example of the rhythmic possibilities with more or less non-narrative shots. The concept of rhythmic montage is being also used by the following filmmaker and theoretician, Sergej Ėjzenštejn.

5. Sergej Michajlovič Ėjzenštejn

Ėjzenštejn (1898-1948) was a director, teacher and theoretician on film. He is, as pointed out by Taylor (1996, x), “by general consent the most important figure in the history of cinema” and “[h]is contribution to the practice of film-making is universally acknowledged”. No other director has written as much on cinema and his own work as he did, and no other director has been so much at the centre of attention of other writers as Ėjzenštejn (cf. Taylor 2010: 1). Whereas other theorists like Pudovkin or Kulešov focused more on isolated details and their effect regarding local expressiveness and narrative construction, Ėjzenštejn had his eye on the stylistic organisation of the whole film. His global and organic perspective is one of his major contributions to film theory (cf. Bordwell 1996: 190). His famous theory of montage is just a piece of a more holistic puzzle. Ėjzenštejn was interested in a psycho-sociological theory of cinema, i.e. he combined psychological and social factors with the theory of film. Ėjzenštejn’s theoretical spectrum included ideas, which would be ascribed to systems theory today. Not only did he theorise, he also tested his ideas using strict empirical studies. One of his basic questions was: Which formal cinematic expressions cause which reaction(s) by the audience? (cf. Albersmeier 1979: 7-15)

Rollberg calls his legacy “gigantic in scope, multifaceted, and conceptually contradictory” (2009: 209). For those reasons the following pages can neither give a comprehensive account of what Ėjzenštejn wrote on shot length – some of his writings are still unpublished –, nor on how and why his ideas changed over time, keeping in mind that his own work remains unfinished and is also characterized by breaks (cf. Bulgakova 1996: 14). When reading Ėjzenštejn, these points should be kept in mind, as well the fact that his terminology is somewhat idiosyncratic. As Bulgakova (1996: 18) noticed: the Russian word for “shot” is “kadr”, sometimes, however, Ėjzenštejn prefers the German “Bildausschnitt”. The word “kusok”, which means something like “piece” or “bit” or “item” in English, is used not as a *terminus technicus*, but in a rather undifferentiated manner to denote a shot, a scene, or a sequence. After those reservations the following paragraphs shall nonetheless try to catch some concepts in the decision making process, which lead to a determination of how long a shot should be.

One central guiding principle in Ėjzenštejn’s work throughout his career was the aim to exert the maximum ideological influence on his audience, and all the parts of his filmic work are supposed to be subordinated to this goal (cf. Taylor 2010: 3f.). To achieve this goal he had a rather pragmatic approach: “Art admits *all* methods except those that fail to achieve their end.” (Eisenstein 2010c, 69, emphasis in the original) It was through the emotions of the audience, which were used as a gate, through which ideology could enter (cf. Taylor 2010: 5). One way to reach the audience is to present stimuli in ever new variations and to use their feedback as a guideline: “There is one thing we have no right to do and that is to *make generalisations*. The current phase of audience reaction determines our methods of influence: what it reacts to. Without this *there can be no influential art and certainly no art with maximum influence*.” (Eisenstein 2010c, 69, emphasis in the original) Therefore, in teaching the art of direction, Ėjzenštejn emphasised the value of creativity and that his pupils ought to be educated in such a way that developed and cultivated their inventive and creative spirits (cf. Ėjzenštejn 1974: 539).

In earlier texts, methods that are more precise are presented on how to arouse feelings by modifications of the shot lengths. In “The Fourth Dimension in Cinema” (1991) he discusses salient features, or *dominants* as he calls them, which form a basis for the combination of shots. One of those formal editing options is the shot length in what he calls *metric* montage. Here, the shots follow a certain repeating pattern when it comes to their length, like the time signature of a waltz or march. In other words, the length of the shot has precedence over the content. To increase suspense in a scene, the shots can be shortened as long as the underlying proportions (3/4, 2/4, etc.) are still in place. It is possible to fall back on a compound fraction, like 16/17, this is, however, not advisable in the eyes of Ėjzenštejn, since the psychological effect ceases to exist due to its complexity. An example of the combination between a simple and a compound metre is the Caucasian dance in his film *Oktjabr’ (October, 1927)*. (cf. Eisenstein 1991: 97ff.) This type of montage is in a way special to his approach to film, because the formal feature of shot length – the rhythm created by several shots through their lengths alone – is of paramount importance; the content of the shot is secondary.

A second editing technique, which is also directly tied to the length of the shots, is called *rhythmic* montage. In this case, the length of a shot is determined by both its content and the length of the adjacent shots. (cf. Eisenstein 1991: 99)¹⁵ When the goal is to raise formal suspense, then shots should become shorter and shorter. An example of this kind of montage, which has not to be discussed in detail here, since it is so famous, is the sequence of the “Odessa Steps” from *Bronenosec Potemkin (Battleship Potemkin, 1925)* (cf. 1991: 99f.).

¹⁵ According to the interpretation Bordwell, this means for instance that wide shots are shown for a longer time period than close-up shots (cf. 1996: 132).

Whereas metric and rhythmic montage are stylistic devices on a micro-level, Ėjzenštejn also made a more general distinction when it came to the function of shots: the depictive function and the rhythmic function. This basic differentiation relates to both the micro-level of certain scenes and to the macro-level of the film at large (cf. Eisenstein 2010a: 227-248). Ėjzenštejn argues that it is sometimes necessary to make shots as short as two or three frames for the sake of rhythm¹⁶ on the one hand, and on the other hand, the descriptive function demands an adequate length of a shot.

It is one thing to consider the rhythmic and depictive function in silent films; it is another thing in sound films. The introduction of sound changed the role of shots lengths concerning their time-bound rhythmic functions because sound could be partially used to take over this very rhythmic function. Another aspect of the introduction of sound is that visual rhythmicity – formerly expressed by certain patterns of shot length – is transferred to the composition within the shot. (cf. Eisenstein 2010a: 229) That is rhythm can now be expressed by the sound and visual rhythmic elements within the shot, like movements of the actors. The idea of not letting longer shots destroy the rhythm formerly produced by specific shot length, but having other elements take over this function, was an idea that seems novel to Ėjzenštejn; at least it was not found in the writings of other directors.

How these theoretical ideas were transferred into practice shows a quote from the composer Sergej Prokof'ev. Both Prokof'ev and Ėjzenštejn worked together during the production of *Aleksandr Nevskij* (*Alexander Nevsky*, 1938) and *Ivan Groznyj, 1 i 2 serija* (*Ivan the Terrible, Part I and II*, (1944) and 1945). Prokof'ev said, "Eisenstein's respect for music was so great that at times he was prepared to cut or add to his sequences so as not to upset the balance of a musical episode." (Schwarz 1972: 136 cited in Bordwell 1996: 220) To illustrate how detailed his plans of a sequence with regard to rhythm were, see Fig.1. This excerpt is an illustration of Ėjzenštejn aligning with each other musical phrases, pictorial composition and movements within the shot for the sake of rhythm.

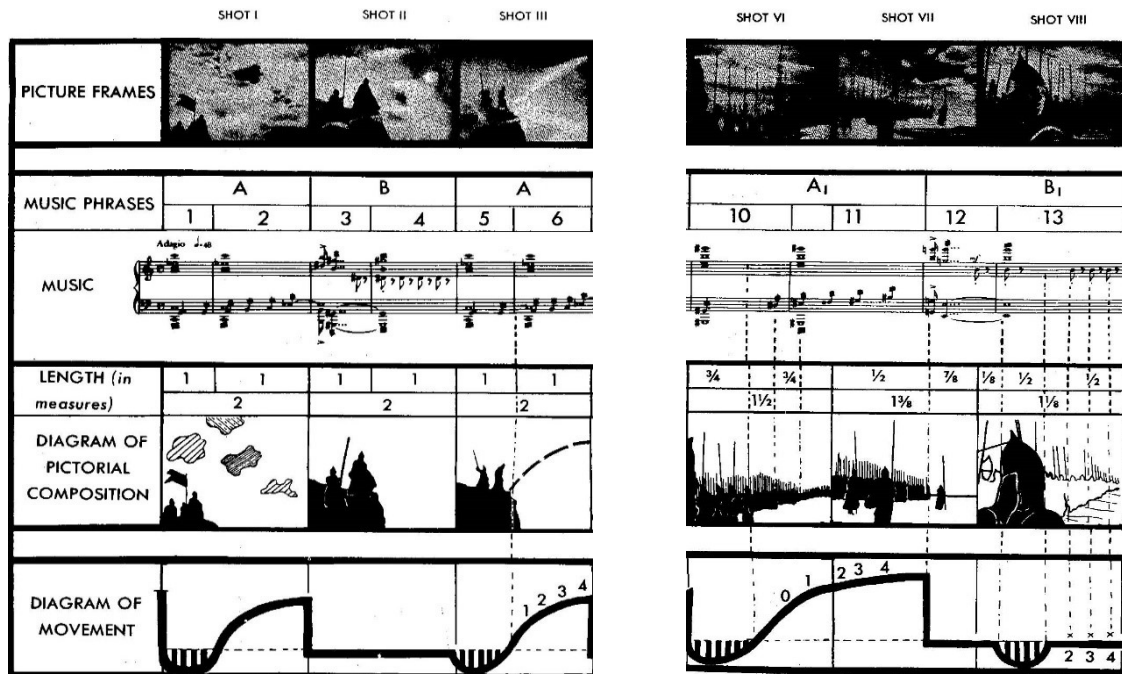


Fig. 1 Parts of the plan for a sequence from the film *Aleksandr Nevskij* (1938) (Leyda & Eisenstein 1957: 283-284).

¹⁶ Three frames are about as long as 0.1 second.

Two more ideas from Ėjzenštejn's writings shall be brought up, to show what is *not* determining the length of shots: censorship and creative ecstasy. A direct reference that formal matters were not entirely up to the director is given in 1940, when Ėjzenštejn discusses his film *Aleksandr Nevskij* (*Alexander Nevsky*, 1938) and the scene "Battle on the Ice". He complains that this battle takes not enough time, since he "was not allowed to cut 200 metres" (Eisenstein 1996: 140). To what extent censorship influenced general regularities of shot length has to be analysed elsewhere.

Creative ecstasy may come up during various stages of making a film. For instance, when the shots are being selected and their consecutive order is chosen. When it comes to shortening the shots, however, only skilfulness and knowledge of techniques are required so there is no more room for 'creative ecstasy'. (cf. Eisenstein 1991: 95) What exactly those skills and techniques are is not mentioned in the text and no thorough discussion has been found either, except the sources mentioned above.

To sum up, the first and foremost goal is the psychological effect on the audience, to wield the strongest possible ideological influence using all available methods. Since the arousal of emotions is related to fresh and new stimuli, the director has to find creative new ways of combining shots. Two minor methods, which are directly related to the length of the shots, are described by Ėjzenštejn: metric and rhythmic montage. A more general classification of shots is the distinction between the depictive function of a shot and the rhythmic function. Depending on the dominant function, the shots may be longer or shorter. The introduction of sound made it necessary that rhythm, which was formerly expressed by certain shot length patterns, was transferred to other levels, like the sound track or to movements within of the shot.

6. Lev Vladimirovič Kulešov

The Russian Kulešov (1899-1970) has been a pedagogue, film theorist, director and screenwriter. Many generations of Soviet directors were strongly influenced by his teachings; among his pupils were Boris Barnet, Michail Kalatozov or Vsevolod Pudovkin, who became famous directors afterwards. One aim of his early teachings in the 1920s was to create a new language of the cinema, which would reflect the spirit of the time. Strongly influenced by the American director David W. Griffith, Kulešov embraced a rhythmic and very dynamic style. Other features of American movies such as minimal psychological motivation of the characters, fast evolving storylines and numerous action sequences were also integrated into Kulešov's work. Some of his theoretical thoughts on film grammar were the basis for the more sophisticated ideas developed by Ėjzenštejn. His two most famous films are *Neobyčajnye priklučenija mistera Vesta v strane bol'shevikov* (*The Unusual Adventures of Mr. West in the Land of the Bolsheviks*, 1924) and *Po zakonu* (*By the Law* also known as *Dura Lex*, 1926). (cf. Rollberg 2009: 379-383)

Even though Kulešov published his first articles on film in 1917, five books over the years as well as over 60 articles (cf. Levaco 1974: 211-215), only a single book will be taken into account here. The reason for choosing his textbook *Osnovy kino režissury* (1995), first published in 1941, is that it is the very first official textbook in the Soviet Union and that it reflects the standards of socialist realism for film. Therefore, not so much Kulešov as the inventor and creative spirit is in the centre of attention here, but Kulešov the pragmatic teacher of directors during the time of socialist realism. According to Kulešov, the appropriate length of a shot can be determined by considering the following three basic factors. First, the content of the film in general and the theme of the respective scene in particular have to be considered. Second, the audience should be able to perceive the shots easily. The third and last factor is rhythm. These three factors will now be dealt with in detail.

The relation between content and shot length ought to be as such that everything unnecessary should be avoided. The story is supposed to unfold in such a way that shots depicting only necessary details are shown in a logical order one after another. (cf. 1995, 23f.) So there is no room for shots, which are not motivated by the theme of the whole film and which do not help the audience understand the story of the film.

The second criterion, ease of perception, is illustrated with the help of two pictures (Fig. 2 and 3) taken from page 275.

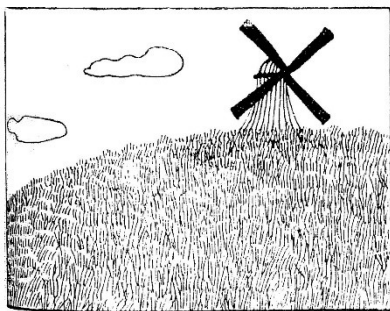


Fig. 2 Fast and easily comprehensible picture (Kulešov 1995: 275).



Fig. 3 Picture with a longer comprehension time (Kulešov 1995: 275).

Due to the lack of details, the first picture can be comprehended in a much faster time than the second picture with its numerous details like the flowers in the foreground, the birds between two hills, the bridge over the river, the various bushes and trees, etc. Therefore, in order to determine the appropriate shot length, the film maker should not only be aware of the purpose of the shot within the narrative of the whole film, but also take the perceptual difficulty of the shot into account. (cf. 1995: 275) A practical advice from Kulešov is to show motionless objects for at least 0.5 seconds (cf. 1995: 15). Just like Timošenko, Kulešov is a strong advocate of planning the future film as precisely as possible in advance (cf. 1995: 65). This includes planning the single shots not only with regard to their content, scale of shot, etc. but also with their respective durations. His experience has shown that directors commonly underestimate the length of the future shots by a fifth, so to the initial estimate one should add another 20% of time (cf. 1995: 122).

The last factor, rhythm, comes into play when planning and editing the single scenes (cf. 1995: 275). The aim of the director is to find the best possible rhythm that is also in tune with the play of the actors. Again, a suitable rhythm is one at which the audience can follow the film in the best possible way. (cf. 1995: 307) As an exemplary director he mentions the outstanding American director D.W. Griffith and he advises his students to watch his films and learn from him (cf. 1995: 30).

To sum up the main points made by Kulešov: the theme of the film and the rhythm of the scenes are two filmic factors, which determine the length of the single shots. The non-filmic factor, which always has to be kept in mind, is the ease of perception by the audience. Finally, the shots should be planned as precisely as possible in advance.

7. Summary

The choice of an appropriate shot length is related to numerous factors: cognitive abilities of the audience, rhythm and content.

Mental skills of the audience to recognize and interpret visual stimuli are taken into account by every director, except Ėjzenštejn. Timošenko refers to approximately 0.2 seconds as a

minimum, and Kulešov gives 0.5 seconds as a guideline for a non-moving object. For Ėjzenštejn this minimum was not a limit, because for his purposes he sometimes wanted not only reach the spectators on a cognitive basis – for which the minimum of the others would have sufficed – but he wanted to influence his audience on an emotional level via rhythm as well. To reach this level, he found it necessary to make some shots even shorter, so the content of the shot would not necessarily reach consciousness.

Attention span is also a limiting factor when it comes to rhythm. All presented directors are unanimous in regarding rhythm very highly, since it is one way to reach the audience on an emotional level. It is also a central factor, which determines shot lengths. If the rhythm is too fast for too long a period of time, i.e. the film has too many short shots, then the audience becomes weary, as Vertov had to find out the hard way. To avoid this problem, faster and slower parts should alternate with one another.

Another limiting factor is related to the sensory channels, the eye and the ear. Visual information like gestures or facial expressions can be processed faster by the audience. Therefore, these shots, whose content is conveyed via non-verbal communication, can be shorter than shots relying on the spoken word, simply because it takes more time to utter those words. This is one reason why shots became generally longer after the introduction of the sound. For some directors these longer shots posed a problem, because long shots would destroy the rhythm of a scene. Two paths were pursued, with one leading to a blind alley: asynchronous sound. Asynchronous sound was unsuccessful, since it expected too much of the audience. The more successful path led to a transfer of the rhythmic function from the level of the shot length to a) the music or sound track and/or b) to visual rhythmicity within the shot. This idea was proposed by Ėjzenštejn.

On the level of the content of the shot, two concepts are brought into play. Pudovkin differentiates between narrative and descriptive shots, which is similar to the stance taken by Ėjzenštejn, who distinguishes between the depictive and rhythmic function of a shot. Generally, narrative shots, or shots with a depictive function are more bound to the action within the shot, because their length is determined by the action. The length of descriptive shots or shots with a rhythmic function can be varied more freely, because factors that are not related to the composition within the shot are more dominant.

It is far beyond the scope of this paper to interpret some of the above mentioned criteria in terms of parameters in theoretical distributions like the Zipf-Alekseev, the negative binomial, the mixed negative binomial, the hyper-Pascal or the double exponential distribution. An answer to this lies in the mists of the future. However, as shown in the introduction, P. Grzybek lay the foundation of quantitative film studies. Future researchers interested in film and the beauty and elegance of synergetics, will find it a rewarding task to devote themselves to this new field. The task of gathering first data is extremely easy when using the data provided by the Cinemetrics database (<http://cinemetrics.lv/database.php> May 1, 2020) and the research design and research questions are easily derivable from quantitative linguistics.

Afternote

I'm sure, Peter would have welcomed every researcher with a warm "Feel at home" (<http://www.peter-grzybek.eu/index.html> May 1, 2020).

References

- Albersmeier, F.-J.** (1979). Filmtheorien im historischen Wandel. In: Albersmeier F.J. (ed.), *Texte zur Theorie des Films: 3-17*. Stuttgart: Reclam.
- Beilenhoff, W.** (ed.) (2005). *Poetika Kino. Theorie und Praxis des Films im russischen Formalismus*. Frankfurt am Main: Suhrkamp.
- Bordwell, D.** (1996). *The Cinema of Eisenstein*. Cambridge, MA: Harvard UP.
- Bulgakowa, O.** (1996). *Sergej Eisenstein - drei Utopien. Architekturentwürfe zur Filmtheorie*. Berlin: Potemkin Press.
- Eisenstein, S.** (1996). The Problem of the Soviet Historical Film. In: Taylor, R. (ed.), *S. M. Eisenstein. Selected Works. Volume III. Writings, 1934-47: 126-141*. London: British Film Institute. (transl. by William Powell)
- Eisenstein, S.** (2010a). Rhythm. In: Eisenstein, S., Glenny, M. Taylor, R. (eds), *Sergei Eisenstein. Selected Works. Volume II. Towards a Theory of Montage: 227-248*. London, New York: I.B. Tauris. (ed. by Michael Glenny and Richard Taylor, transl. by Michael Glenny)
- Eisenstein, S.** (2010b). The form of the script // O forme scenarija. In: Eisenstein, S. & Taylor, R. (eds.), *Selected Works. Volume I. Writings, 1922-34: 134-135*. London, New York: I.B. Tauris. (ed. and transl. by Richard Taylor)
- Eisenstein, S.** (1991). Die vierte Dimension im Film. In: Eisenstein, S. (ed.), *Das dynamische Quadrat: 90-108*. Leipzig: Reclam.
- Eisenstein, S.** (2010c). Constanta (Wither 'The Battleship Potemkin'). In: Eisenstein, S., Taylor, R. (eds), *Selected Works. Volume I. Writings, 1922-34: 67-70*. London, New York: I.B. Tauris. (ed. and transl. by Richard Taylor)
- Ėjchenbaum, B.M.** (ed.) (1927). *Poëtika kino*. Moskva, Leningrad: Academia.
- Ėjzenštejn, S.** (1974). Lehrprogramm für Theorie und Praxis der Regie. *Filmkritik* 18, 216, 538-569. (transl. by Hans-Joachim Schlegel and Gabriele Hübner)
- Ėjzenštejn, S.** (1925). *Bronenosec Potemkin // Battleship Potemkin*. Sovkino. Soviet Union.
- Ėjzenštejn, S.** (1944). *Ivan Groznyj I // Ivan the Terrible, Part I*. Mosfil'm COKS. Soviet Union.
- Ėjzenštejn, S.** (1945). *Ivan Groznyj II // Ivan the Terrible, Part II*. Mosfil'm COKS. Soviet Union.
- Ėjzenštejn, S. & Aleksandrov, G.** (1927). *Oktjabr': Desjat' dnej kotorye potrjasli mir // October: Ten Days that Shook the World*. Sovkino. Soviet Union.
- Ėjzenštejn, S. & Vasil'ev, D.** (1938). *Aleksandr Nevskij // Alexander Nevsky*. Mosfil'm. Soviet Union.
- Grzybek, P. & Koch, V.** (2012-2013). *Quantitative Analyse jugoslawischer Filme*. <http://www.peter-grzybek.eu/science/projects/index.html>. April 04, 2020.
- Grzybek, P. & Koch, V.** (2012). Shot Length: Random or Rigid, Choice or Chance? An Analysis of Lev Kulešov's "Po zakonu" [By the Law]. In: Hess-Lüttich, E. (ed.), *Sign Culture. Zeichen Kultur: 169-188*. Würzburg: Königshausen & Neumann.
- Köhler, R.** (2005). Synergetic linguistics. In: Köhler, R., Altmann, G. & Piotrowski, R. G., (eds), *Quantitative Linguistik: Ein internationales Handbuch // Quantitative Linguistics: 760-774. An International Handbook*. Berlin, New York: Gruyter.
- Kulešov, L.V.** (1924). *Neobyčajnye priklučenija mistera Vesta v strane bol'shevikov // The Unusual Adventures of Mr. West in the Land of the Bolsheviks*. Goskino. Soviet Union.
- Kulešov, L.V.** (1926). *Po zakonu // By the Law*. Goskino. Soviet Union.
- Kulešov, L.V.** (1995 [1941]). *Osnovy kino režissury. Reprintnoe izdanie*. Moskva: Vserossijskij gosudarstvennyj institut kinematografii im. S.A. Gerasimova.

- Levaco, R.** (ed.) (1974). *Kuleshov on Film. Writings by Lev Kuleshov*. Berkeley, Los Angeles, London: University of California Press. (sel., transl. and ed., with an introduction by Ronald Levaco)
- Leyda, J. & Eisenstein, S.** (ed.) (1957). *The Film Sense*. New York: Meridian. (ed. and transl. by Jay Leyda)
- Pudovkin, V.** (1926a). *Kinorežisser i kinomaterial*. Moskva: Kinopečat'.
- Pudovkin, V.** (1926b). *Kino-scenarij. Teorija scenarija // The Film Script. The Theory of the Script*. Moskva: Kino-izdatel'stvo R.S.F.S.R.
- Pudovkin, V.** (1926c). *Mat' // Mother*. Mežrabpom-Rus'.
- Pudovkin, V.** (1927). *Konec Sankt Peterburga // The End of St. Petersburg*. Mežrabpom-Rus'.
- Pudovkin, V.** (1958a). Film Technique. In: Pudovkin, V. & Montagu, I. (eds.), *Film Technique and Film Acting. Memorial Edition: 19-164*. London: Vision Mayflower. (transl. and ed. by Ivor Montagu)
- Pudovkin, V.** (1958b). Rhythmic Problems in my First Sound Film. In: Pudovkin, V. & Montagu, I. (ed.), *Film Technique and Film Acting. Memorial Edition: 194-202*. (transl. and ed. by Ivor Montagu). London: Vision Mayflower.
- Pudowkin, W.** (ed.) (1928). *Filmregie und Filmmanuskript*. Berlin: Lichtbildbühne. (transl. by Georg and Nadja Friedland)
- Pudowkin, W.** (1983a). Die Zeit in Großaufnahme. In: Pudowkin, W. (ed.), *Die Zeit in Großaufnahme. Aufsätze, Erinnerungen, Werkstattnotizen: 305-312*. Berlin: Henschelverlag.
- Pudowkin, W.** (1983b). Ton und Bild. In: Pudowkin, W. (ed.), *Die Zeit in Großaufnahme. Aufsätze, Erinnerungen, Werkstattnotizen: 80-81*. Berlin: Henschelverlag.
- Pudowkin, W.** (1983c). Wie wir den Film "Das Ende von St. Petersburg" produzierten. In: Pudowkin, W. (ed.), *Die Zeit in Großaufnahme. Aufsätze, Erinnerungen, Werkstattnotizen: 64-38*. Berlin: Henschelverlag.
- Rollberg, P.** (2009). *Historical Dictionary of Russian and Soviet Cinema*. Lanham / Maryland, Toronto, Plymouth / UK: The Scarecrow Press.
- Schmidt, V.** (2019). *Quantitative Film Studies. Regularities and Interrelations Exemplified by Shot Lengths in Soviet Feature Films*. Hamburg: Kovač.
- Schwarz, B.** (1972). *Music and Musical Life in Soviet Russia, 1917-1970*. New York: Norton.
- Taylor, R.** (1996). General Editor's Preface. In: Taylor, R. (ed.), *S. M. Eisenstein. Selected Works. Volume III. Writings, 1934-47: x-xi*. London: British Film Institute. (transl. by William Powell)
- Taylor, R.** (2010). Introduction. In: Sergei Eisenstein & Richard Taylor (ed.), *Selected Works. Volume I. Writings, 1922-34: 1-24*. London, New York: Tauris. (ed. and transl. by Richard Taylor)
- Timoschenko, S.** (1928). Filmkunst und Filmschnitt // Iskusstvo kino i montaž fil'ma // Cinema Art and the Montage of a Film. German Parts of the Russ. Original. In: Pudowkin, W. (ed.), *Filmregie und Filmmanuskript: 148-204*. Berlin: Lichtbildbühne. (transl. by Georg and Nadja Friedland)
- Timošenko, S.** (1929). *Čto dolžen znat' kino-režisser // What a Film Director needs to Know*. Moskva. Leningrad: Tea-kino-pečat'.
- Timošenko, S.** (1935). *Tri tovarišča / Three Friend*. Lenfil'm. Soviet Union.
- Vertov, Dz.** (1920). *Boj pod Caricynom // The Battle before Caricyn*. unknown studio.
- Vertov, Dz.** (1924). *Kino-glaz - I-ja serija cicla »Žizn' vrasploch« // Cinema Eye - Part I »Life Caught Unawares«*. Goskino. Soviet Union.
- Vertov, Dz.** (1929). *Čelovek s kinoapparatom // The Man with a Movie Camera*. VUFKU. Soviet Union.

Theoretical Thoughts and Practical Advice on the Length of Shots

- Vertov, Dz.** (1931). *Éntuziazm: Simfonia Donbassa // Enthusiasm aka Symphony of Donbass*. Ukrainfilm. Soviet Union.
- Vertov, Dz.** (1973a). Aus der Geschichte der Kinoki. In: Beilenhoff, W. (ed.), *Dziga Vertov: Schriften zum Film*: 82-88. München.
- Vertov, Dz.** (1973b). Vorläufige Instruktion an die Zirkel des "Kinoglaz". In: Beilenhoff, W. (ed.), *Dziga Vertov: Schriften zum Film*: 41-53. München.

Sentence and Paragraph in the Light of Menzerath-Altmann's Law

Volker Gröller¹

Abstract

The present study focuses on the relation between paragraph and sentence from the perspective of Menzerath–Altmann's law. It can be shown that sentence length is a function of paragraph length. For paragraphs, different properties can be observed when measuring its length in sentences or in words, the latter yielding a smoother, simpler fitting curve.

Keywords: Menzerath-Altmann's law; paragraph; sentence

1. Introduction

The linguistic law of the relation between language constructs and their components was first discovered by Paul Menzerath and later formalized by Gabriel Altmann. The Menzerath–Altmann law became one of the most important laws in quantitative linguistics (Altmann & Schwibbe 1989). The question which will be discussed here is whether Menzerath–Altmann's law holds also for the relation between sentence and paragraph length. While an impact of text length on at least some linguistic units has been proven (Jin 2017), the paragraph was, until now, not analyzed in the light of Menzerath–Altmann's law. It is at times seen as something beyond grammar, maybe something too big for traditional linguistic analysis. It is rather associated with the text on some higher level. Altmann (2014) discusses supra-sentential structures of text, leaving out the paragraph. Therefore, introducing the paragraph as a relevant construct to the component sentence could be of further theoretical and empirical interest. The question is whether Menzerath–Altmann's law also holds true for “bigger” language constructs.

In the next chapter we will briefly describe Menzerath–Altmann's law and its development. Afterwards we give an overview of the hypothesis and the way we measure linguistic units, which will be followed by the actual analysis and finished up by some conclusions.

2. A short overview of Menzerath–Altmann's law

Menzerath-Altmann's law describes the relation between the size of a language construct and its components. It was first observed by Paul Menzerath in his studies about sound lengths in Spanish, where he writes (Menzerath/de Oleza 1928: 68):

Die relative Lautzahl nimmt mit steigender Silbenzahl ab, oder mit anderer Formel

¹ volker.groeller@gmail.com.

gesagt: je mehr Silben ein Wort hat, um so (relativ) kürzer (lautärmer) ist es².

So, the more syllables a word has, the lower is its relative number of sounds. This and similar observations motivated P. Menzerath to formulate some “quantitative laws” (“Quantitätsgesetze”). Note that they are still linear in nature, and predict smaller sizes of components, the larger the construct. Menzerath already mentions irregularities in the monotony of the size relations though (ibid.: 68). He later reformulated his quantitative laws into a more general statement, which he termed “Sparsamkeitsregel”, or in English “rule of economy (brevity)” (Menzerath 1954: 101):

(...) je größer das Ganze, umso kleiner die Teile!

I.e., the greater the whole, the smaller its parts. Some decades later Gabriel Altmann later expanded on Menzerath's studies and formalized them into a mathematically based hypothesis (Altmann 1980: 1). His starting point was the (simple) formula:

$$\frac{y'}{y} = -c$$

It represents the rule of brevity. The relation of the construct “y” and its component “y” is described by the negative variable “c”, to account for decreasing lengths. Transformation and integration yields:

$$y = a e^{-cx}$$

Here, “x” is an expression of the language construct “X”, and “y” of X's component “Y”. Adding an additional element to the original formula allows non-monotonic relations to be described as well:

$$\frac{y'}{y} = -c + \frac{b}{x}$$

Thus Menzerath–Altmann's law was “born”. Its definition being (ibid.: 3):

The length of the components is a function of the length of language constructs.

Two things are noteworthy here concerning the definition: Firstly, there are no details given as to the nature of the relation between the construct and its components. The monotonous nature of the rule of brevity was abandoned, and gives way to more complex relations. Secondly, this definition also allows constructs to be related with their indirect components. This will become relevant in the later stages of our analysis.

Going back to the mathematical representation, Altmann's work results in three possible formulae, which were expanded upon further (Wimmer & Altmann 2005), yielding in total six possible functions (taken from Grzybek 2011: 66):

I	$y = K * e^{ax}$	$a < 0 ; b, c = 0$	2
II	$y = K * x^b$	$b < 0 ; a, c = 0$	2
III	$y = K * e^{ax} * x^b$	$a, b \neq 0 ; c = 0$	3
IV	$y = K * e^{-cx}$	$c > 0$	2

² (The relative number of sounds decreases with increasing number of syllables, or, to put it in other words: the more syllables a word has, the (relatively) shorter it is (the fewer sounds the individual syllables have).

V	$y = K * x^b * e^{c/x}$	$b, c \neq 0$	3
VI	$y = K * e^{ax} * x^b * e^{-c/x}$	$a, b, c \neq 0$	4

Note that formulae I through V can be considered as special cases of VI. In that sense, Menzerath's earlier quantitative laws and rule of brevity can be considered to be marginal cases of Menzerath–Altmann's law.

3. Units of measurement, hypothesis

Abiding by the just discussed definition of Menzerath–Altmann's law, we can formulate our hypothesis as follows:

The length of sentences is a function of the length of paragraphs.

We predict that increasing paragraph length affects the length of sentences, which compose these paragraphs in a systematic way, that can be described by one of the six formulae presented above. In order to prove this hypothesis, the Russian novel “Prestuplenie i nakazanie” (*Crime and Punishment*) (1866) by Fyodor Dostoevsky was chosen. For every paragraph within it, the number of sentences and words is measured, after which the paragraph and sentence length is calculated.

The length of sentences will be measured in words. It has to be mentioned here that counting sentence length in clauses has already been successfully done (Buk & Rovenchak 2008, Sanada 2016), and this makes more “sense” from the perspective of Menzerath–Altmann's law. Still we shall stick to measuring sentences in words for this analysis, as it is not only easier to compute, but the unit of word is also more stable and simplistic in its definition(s) than the clause.

Paragraph length will be measured in sentences as well as in words. The latter is done to normalize paragraph length: Most words will not exceed a length of five to ten letters (or 2-4 syllables), sentences however can be comprised of one word or 150 and more. Obviously measuring paragraphs in sentences is more in the spirit of Menzerath–Altmann's law, while counting them in words allows for a more controlled measurement of length.

Below are the definitions of word, sentence and paragraph, as used in this analysis:

A word is a sequence of alphanumeric characters.

A sentence is a sequence of characters bounded by ".", "!" or "?".

A paragraph is a sequence of characters bounded by an empty line.

The text we used for the analysis is Russian, so words that do not contain vowels were not considered as full words (Antić, Kelih & Grzybek 2006: 126), but these words (mainly prepositions) are interpreted as clitics. As for the sentences, in addition to the three standard sentence markers, the semicolon (;) and ellipsis (...) might be considered as well. We did not include them, as although they do both indicate an interruption in the sentence structure in some way, they arguably do not always end the sentence entirely.

There are, of course, other ways to measure paragraphs. An obvious one is stopping the paragraph at any line break, even if there is no empty line following it. There cannot be any strong argument against this way of counting, but we still decided against it, as an empty line

is the clearest sign of a paragraph ending.

Here might be a good point for a small digression concerning the choice of language constructs as a whole. As very clearly stated by Altmann (2014), there are no naturally occurring language constructs as such. Language is human, and so are the units to measure and analyze. And, while they grasp some of the reality that forms the structures of language, they miss many others. In the end we always observe Menzerath–Altmann's law through a hazy glass window of man-made language constructs. Therefore, we argue that while choosing the right language constructs for quantitative analysis is important, it is not essential.

4. Analysis

Counting the paragraphs, sentences and words of the selected book as described above yields 3,537 paragraphs, 13,486 sentences and 160,504 words. As the frequency of sentences and paragraphs declines rapidly with decreasing size, paragraphs with fewer than 100 words are pooled together. This avoids giving statistical outliers too much weight during curve fitting.

As already mentioned, paragraph length was measured in sentences as well as in words. For both approaches, all six formulae were tried for fitting, choosing the one with the least parameters where possible, in accordance with Occam's razor. Two graphs with the best fitting functions can be found below.

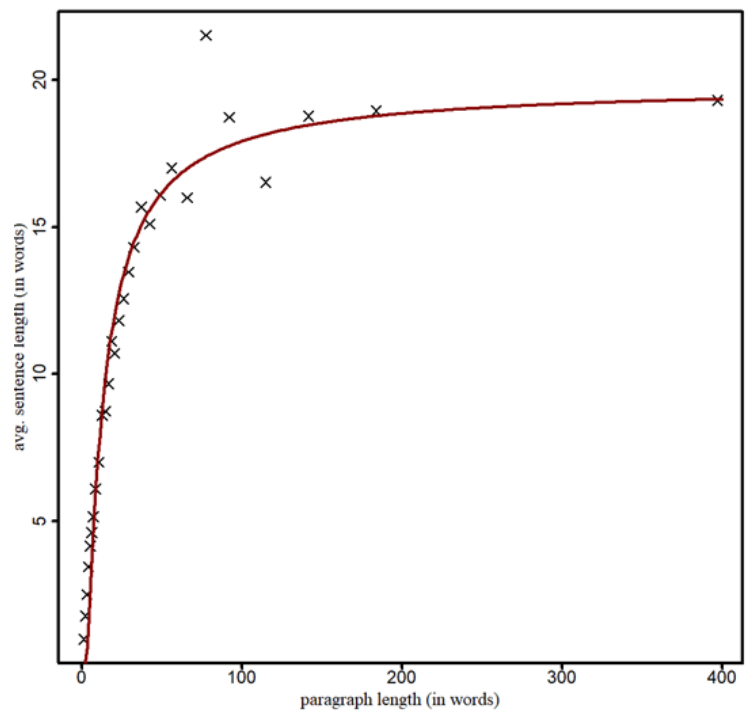


Figure 1 Paragraph length in words vs. sentence length (formula IV: $R^2 = 0.9577$, $K = 19.8411$, $c = 10.210$)

Comparing the two graphs makes it clear that measuring paragraphs in words rather than sentences gives a much smoother, “clearer” picture. Accordingly, the used formula IV in the first graph has only two parameters, which makes it easier to describe and to interpret. In comparison, if paragraphs are measured in sentences, the fitted function has four parameters, with lower goodness of fit.

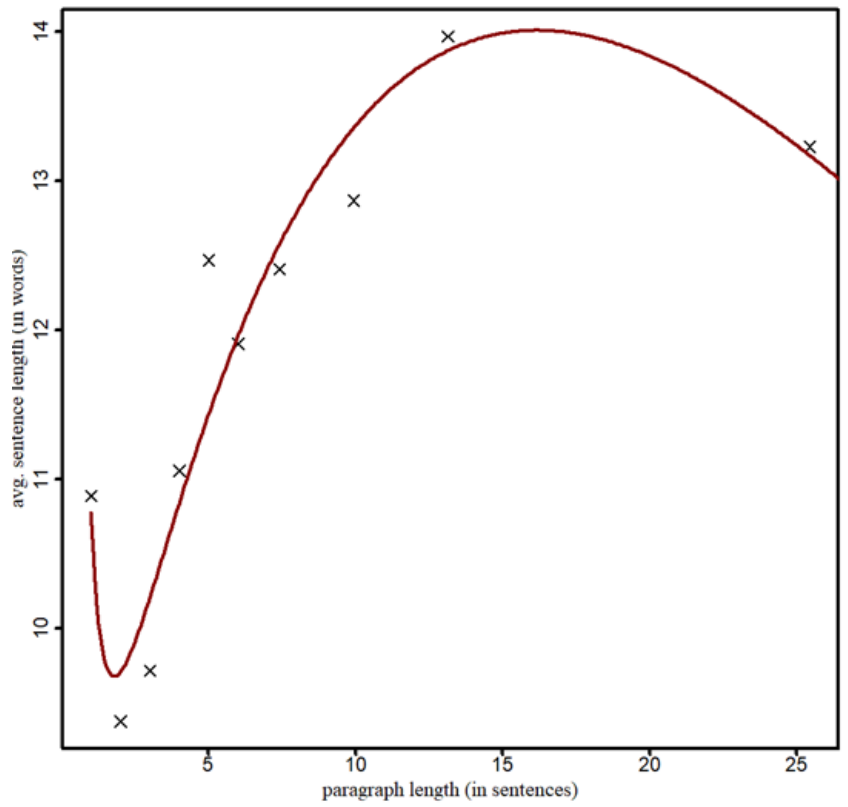


Figure 2 Paragraph length in sentences vs. sentence length (formula VI: $R^2 = 0.9135$, $K = 4.1176$, $a = -0.0344$, $b = 0.6179$, $c = -0.996$)

Thus, measuring paragraphs in words rather than sentences yields a simpler and stronger relation.

5. Conclusion

It could be shown that for paragraphs measured in words, our hypothesis holds, and so does Menzerath–Altmann's law for the relation of sentence and paragraph. Sentence length is a function of paragraph length.

It also shows the usefulness of formula VI to describe complex relations. Language constructs with a length/size ratio of 1 often seem to have special properties, which sometimes might only be described properly by that formula. Figure 2 shows this clearly.

Moving down the hierarchy of language constructs below the phoneme might prove impossible, moving up higher however should be an interesting test of Menzerath–Altmann's law. The length of chapters measured in paragraphs, or the lengths of books measured in chapters, are just two possibilities that spring to mind.

References

- Altmann, G.** (1980): Prolegomena to Menzerath's law. In: Grotjahn, R. (ed.): *Glottometrika 2: 1-10*. Bochum: Brockmeyer. (Quantitative Linguistics, 3)
- Altmann, G. & Schwibbe, M.** (eds.) (1989): *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Zürich, New York: Hildesheim.
- Altmann, G.** (2014): Supra-sentence levels. *Glottology 5, 1*, 25–39.
- Antić, G.; Kelih, E. & Grzybek, P.** (2006): Zero-syllable Words in Determining Word Length. In: Grzybek, P. (ed.): *Contributions to the Science of Language. Word Length Studies and Related Issues: 117-156*. Boston: Kluwer.
- Buk, S. & Rovenchak, A.** (2008): Menzerath-Altmann Law for Syntactic Structures in Ukrainian. *Glottology 1, 1*, 10-17.
- Grzybek, P.** (2011): Der Satz und seine Beziehungen. I: Satzlänge und Wortlänge im Russischen (Am Beispiel von L.N. Tolstojs «Анна Каренина»). *Anzeiger für Slavische Philologie 39*, 39-74.
- Jin, H. & Liu, H.** (2017): How will text size influence the length of its linguistic constituents? In: *Papers and Studies in Contrastive Linguistics 53, 2*, 197-225. DOI: 10.1515/psicl-2017-0008.
- Menzerath, P. & de Oleza, J.M.** (1928): *Spanische Lautdauer. Eine experimentelle Untersuchung*. Berlin: de Gruyter.
- Menzerath, P.** (1954): *Die Architektur des deutschen Wortschatzes*. Bonn: Dümmler. (Phonetische Studien, 3)
- Sanada, H.** (2016): The Menzerath-Altmann Law and Sentence Structure. *Journal of Quantitative Linguistics 23, 3*, 256-277.
- Wimmer, G. & Altmann, G.** (2005): Unified derivation of some linguistic laws. In: Köhler, R.; Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 791-801*. Berlin: de Gruyter. (Handbücher zur Sprach- und Kommunikationswissenschaft, 27)

The Impressive Story of a Unique Collaboration

Ernst Stadlober¹

Abstract

In this paper, we discuss the long and very fruitful close collaboration between Peter Grzybek and our working group. Together we investigated different texts of three Slavic languages and obtained interesting results concerning the distribution of word length. An important feature was the classification of texts in different text types. This categorization opened the opportunity to discriminate texts. We also studied the relationship between word length and sentence length. In our last project we considered and described the characteristics of shot durations in silent movies by using lognormal distributions.

Keywords: word length; sentence length; shot duration in films; lognormal distribution

1. Introduction

Our collaboration started with an e-mail from Peter Grzybek asking about statistical details in a text sequence. As I recall, this was around 1999 and 2000. Together with a colleague, I carried out some statistical analyses of texts. Peter was very satisfied with the results and eager to stay in contact working together. To continue our collaboration Peter had the idea to start with a project on *word length (frequencies) in Slavic texts*. In order to get the project funded, we formulated a research proposal for the Austrian Fund for Scientific Research (FWF). Our application was successful and we received funding for three years (see the description of the project in Grzybek & Stadlober 2002). Hence, we were able to engage a group of students from different fields of study: quantitative linguistics, text scholarship, computer science and statistics. As a fortunate coincidence, Gordana Antić (now married Djuraš) from Novi Sad applied, at the same time, for a PhD study in Statistics at the Graz University of Technology. She was admitted to the PhD program and engaged in our project.

The aim of the project concentrated on three crucial aspects: (i) a systematic study of word length frequency distributions in texts of the three Slavic languages Croatian, Russian and Slovenian, (ii) studying an individual author's style and text typology, i.e. finding factors which influence word length, (iii) applying the knowledge obtained in (ii) to attribute individual texts to specific authors and text types.

The first necessary step was to create a text database with around 1000 texts in each of the three languages to be studied. Until the year 2010 the so-called QuanTA database included 1100 Croatian texts, 1800 Russian texts, 1600 Slovenian texts, extended by 600 Serbian and 800 Slovakian texts. The next task consisted in preparing the texts for the analysis. This included a unitary treatment of abbreviations, headings, numbers, foreign words and the definition of a sentence. Then, statistical analysis provided the raw data for each text stored in an adequate file format. After these efforts, we were able to find distribution models together with factors influencing the parameters of the model, resulting in text classification and text discrimination. Some more details will be discussed in section 2.

Peter was an excellent organizer of conferences. He coordinated two very successful symposia in the castle of Seggau, south of Graz, (Quantitative Textanalyse, 2002 and The Science of Language, 2005). Especially two aspects of these conferences were outstanding: the profound professional discussions after the scientific speeches and the relaxed conversations on the

¹ Institute of Statistics, Graz University of Technology Kopernikusgasse 24/III, A-8010 Graz, Austria, e.stadlober@tugraz.at.

walk to the Buschenschank (typical Austrian wine tavern) nearby, where we enjoyed excellent Styrian wine on warm summer nights under the full moon. The conference Qualico 2009 took place in Graz where Peter had booked rooms in an intimate monastery for both the conference talks and the accommodation.

He got always very creative when it came to booking extraordinary accommodations (pension, farm house, rustic guesthouse) and he was an excellent trip adviser in his German homeland (former German borderline, Harz, Berlin-Dahlem, Black Forest). In 2003 we visited Göttingen (*Graz-Göttinger Gespräche zur quantitativen Linguistik*) and presented our ongoing research results in a series of conferences (Dortmund 2004, Magdeburg 2005, Berlin 2006, Freiburg 2007), all organized by the *Gesellschaft für Klassifikation* that offered special sessions on quantitative linguistics. We also participated in the conference Qualico 2012 in Belgrade and in the Qualico 2016 in Trier, where we gave our last joint conference speech about quantitative film analysis. The results of these investigations will be given in section 4.

2. Word Length Studies

2.1. Discrimination and Classification

We started with a case study of 153 Slovenian texts (51 poems, 52 literary prose texts, 50 journalistic texts) (see the diploma thesis of Djuzelic 2002 and Stadlober & Djuzelic 2005). The distribution of the word lengths, measured by the number of syllables, is described by a number of characteristics reflecting the moments of the distribution. By means of discriminant analysis, we gave answers to the following questions: Is it possible to discriminate between the texts with the help of the measures such that most texts are assigned to the right category? Which characteristics determine the classification? The answers to these questions can be formulated as follows:

The distribution of the word length is described by m_1 (average word length), m_2 (variance of the word length), $I = m_1/m_2$ (1st criterion of Ord) and $S = m_3/m_2$ (2nd criterion of Ord, m_3 the third moment). The text length is given by TLS (text length in syllables) and $\log(TLS)$. All three categories of texts can be compared by canonical discrimination. The best results were obtained by establishing canonical discriminant functions Z_1 and Z_2 based on the three most important variables m_1 , I and $\log(TLS)$. It turned out that the variance between the groups is much higher than the variances within the groups. Hence, both functions Z_1 and Z_2 are good measures for the separation of the categories. This procedure was able to classify all 52 literary texts correctly. Only one of the 50 journalistic texts was wrongly assigned to poetry. Additionally, two of the 51 poetic texts were misclassified: one as journalistic text, the other one as literary text. Overall 98% of the texts were classified correctly.

Another case study of 398 Slovenian texts (see Grzybek et al. 2005) considers texts from different genres and authors. Four relevant predictor variables, characterizing word length distribution, were part of the discriminant functions used to discriminate all eight text sorts (poems, short stories, private letters, drama, epistolary novel, reader's letters, comments, open letters). However, due to many overlapping features, only 56.3% of the texts were correctly discriminated.

So it was necessary to establish a new typology for two major groups (private letters–epistolary novel) and (open letters–readers' letters) which were as a first attempt treated as special cases of private–everyday style and public–official style. The reintroduction of journalistic comments to the group of public texts did not decrease the good discrimination result where 91.6% of the 248 texts were correctly discriminated.

In the end, the analysis of eight text sorts partitioned into three categories (unofficial–oral,

with private letters, epistolary novel, drama, short stories = literary prose), (official-written, with open letters, readers' letters, comments), (poetry with poems) lead to 92.7% correctly allocated texts. These results suggest the existence of specific discourse types different from traditional functional styles.

2.2. Modeling Word Length Frequencies

The problem of modeling the distribution of word length was of general interest to scientists in different areas: linguistics, physics, mathematics and statistics. De Morgan, an English mathematician discussed word length as characteristic to determine authorship. At the beginning of our studies we paid a lot of attention to the work of Fucks (e.g. Fucks 1956), a German physicist who introduced a mixture of Poisson probabilities, known as Fuck's Generalized Poisson distribution to describe word formation through syllables. The contribution of Antić et al. 2005 is a detailed systematic study of the mathematical and statistical aspects of this approach. It turned out that the models of Fucks, despite of their complexity, are only applicable to particular types of empirical distributions, hence they cannot serve as overall models for language. The paper presents theoretical and practical drawbacks of Fucks's GPD.

This is why our research focused on useful generalizations of the Poisson distribution which would be able to fit a broad range of word length frequency distributions. Our aim was to find a general model able to cover the whole range of the index of dispersion (i.e. the variance to mean ratio) $\delta = \sigma^2/(\mu - 1)$ and estimated by its empirical value $d = s^2/(\bar{x} - 1)$. We were seeking to create a model with at most two parameters and unique for all texts to study.

Three generalizations of the Poisson distribution fulfilled these requirements: the Generalized Poisson distribution (GPD), the Hyper-Poisson distribution (HPD) and the Singh-Poisson distribution (SPD) (Djuraš & Stadlober 2010, Djuraš 2012, Djuraš et al. 2013a, Djuraš et al. 2013b). The analysis of 120 Slovenian and 120 Russian texts and a simulation study to investigate the stability of the parameter estimates for both the GPD and the SPD showed some benefits for the SPD. This model was able to better discriminate four different text types (private letters, journalistic, poems, prose). Also the text types could be quantified and characterized by adequate parameter regions. Moreover, the calculation of the maximum likelihood estimates was a very simple task: it coincides with the method based on the sample mean and the first frequency class. Additionally, the simulation study demonstrated the usefulness of the parameter estimates under the three data-driven dispersion scenarios. In contrast, the results for the HPD in Djuraš 2012 simulation study showed unreliable estimates for all three different estimation procedures. For this distribution also the estimation via maximum likelihood is a rather elaborate challenge.

3. Sentence Length Studies

3.1. Quantitative Text Typology

Kelih et al. 2006 presented a first systematic approach to the issue of text classification based on sentence length as most important discriminating factor. Here, for each individual text sentence lengths are measured by the number of words per sentence. Thus, a frequency distribution of x -word sentences is obtained. They analyzed 333 Slovenian texts where four statistical characteristics of sentence length turned out to be decisive for discriminating the texts within the framework of multivariate discriminant analysis: mean and standard deviation of sentence length, 2nd Ord's criterion S and entropy h .

These investigations yield a new text typology containing six different types of texts. Dramatic texts (oral discourse), cooking recipes (technical discourse) and novel texts (everyday narration) can be clearly separated, but the discrimination of scientific and journalistic texts is weaker. Private letters build a rather heterogeneous group which cannot unequivocally be assigned to one of the major discourse groups.

3.2. The Relationship between Sentence Length and Word Length

The relation between sentence length (SL x , measured as number of words per sentence) and word length (WL y , measured as number of syllables per word) may be expressed by the so-called Menzerath-Altmann law (see Altmann 1980) as $y = Ax^b$. Re-analysis of 117 German literary prose texts in Grzybek & Stadlober (2006), originally considered by Arens (1965), showed a rather weak evidence of this model ($r^2 = 0.70$).

This observation was the starting point for a general and more systematic analysis of the WL and SL issue. We pointed out that the Menzerath-Altmann law was constructed as intra-textual relationship, significant for the internal structure of a given sample of texts. In contrast, the data of Arens dealt with inter-textual connections comparing mean lengths of words (\bar{x}_i) with mean lengths of sentences (\bar{y}_i) resulting in two vectors of arithmetic means (\bar{x} and \bar{y}).

Additionally, Grzybek and Stadlober (2006) discovered that pooling of data is highly efficient, resulting in a coefficient of determination of $r^2 > 0.9$. However, pooling alone is not sufficient. Simply adding more data may even worsen the outcome. As a consequence, we concluded that data homogeneity has to be considered, since texts of different types are likely to follow different rules.

The study of Grzybek et al. 2007 analyzed 199 Russian texts grouped in six text types according to the inter-textual relationship, i.e. the mean WL and mean SL of each text resulted in 199 data points controlled by the six text types. The fitting results were poor for each case ($0.02 < r^2 < 0.26$) reflecting only a weak relation between the means of WL and SL .

Grzybek et al. 2008 continued the work on this subject by looking at the intra-textual-level with particular emphasis on different text types. All sentences with the same length SL were compared with the mean values m_{WL} of the word lengths of all sentences with length SL .

In order to obtain reliable statistical results, the following restrictions were introduced. (i) Frequencies of sentences with length SL (f_{SL}) had to be larger than 30, (ii) mean word length $m_{WL} > LCP$, where $2 < LCP < 7$, depending on the text type, (iii) sentence length SL had to be smaller than 30. More than 90% of the data met these requirements.

Under the restrictions above, a number of surprising results were accomplished. For three of the four text types of Russian texts (drama, comment, letters) no Menzerath tendency could be confirmed.

Only for literary texts a Menzerath tendency could be observed. For a partial corpus consisting of these three text types and for the whole corpus, the Menzerath model is suitable. Hence, it seems reasonable to look at the important factor of data heterogeneity, because it is relevant when different text types are merged together. Therefore, building an *external textual heterogeneity* seemed to be helpful. Literary texts however, have an *internal textual heterogeneity* typically composed of dialogues, descriptive parts and narrative sequences, sharpened by different WL and SL relationships.

4. Studies of Shot Duration in Silent Movies

At the end of the year 2015 Peter appeared in my office and told me about interesting studies on shot length frequencies in movies, a research topic that culminated in a vivid and controver-

sial scientific discussion (Baxter 2015, DeLong 2015, Redfern 2015). These scientific contributions inspired and motivated me to become involved in this interesting research area. As last joint appearance, we presented our contribution at the Qualico 2016 Conference in Trier, but the results have not been published so far. So I will take this opportunity to pay a last tribute to Peter.

Given that the frequency distribution of shot length displays an asymmetric left-skewed shape, the aforementioned discussions focused on the question of whether the lognormal distribution is an adequate model or not. Most participants in this many-voiced choir united to the unison support of this model as a suitable one, although the crucial question if there is some "philosophy" behind it, which might serve as an explanation for the observations made, has remained unanswered. In her dissertation, published in 2014, Veronika Koch conducted a systematic study of Russian Soviet movies. For the analysis she favored the Zipf-Alekseev function as suitable model, not knowing that this model is not different from the lognormal distribution. Indeed, below we will show that the Zipf-Alekseev distribution is equivalent to the lognormal distribution where the rates of change express the relationship of the parameters.

Our theoretical arguments will be sided by a re-analysis of shot length frequencies in twelve early Laurel and Hardy silent movies, providing additional empirical arguments for the lognormal distribution as a partially good choice in this context.

4.1. Some Theoretical Results of the Lognormal Distribution

In the following we will discuss some crucial theoretical properties of the lognormal distribution: its well-known connection to the normal distribution, and its widely unknown equivalence to the Zipf-Alekseev distribution as well as to an alternate form of the lognormal distribution.

The normal random variable $Y(N(\mu, \sigma))$ has expectation $E(Y) = \mu$, and variance $Var(Y) = \sigma^2$. Its probability density is given as

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, \quad y \in \mathbb{R}; \mu \in \mathbb{R}, \sigma > 0.$$

Then the random variable $X = \exp(Y)$ is lognormal ($LN(\mu, \sigma)$) with probability density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(\log(x)-\mu)^2}, \quad x > 0; \mu \in \mathbb{R}, \sigma > 0 \quad (1)$$

and $E(X) = e^\mu e^{\sigma^2/2}$, $Var(X) = e^{2\mu} e^{\sigma^2} (e^{\sigma^2} - 1)$.

We introduce the parameters $b_N = e^\mu$ and $c_N = \sigma$ ($LN(b_N, c_N)$) and get the density as

$$f(x) = \frac{1}{\sqrt{2\pi}c_N} e^{-(\log(x) + \frac{1}{2c_N^2}(\log(x/b_N))^2)}, \quad x > 0; b_N > 0, c_N > 0. \quad (2)$$

The exponential form of the Zipf-Alekseev distribution $ZA(a_Z, b_Z)$ can be written as

$$f(x) = C e^{-(a_Z \log(x) + b_Z (\log(x))^2)}, \quad x > 0; a_Z \in \mathbb{R}, b_Z > 0 \quad (3)$$

with normalizing constant

$$C = \sqrt{\frac{b_Z}{\pi}} \exp\left(-\frac{(1-a_Z)^2}{4b_Z}\right).$$

Let us consider the rates of change of $LN(b_N, c_N)$ and $ZA(a_Z, b_Z)$ given as

$$rc(x) = f'(x).$$

Because of $f(x) = \exp(h(x))$ we get $f'(x) = f(x)h'(x)$ and

$$rc(x) = h'(x).$$

We show that the lognormal $h'_N(x)$ is equal to the Zipf-Alekseev $h'_Z(x)$.

$$h'_N(x) = -\frac{c_N^2 + \log\left(\frac{x}{b_N}\right)}{c_N^2 x} = h'_Z(x) = -\frac{a_Z + 2b_Z \log x}{x}$$

$$\log x(1 - 2c_N^2 b_Z) = c_N^2(a_Z - 1) + \log b_N.$$

The left part of the equation is zero if

$$1 - 2c_N^2 b_Z = 0 \Rightarrow b_Z = \frac{1}{2c_N^2} \Rightarrow a_Z = \frac{c_N^2 - \log b_N}{c_N^2}.$$

Hence we can formulate the following **crucial property**.

The Zipf-Alekseev distribution $ZA(a_Z, b_Z)$ is identical to the lognormal distribution $LN(b_N, c_N)$ if the parameters are given as

$$a_Z = \frac{c_N^2 - \log b_N}{c_N^2} \quad \text{and} \quad b_Z = \frac{1}{2c_N^2},$$

$$b_N = \exp\left(\frac{1 - a_Z}{2b_Z}\right) \quad \text{and} \quad c_N = \frac{1}{\sqrt{2b_Z}}.$$

The software package `TableCurve` (not used in this paper) includes Ron Brown's lognormal function which is defined as probability density

$$f(x) = Ae^{\left(-\frac{1}{2c_L^2}(\log(x/b_L))^2\right)}, \quad x > 0; b_L > 0, c_L > 0, \quad (4)$$

with normalizing constant

$$A = \frac{1}{\sqrt{2\pi}b_L c_L} \exp(-c_L^2/2).$$

$$E(X) = b_L \exp(3c_L^2/2), \text{Var}(X) = b_L^2 \exp(3c_L^2)(\exp(c_L^2) - 1).$$

Equating the rates of change Lognormal $h'_N(x) =$ alternate Lognormal $h'_L(x)$ leads to the second *crucial property*:

The alternate lognormal distribution $AL(b_L, c_L)$ is identical to the lognormal distribution $LN(b_N, c_N)$ with parameters given as

$$c_L = c_N \text{ and } b_L = b_N \exp(-c_N^2),$$

$$c_N = c_L \text{ and } b_N = b_L \exp(c_L^2).$$

Because of these equivalences our statistical analysis will fall back to the lognormal distribution in its original form as defined in equations (1) and (2).

4.2. Data Material

We investigate twelve Laurel and Hardy silent movies (1927–1929) introducing the following notation: The film length is the number of shots per film, the film duration is the time (in min:sec) from the first shot to the end of the very last shot. Shot length is the number of frames between two consecutive transitions (24 frames per second, i.e. 24 frames \equiv 1 second) and shot duration is the temporal interval between two consecutive transitions (in seconds). We define shot duration as continuous random variable X which will be fitted by an adequate lognormal distribution. Table 1 lists the twelve films with their title, production, release, film duration fd , film length fl , duration/length fd/fl , average shot duration \bar{x} and standard deviation of shot duration sd .

Table 1
Characteristics of Twelve Laurel and Hardy Silent Movies

No.	Title	Prod.	Release	fd	fl	fd/fl	\bar{x}	sd
1	<i>The Second Hundred Years</i>	06 1927	10 1927	20:26	242	5.07	5.00	3.60
2	<i>Putting Pants on Philip</i>	08 1927	12 1927	18:58	175	6.50	8.85	3.80
3	<i>Leave 'Em Laughing</i>	10 1927	01 1928	21:24	222	5.78	5.65	3.80
4	<i>From Soap to Nuts</i>	12 1927	01 1928	18:06	197	5.51	5.71	3.30
5	<i>The Finishing Touch</i>	11/12 1927	01 1928	19:38	207	5.69	5.44	3.50
6	<i>You're Darn Tootin'</i>	01 1928	04 1928	20:53	189	6.63	7.63	4.00
7	<i>Early to Bed</i>	05 1928	10 1928	19:04	248	4.61	4.06	3.15
8	<i>Habeas Corpus</i>	07 1928	12 1928	19:40	231	5.11	7.56	2.90
9	<i>Liberty</i>	09 1928	01 1929	18:14	169	6.47	8.32	3.30
10	<i>Wrong Again</i>	10/11 1928	01 1929	19:45	171	6.93	9.30	3.70
11	<i>Bacon Grabbers</i>	02/03 1929	10 1929	19:31	232	5.05	4.82	3.30
12	<i>Angora Love</i>	03 1929	12 1929	20:38	262	4.73	4.82	3.35

4.3. Statistical Analysis of 12 Silent Movies

We investigate whether the variable X (shot-length duration) can be adequately described by a lognormal distribution. This is equivalent to the question whether the transformed variable $Y = \log(X)$ is approximately normal distributed. To check this with the software package R several diagnostic tools are applied: graphical representations and goodness of fit tests.

- Histogram with normal density and smooth function
- Boxplot
- Empirical distribution function with normal distribution
- Quantile-Quantile-Plot for diagnosis (reference linear pattern)
- Tests for normality: Shapiro-Francia (Shapiro & Francia 1972), Jarque-Bera (Bera & Jarques 1980)

Based on the tests of normality we define the following decision rules.

- *Strictly normal* if p -values of both tests larger than 0.05

- Close to normal if p -values of both tests larger than 0.01 and one test larger than 0.05 for truncated data (97.5%)
- Deviations from normal if p -value of at least one test smaller than 0.01 for truncated data (97.5%)

For each group we choose one representative data set and show its results in full detail. The silent movie *The Second 200 Years* shows a perfect fit, close to normal is *You're Darn Tootin'*, and deviations from normal can be observed in *Wrong Again*.

As an overall result we can classify 3 movies as strictly normal and 6 movies as close to normal. However, 3 movies seem to have significant deviations from the normal (or lognormal) distribution.

4.3.1. Silent Movie: The Second Hundred Years (1927)

Length in seconds: 1226, Sample size $N = 242$ shot-lengths,

estimated parameters: $\hat{\mu} = \bar{x} = 5.07$, $\hat{\sigma} = sd = 5.00$

Shapiro-Francia-Test: $p = 0.48$, Jarque-Bera-Test: $p = 0.33$

Movies 7 (*Early to Bed*) and 12 (*Angora Love*) show also a perfect fit.

log(X): The Second 100 Years (1927)

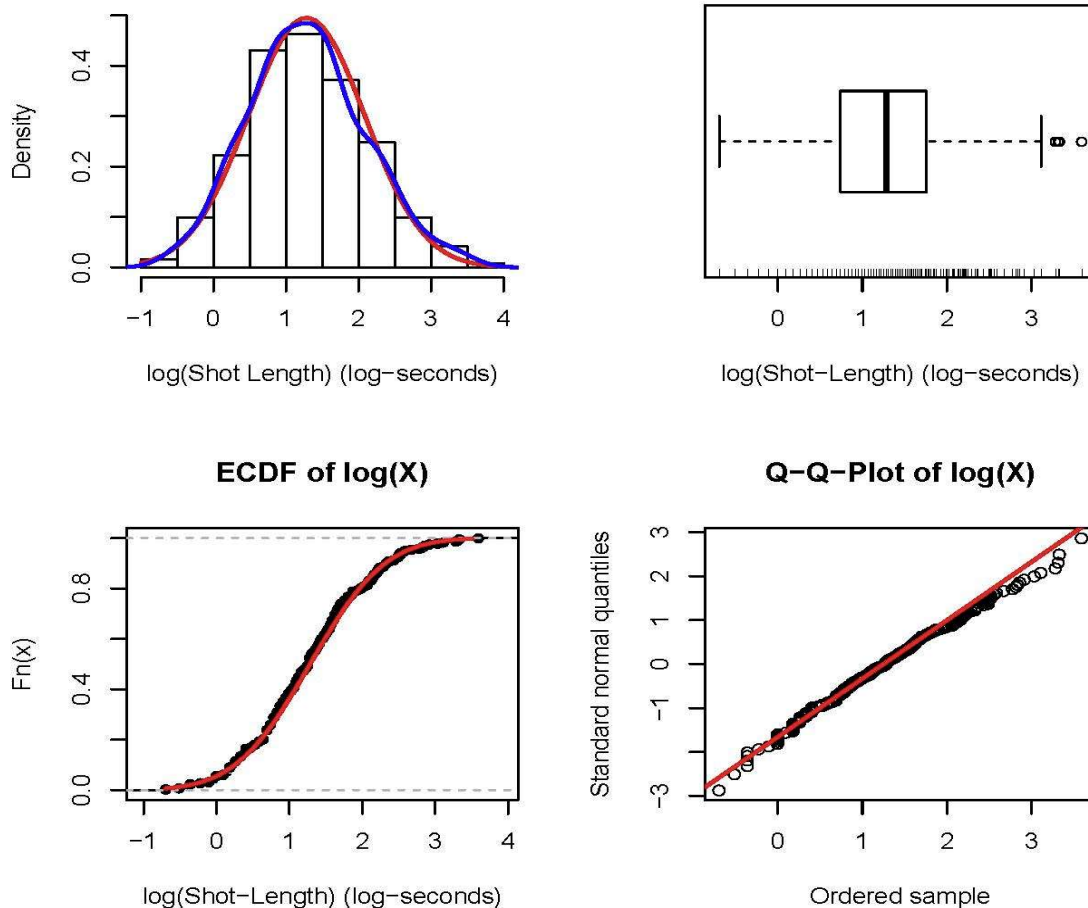


Figure 1 Histogram of $\log(X)$ with smoother (blue), normal density (red) and Box Plot (above), empirical and normal distribution function (red), Q-Q-Plot (below)

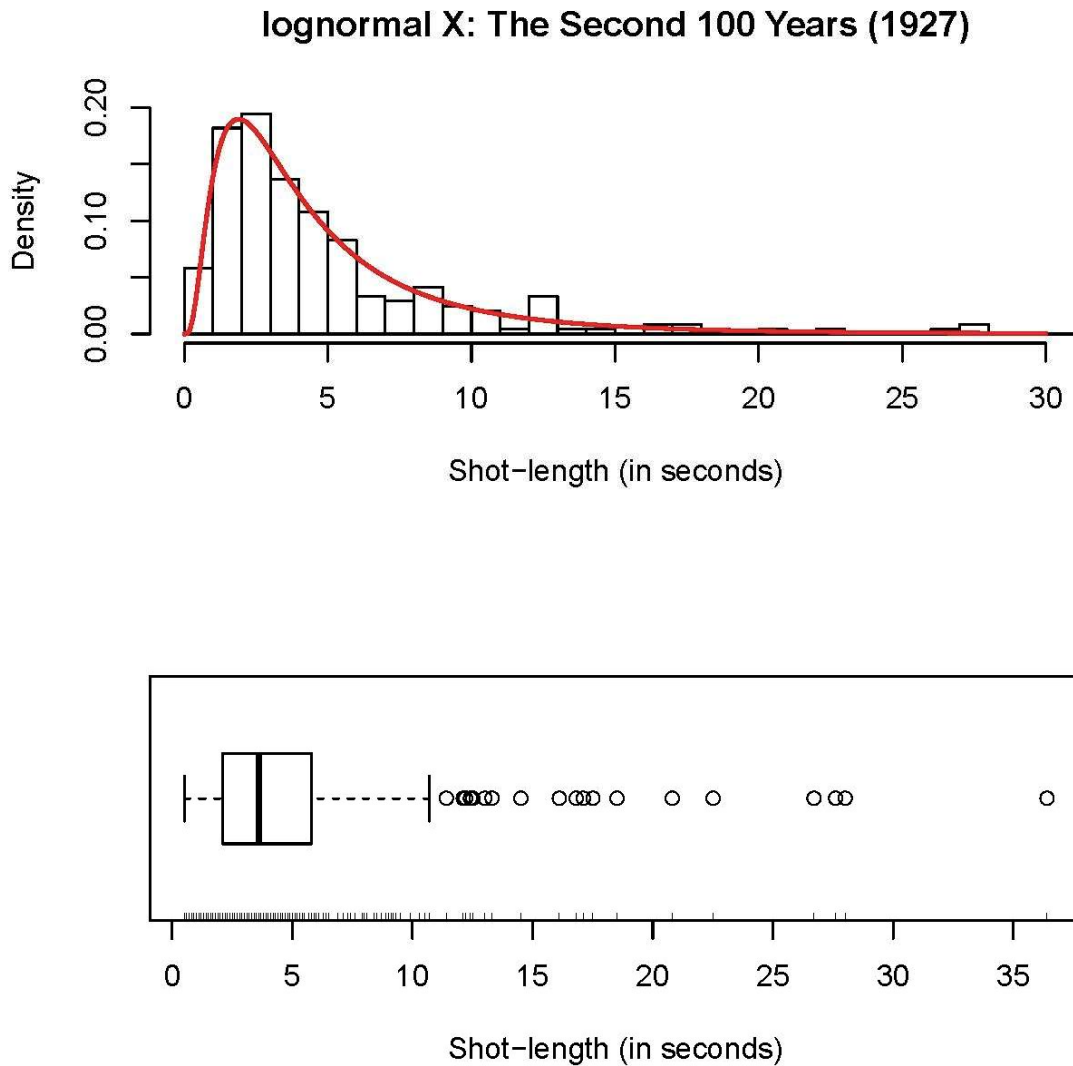


Figure 2 Histogram of X with lognormal density (red), Boxplot

4.3.2. Silent Movie: *You're Darn Tootin'* (1928)

Length in seconds: 1253, Sample size $N = 189$ shot-lengths,
 estimated parameters $\hat{\mu} = \bar{x} = 1.47$, $\hat{\sigma} = sd = 0.88$

S-F-Test: $p = 0.004(0.17)$, J-B-Test: $p = 0.03(0.15)$ (values in brackets with 0.975-filter)

Movie 2 (*Putting Pants on Philip*), movie 4 (*From Soap to Nuts*), movie 8 (*Habeas Corpus*),
 movie 9 (*Liberty*) and movie 11 (*Bacon Grabbers*) are also close to normal (or lognormal).

The central body of the empirical distribution fits very well. However, more extreme values appear as one would expect under the lognormal model. Hence filtering the highest 2.5% of the values improves the overall fit remarkably.

log(X): You're Darn Tootin' (1928)

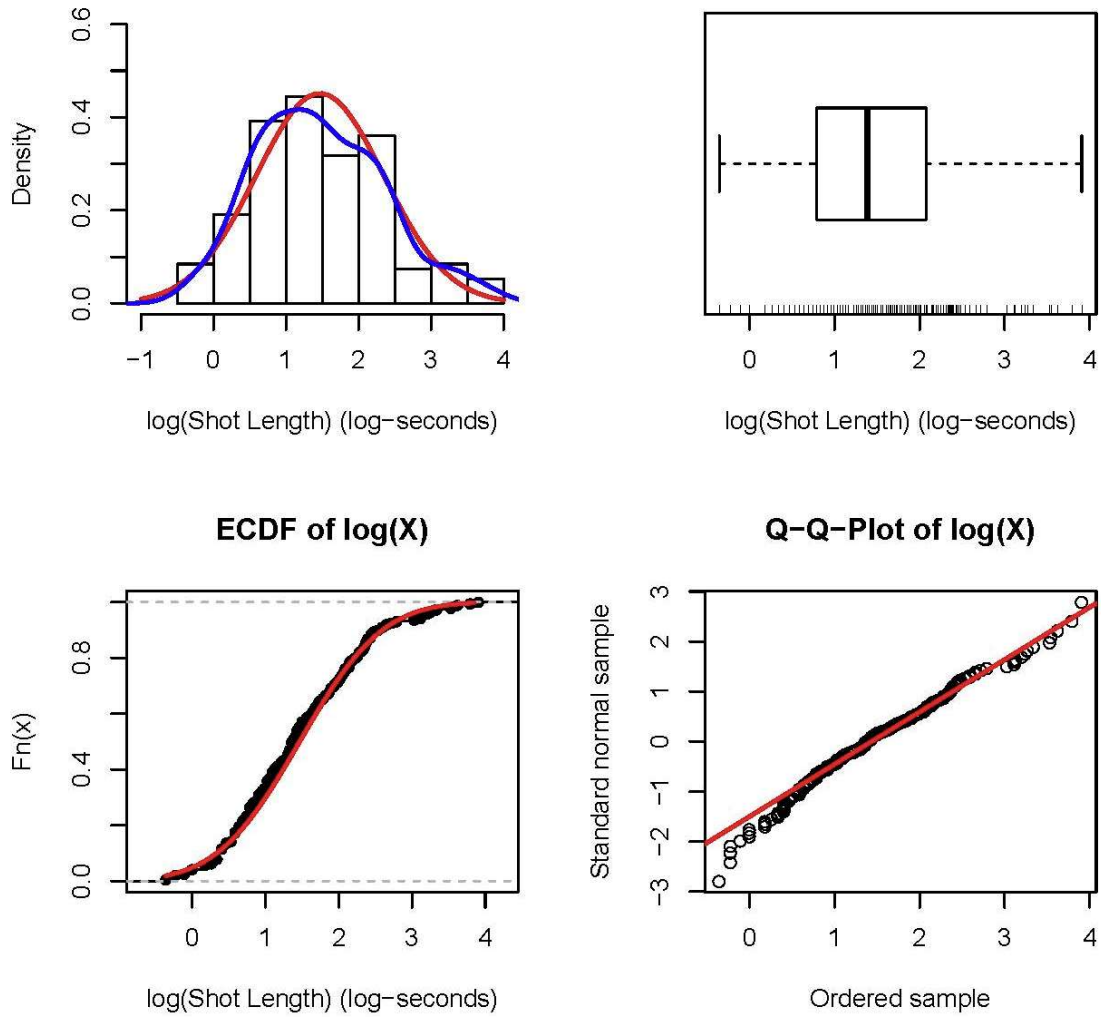


Figure 3 Histogram of $\log(X)$ with smoother (blue), normal density (red), Box Plot (above), empirical and normal distribution function (red), Q-Q-Plot (below)

lognormal X: You're Darn Tootin' (1928)

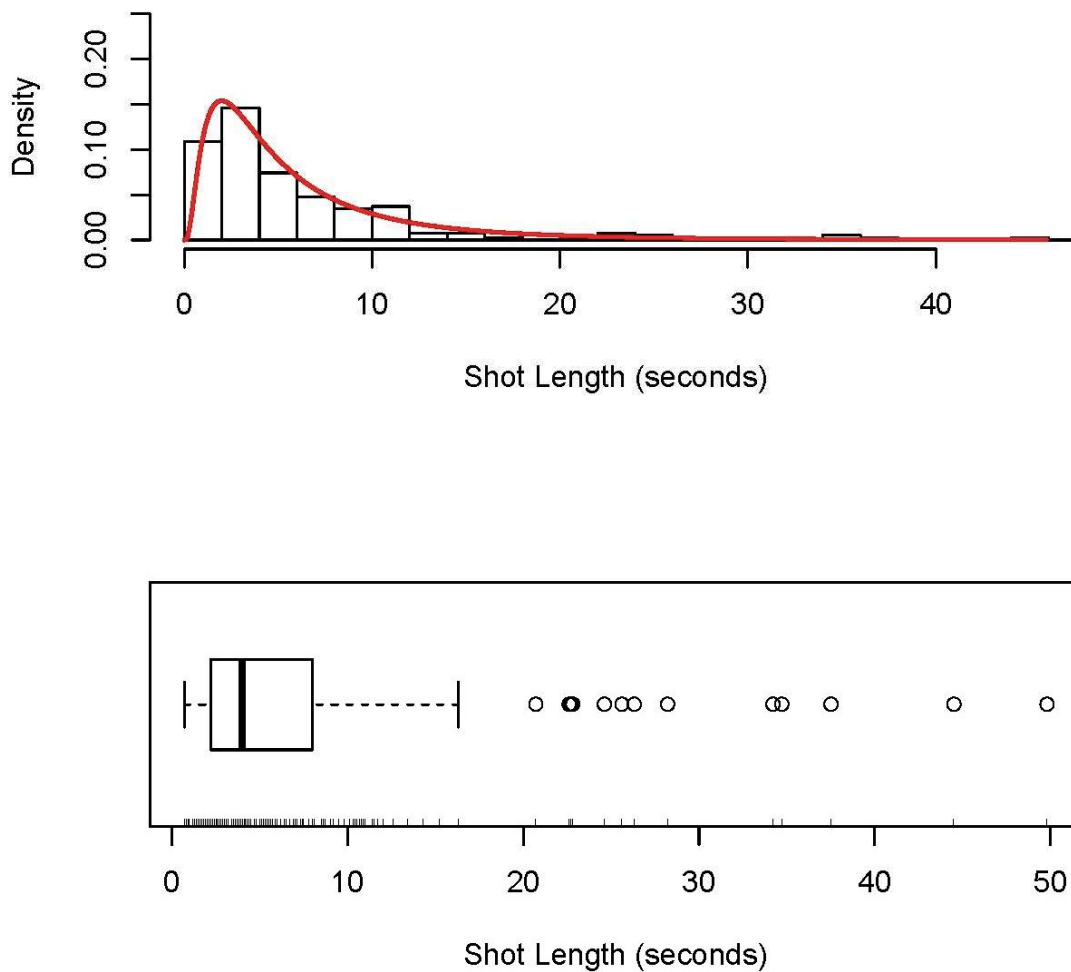


Figure 4 Histogram of X with lognormal density (red), Boxplot

4.3.3. Silent Movie: Wrong Again (1929)

Length in seconds: 1185, Sample size $N = 171$ shot-lengths,
 estimated parameters $\hat{\mu} = \bar{x} = 1.45$, $\hat{\sigma} = sd = 0.93$

Shapiro-Francia-Test: $p = 0.008$, Jarque-Bera-Test: $p = 0.039$ (0.975 filter)

Movie 3 (*Leave 'Em Laughing*) and movie 5 (*The Finishing Touch*) have a significant deviation from normal (or lognormal).

The characteristics of these distributions may be rather described by a mixture of two distributions than by a single unimodal model. One part contains short shot durations around 5 seconds, the other includes a large range of durations around 10 seconds up to 60 seconds (see Figure 6).

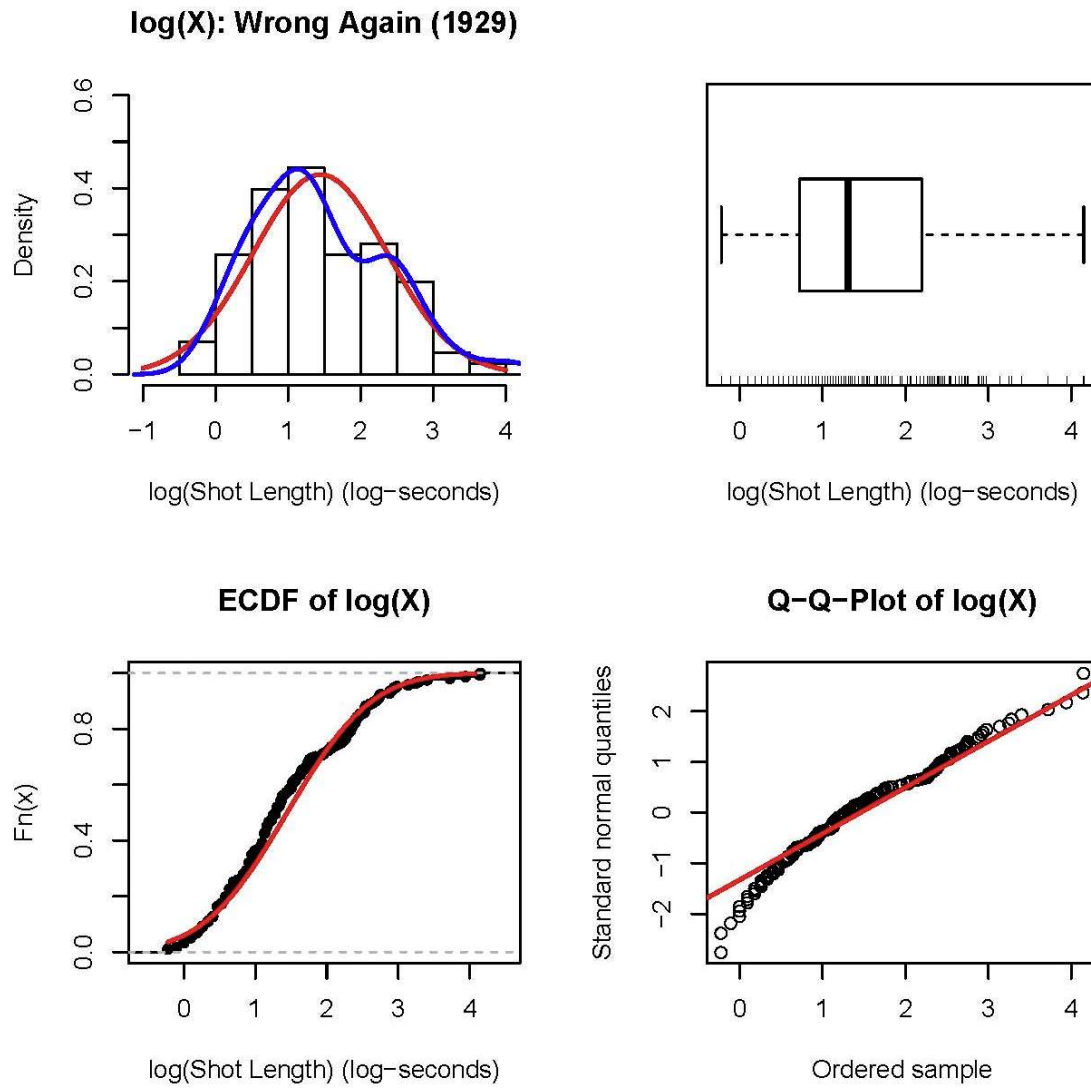


Figure 5 Histogram of $\log(X)$ with smoother (blue), normal density (red), Box Plot (above), empirical and normal distribution function (red), Q-Q-Plot (below)

lognormal X: Wrong Again (1929)

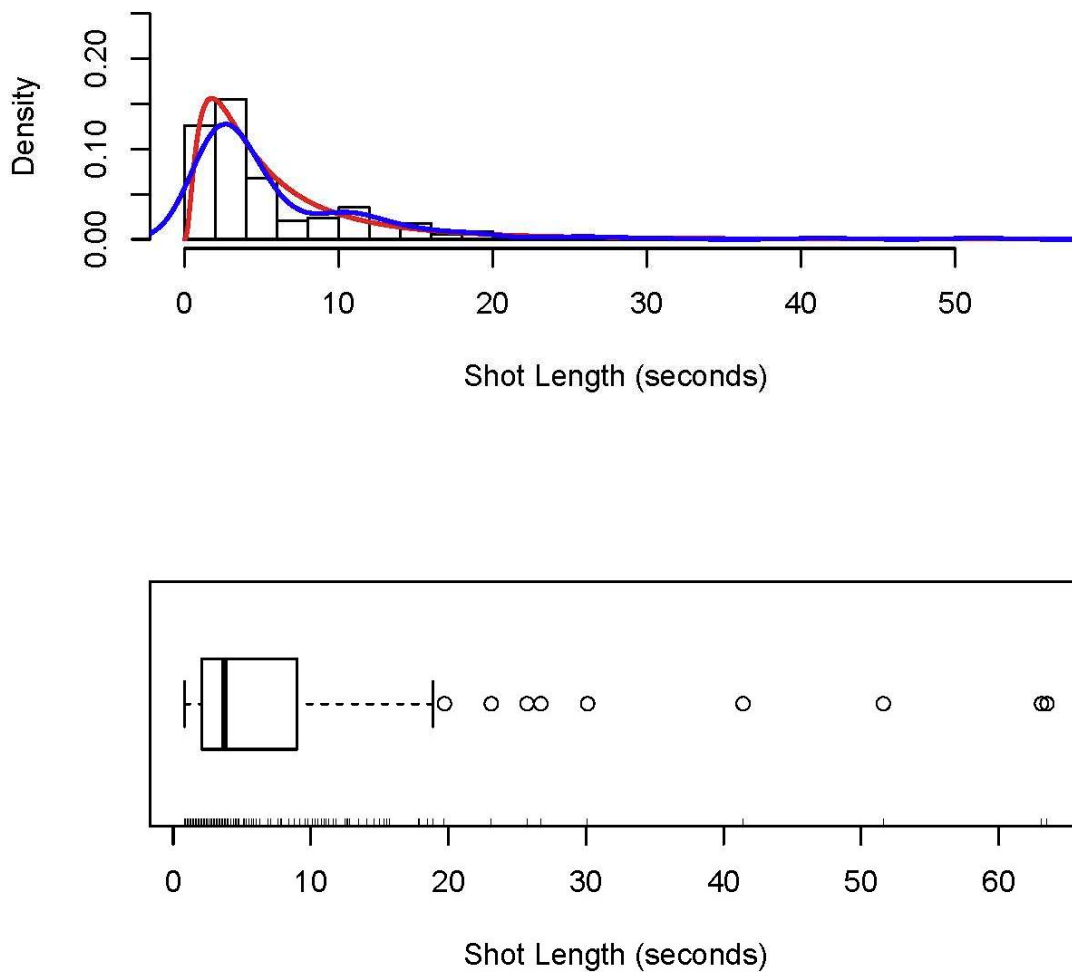


Figure 6 Histogram of X with lognormal density (red), Boxplot

5. Conclusions

With our contributions, we attempted to introduce a more systematic view into the studies of word length of Slavic languages. Dealing with the problem of discrimination and classification of different text sorts led to a new discussion about the existence of specific discourse types different from traditional functional styles.

The exhaustive studies of word length frequencies resulted in the Singh-Poisson distribution, a distribution with remarkable features: (i) the whole range of the index of dispersion can be covered, (ii) it only contains two parameters, (iii) it can differentiate different text types, (iv) the calculation of the maximum likelihood estimates is a simple task.

By relating sentence length to word length, we differentiate between two approaches: the inter-textual and the intra-textual connections. This strategy yielded a new interpretation of the Altmann-Menzerath law.

Finally, we investigated the shot duration of 12 silent movies created by Laurel and Hardy and we were able to show that the shot durations of 9 movies (75%) can be adequately described

by the lognormal distribution. The remaining 3 movies seem to have characteristics that cannot be modeled by a unimodal distribution, but rather by a mixture of two distributions.

Acknowledgements

My heartfelt thanks go to Peter and especially to my former PhD student Gordana Djuraš as well as to Peter's colleague and friend Emmerich Kelih. Due to our highly rewarding collaboration, I was able to expand my professional horizons considerably. I became fascinated by the potential of statistical modeling in the field of quantitative linguistics. It opened my mind to the way of thinking in the humanities which is different to the approach in natural science. I got the opportunity to introduce a pragmatic and systematic approach into the field, because of my mathematical and statistical expertise. It was a great challenge to absorb spectacular ideas, give them a solid theoretical basis and conduct experimental studies based on selected data.

I want to thank Peter for his never-ending enthusiasm and motivation. He enriched my life and I miss him. Through our joint work his spirit will be present forever and I will always remember him for our unique collaboration.

References

- Altmann, G.** (1980). Prolegomena in Menzerath's Law. *Glottometrics* 2, 1-10.
- Antić, G., Grzybek, P., Stadlober, E.** (2005). Mathematical Aspects and Modifications of Fuck's Generalized Poisson Distribution (GPD). In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 158-180*. Berlin, New York: de Gruyter.
- Arens, H.** (1965). *Verborgene Ordnung. Die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute*. Düsseldorf: Pädagogischer Verlag Schwann.
- Baxter, M.** (2015). On the Distributional Regularity of Shot Lengths in Film. *Digital Scholarship in the Humanities* 30/1, 119-128.
- Bera, A.K., Jarques, C.M.** (1980). Efficient Test for Normality, Homoscedasticity and Serial Independence of Regression Residuals. *Economics Letters* 6, 255-259.
- DeLong, J.** (2015). Horseshoes, Handgrenades, and Model Fitting: the Lognormal Distribution is a Pretty Good Model for Shot Length Distribution of Hollywood Films. *Digital Scholarship in the Humanities* 30/1, 129-136.
- Djuraš, G.** (2012). Generalized Poisson Models for Word Length Frequencies in Texts of Slavic Languages. Graz: Dissertation, Technische Universität.
- Djuraš, G., Stadlober, E.** (2010). Modeling Word Length Frequencies by the Singh-Poisson Distribution. In: Grzybek, P. et al. (eds.), *Text and Language. Structures. Functions. Interrelations. Quantitative Perspectives: 37-48*. Wien: Praesens.
- Djuraš, G., Stadlober, E., Kelih, E.** (2013). The Generalized Poisson Distributions as Models of Word Length Frequencies. in Obradović, I. et al. (eds.), *Methods and Applications of Quantitative Linguistics: 107-118*. Belgrade: Academic Mind.
- Djuraš, G., Stadlober, E., Kelih, E., Grzybek, P.** (2013). Komplexität sprachlicher Formen. Die Singh-Poisson-Verteilung: ein Modell in der Wortlängenforschung? In: Köhler, R., Altmann, G. (eds.), *Studies in Quantitative Linguistics 13, 2013, dedicated to Karl-Heinz Best 70th birthday: 291-308*. Lüdenscheid: RAM-Verlag.
- Djuzelic, M.** (2002). *Einflussfaktoren auf die Wortlänge und ihre Häufigkeitsverteilung am Beispiel von Texten slowenischer Sprache*. Graz: Diplomarbeit, Institut für Statistik, Technische Universität Graz.
- Fucks, W.** (1956). Mathematical Theory of Word Formation. In: Cherry, C. (ed.), *Information theory: 154-170*. London: Butterworth.
- Grzybek, P., Kelih, E., Stadlober, E.** (2008). The Relation between Word Length and Sentence Length: An Intra-systemic Perspective in the Core Data Structure. *Glottometrics* 16, 111-121.
- Grzybek, P., Stadlober, E.** (2002). The Graz Project on Word Length (Frequencies). *Journal of Quantitative Linguistics* 9, 2, 187-192.
- Grzybek, P., Stadlober, E.** (2006). Do We Have Problems with Aren's Law? The Sentence-Word Relation Revisited. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 205-218*. Berlin: de Gruyter.
- Grzybek, P., Stadlober, E., Kelih, E., Antić, G.** (2005). Quantitative Text Typology: The Impact of Word Length. In: Weihs, G., Gaul, W. (eds.), *Classification the Ubiquitous Challenge: 498-505*. Berlin: Springer.
- Grzybek, P., Stadlober, E., Kelih, E.** (2007). The Relationship of Word Length and Sentence Length: The Inter-textual Perspective. In: Decker, R., Lenz, H.-J. (eds.), *Advances in Data Analysis: 611-618*. Berlin: Springer.
- Kelih, E., Grzybek, P., Stadlober, E., Antić, G.** (2006). Quantitative Text Typology: The Impact of Sentence Length. In: Spiliopoulou, M. et al. (eds.), *From Data and Information*

- Analysis to Knowledge Engineering*: 382-389. Berlin: Springer.
- Koch, V.** (2014). *Quantitative Film Studies: Regularities and Interrelations Exemplified by Shot Lengths in Soviet Feature Films*. Graz: PhD Dissertation, University of Graz. [<https://unipub.uni-graz.at/download/pdf/242930>]
- Redfern, N.** (2015). The Log-Normal Distribution is not an Appropriate Parametric Model for Shot Length Distributions of Hollywood Films. *Digital Scholarship in the Humanities* 30/1, 137-151.
- Shapiro S.S., Francia, R.S.** (1972). An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association* 67, 215-216.
- Stadlober, E., Djuzelic, M.** (2005). Multivariate Statistical Methods in Quantitative Text Analyses. In: Grzybek, P. (eds.), *Contributions to the Science of Language*: 237-253. Berlin: Springer.

The Peter Grzybek Memorial Archive of Slavic Studies Publications

Emmerich Kelih¹, Hermann Moisl²

Abstract

The present paper gives a brief description of the Peter Grzybek Memorial Archive (<http://www.peter-grzybek-archive.org>). It contains mostly research papers on quantitative linguistics and quantitative text analysis, collected by Peter Grzybek and his colleagues in the past twenty years.

Keywords: Menzerath-Altmann's law; sentence length; word length; grapheme frequency; phoneme frequency; Peter Grzybek

The Peter Grzybek Memorial Archive is dedicated to the memory of Peter Grzybek, who was the leading light in the assembling of the collection of publications on which it is based. The archive and its contents are directly related to Peter's central place of work in the University of Graz, where he initiated the project QuanTa, Quantitative Text Analysis (for details cf. the commented bibliography by Kelih/Schmidt in this volume). The core of QuanTa was the project "Word Length in Slavic Languages", which was active between 2002 until 2005 and whose main collaborators were Gordana Đuraš (formerly Antić), Emmerich Kelih and Rudi Schlatte, with conceptual and statistical support from Ernst Stadlober.

The archive does not contain any publications by Peter himself, but is a collection of papers and articles collected by Peter and his colleagues over the past 20 years.

Peter did not suffer from bibliomania in the usual sense of the word. He was, however, an enthusiastic miner of the research literature on quantitative text analysis in general and as applied to Slavic studies in particular, and was prepared to dive as deep as necessary into that literature to establish the current state of discussion in whatever topic had captured his attention. This taught him the important lesson, familiar to current academic researchers in all disciplines, that the search of burgeoning technical literatures must end at some point to forestall exhaustion – or, as he himself put it, "nicht schon wieder ein neues Faß aufmachen!" ("Let's not tap yet another barrel!"). The distribution of material included in the archive reflects this outlook in that there is no clearly discernible single focus within the general subject domain, but rather an accumulation of literature on topics which interested Peter over many years.

The archive in its given form has in respect to the "input" no sharply describable focus, but some more or less core areas of interest, which are of course a more or less successful reflection of Peter's personal interests in quantitative text analysis and quantitative linguistics.

There are 1,890 entries, of which 1,640 have associated full-text, online-accessible PDFs, many of them scanned from hard copies. In selecting what to scan, the focus was on materials that are now difficult to obtain in their original physical format either on account of their relatively early date of publication or of restricted representation in academic libraries. Although, as noted, there is no single focus in the distribution, various concentrations of topics are discernible. The clearest of these are works on word length and on modelling of word length distributions (cf.

¹ University of Vienna (Austria), emmerich.kelih@univie.ac.at.

² hermann.moisl49@gmail.com.

Grzybek 2006). These relate mainly to the Slavic languages, primarily Russian, Polish, Czech and Serbo-Croatian, but also include German and English. Another is work on grapheme and phoneme frequencies. Peter and his colleagues, mainly Ján Macutek and Emmerich Kelih, have themselves published extensively in this area; a comprehensive monograph was in preparation when Peter passed away, and it remains unpublished (cf. Grzybek/Kelih/Macutek 2020). The archive also contains many older publications relating to phoneme distribution (phonotactics) in the tradition of the Prague school of (linguistic) structuralism, which will be of interest from both theoretical and empirical points of view.

One of Peter's many scientific passions was the history of quantitative linguistics and quantitative approaches to text analysis. The interlibrary exchange service staff at the University of Graz knew Peter very well, since he used the service extensively to track down and obtain now-obscure publications from libraries all over the world. He refused to accept that any given item "could not be found" or was "unavailable via the exchange service", and was known to contact libraries directly until his objective was achieved. Peter's disposition to hunt for apparently peripheral publications is represented in the online archive mainly via Russian works on quantitative approaches to text analysis and quantitative linguistics, on which see Grzybek & Kelih (2005). Russian philology at the end of the 19th and the beginning of the 20th centuries can be regarded as a prototypical laboratory for quantitative approaches, without sharp demarcation of literary studies from linguistics and productive of many innovative ideas and methodologies.

The archive also contains some examples of German work on quantitative text analysis, mainly in the form of older papers by Wilhelm Fucks (1902–1990), a pioneer of German quantitative linguistics and quantitative text, music and painting analysis, whose works are scattered among many different journals.

The stored articles have served as input for an outstandingly rich output, i.e. the bulk of the published scientific papers written by Peter, many of them in collaboration with his colleagues. His output need not be presented in detail here, since Peter created an archive of his publications during his lifetime which is available at <http://peter-grzybek.eu/> and contains almost all of his scientific output from 1983 to 2016. There is only one part of his research interest which is not reflected in his output: in the last years of his life he was quite fascinated by rhythm, i.e. repetitive structures in language, text, music, paintings and film. This interest in rhythm and repetitive structures could be seen as a bridge between biological and cognitive laws, without any border between nature and culture. Through extensive legwork Peter collected many interesting papers and contributions, mainly of (older) German and Russian scholars, about this fascinating phenomenon. However, as already pointed out at the beginning of this paper, the Peter-Grzybek memorial archive of course reflects only a small proportion of Peter's research interests and many of the started projects and ideas remain unfinished.

It is as it is – Peter, thank you for all your efforts and far-sightedness!

References

- Grzybek, P.** (ed.) (2006). *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht: Springer (Text, Speech and Language Technology, 31).
- Grzybek, P., Kelih, E., Mačutek, J.** (2020). *Alphabet Analyses. Quantitative Studies on Slavic Letter and Grapheme Frequencies*. (unpublished manuscript 220pp.)
- Grzybek, P.; Kelih, E.** (2005). Zur Vorgeschichte quantitativer Ansätze in der russischen Sprach- und Literaturwissenschaft. In: Köhler, R., Altmann, G., Piotrovskij, R.G. (eds.): *Quantitative Linguistik: Ein internationales Handbuch. Quantitative Linguistics: An International Handbook: 23-64*. Berlin: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).
- Kelih, E.; Grzybek, P.** (2005). Neuanfang und Etablierung quantitativer Verfahren in der sowjetischen Sprach- und Literaturwissenschaft (1956-1962). In: Köhler, R., Altmann, G., Piotrovskij, R.G. (eds.): *Quantitative Linguistik: Ein internationales Handbuch. Quantitative Linguistics: An International Handbook: 65-82*. Berlin: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Acknowledgments

The design and creation of the archive (<http://www.petergrzybek-archive.org>) was generously supported by the steadfastness, patience and precision of numerous volunteers. In particular we would like to mention Veronika Schmid (formerly Koch), Nino Auer, Bianca Sieberer and Karolina Wieserová.

Commented Bibliography of Peter Grzybek

(Or a Short Contribution to His Impact on Quantitative Text Analysis)

Emmerich Kelih¹, Veronika Schmidt²

Abstract

The article gives a short description and summary of the main contributions of Peter Grzybek on quantitative text analysis and quantitative linguistics. Moreover, the full bibliography of Peter Grzybek is given.

Keywords: Menzerath-Altmann law; sentence length; word length; grapheme distribution; phoneme distribution; quantitative film analysis; prose rhythm: Peter Grzybek

1. Introduction

To give a paean of praise to Peter Grzybek's tremendous merits in quantitative text analysis and quantitative linguistics would absolutely not be in line with his nature or character. Quite the opposite seems to be necessary, and thus we prefer a factual and unemotional reconstruction of his oeuvre. Based on his numerous written contributions, some remarkable foci and in particular the development of his approaches and way of thinking in quantitative text analysis shall be identified and reconstructed. We believe that a "sober view" on his contributions, without pathos and emotions, is what Peter would like to read and see. However, one has to remind the kind reader that Peter had many different academic interests and key areas, including literary and cultural theory, general semiotics, phraseology and in particular paremiology. After his passing away, numerous obituaries (cf. Eismann 2019a, 2019b, Eismann/Deutschmann 2019, Kelih/Köhler/Altmann 2019, 2020, Kõiva 2019, Lotman/Pilshchikov/Lotman 2019, Mieder 2019a, 2019b) have been published in all his different academic communities, realistically reflecting the broadness of his merits. Peter's impact is reflected by his comprehensive oeuvre, collected in the almost complete bibliography of his published monographies, articles and papers as well as numerous edited omnibus volumes, which is published in the appendix of this article. Based on this bibliography his way of doing quantitative text analysis and his profile of applying statistical methods in linguistics and text analysis can be easily summarized as follows.

For sure Peter's main area of scientific engagement was phraseology, and more precisely paremiology. It is this engagement with idiomatic expressions which was in some respects the door opener to the "world of quantitative analysis". The empirical analysis of proverbs, namely empirical studies about their popularity, i.e. whether proverbs are known to speakers or not, is the starting point for his empirical-statistical work. Furthermore, he was attracted by the problem of variation in proverb usage, since speakers are hardly able to reproduce proverbs only in a single given canonical form, but with variations. Peter became more and more interested in factors like the length of proverbs, their familiarity or frequency, influencing the popularity of proverbs among speakers of a given language. In this respect phraseology was the trigger for a more thorough and intensive occupation with statistical methods in general. This

¹ University of Vienna, Austria, emmerich.kelih@univie.ac.at.

² Schmidt_Vero@web.de.

interest in the lengths of linguistic units and constructs, i.e. the length of proverbs, was soon followed and deepened by a strong interest in word length studies. Relatively early, in the mid-nineties, Peter became intrigued by the analysis of functional dependencies and interrelations. He focussed on Menzerath's law and later Arens' law, and by doing so he definitely entered the core area of quantitative linguistics.

In the late nineties Peter became engaged more and more in word length and word length frequency. This was a time when several projects on word length had been launched simultaneously. Especially the Germany-based Göttingen project, headed by Karl-Heinz Best, played an important role. Peter, being exceptionally outgoing, quickly established fruitful contacts with leading proponents of word length studies. At that time quantitative linguistics was dominated by "holistic approaches", in particular regarding the statistical modelling of linguistic data. The unified derivation of linguistic laws, mainly proposed by Gabriel Altmann and his fellows (cf. Wimmer/Altmann 2005), was vividly discussed by the quantitative linguistics community. In addition, some technological and software developments in text analysis and statistical modelling (among others Altmann-Fitter 1.0) also facilitated the empirical falsification of quantitative linguistic hypotheses. Most of them were related to ideas coming from synergetic linguistics, which offered a remarkable theoretical framework for the understanding of linguistic laws.

Against this background the project "QuanTA – Quantitative Text Analysis" was born at the Institute of Slavic Studies at the University of Graz, Austria. Its most important project, "Word Length (Frequencies) in Slavic Language Texts", was the official start of a fruitful series of international projects in quantitative linguistics and quantitative text analysis. The Graz project on word length was in some sense Peter's favourite, and without any exaggeration in some aspects a best-practice model for quantitative linguistics. Being interdisciplinary from its beginning, it assembled people coming from linguistics, statistics and computer sciences, having one common goal: an in-depth analysis of word lengths in Slavic languages. The empirical basis of this project was a gradually developed balanced corpus of Croatian, Russian and Slovene texts, which served as the database for several empirical and theoretical studies, many of them published by Peter himself or jointly with his collaborators.

One key aspect was the sounding of boundary conditions, influencing the "shape" of word length frequency distributions in languages. This included the discussion of genuine linguistic factors, like the kind of text pre-processing, word definition, choice of measuring units, possibilities and limits of automatic word length determination and finally the main question to what extent an author, text or discourse type can determine the word length of a text. Some results can be found in Grzybek, Stadlober & Kelih (2005e), where mainly by means of linear discriminant analysis, based on word length measures (the mean, standard deviation, kurtosis, Ord's criteria etc.) the possibility of a classification of the "world of texts" has been challenged. Another salient point was the modelling of word length frequency distributions, where Peter in liaison with colleagues from statistics started to reconstruct statistical models already discussed in the past. Peter and his colleagues contextualized these works (going back to Wilhelm Fucks, R. Grotjahn, V. Kromer, A. Bartkowiakowa, B. Gleichgewicht) and tried to show their attractiveness and suitability for the recent discussion in word length modelling.

It was quite remarkable with how much obstinacy and tenaciousness Peter "discovered" older, in particular in the Western hemisphere fewer known, works on word length modelling from Poland and the countries of the former Soviet Union. Some of these hardly accessible works and papers, are now digitally available in the Peter Grzybek memorial archive of Slavic Studies Publications (cf. Kelih/Moisl 2020 in this volume for further details). For a more thorough overview of some results of this reconstruction of appropriate models from modelling word length distributions see Antić/Grzybek/Stadlober (2005m).

During the time of the word length project (which officially lasted from 2002 until 2005) his activities and interests expanded to many other areas of quantitative text analysis and linguistics. Peter's energy and passion were at its peak at this time; he pushed many projects and ideas at the same time. Particular attention was paid to sentence length and its relation to word length, which is an integral part of Menzerath's law and Arens' law. As is well known, Menzerath's law describes the relation between the constituting components of a given construct. Originally it was designed in terms of an intra-textual law, relevant for the internal structure of a given text sample. Arens' data, however, were of a different kind, implying inter-textual relations. It was Peter's persistence to explore this relation more thoroughly, since it is also a well-known fact that the relation between word and sentence length is much looser (in terms of the fitting results) than for instance the relation between word length and syllable length (being the "classical" case of Menzerath's law). The most common and probable explanation for the weak relation between word length and sentence length is the hierarchical distance between the word and the sentence, i.e. the high number of intermediate levels (like clauses, phrases, multi-word expressions etc.). However, as pointed out by Grzybek/Stadlober (2007d), there are many other factors which have to be taken into consideration, not least pooling procedures (especially of sentence length classes) and data sparsity (in particular relevant for longer sentences which simply do not occur so often). All in all, data homogeneity of the analysed texts played a great role in modelling. In any case a "lucky" mix of different, carefully selected text types (for instance private letter, drama, short story, novel, comment, scientific text) increased the homogeneity of the corpus, which was again reflected in the obtained results.

Peter was not only focussed on word length studies and Menzerath's and Arens' law alone, but bearing in mind his character and nature, one has to mention his openness and willingness to "dive" quickly into various areas of research, such as text difficulty studies and related concepts of measuring it by using various quantitatively determinable empirical parameters. In Grzybek (2010d) he showed that the normally taken ad hoc parameters are not necessary, since a reduction to word or sentence length can already show convincing results of text difficulty.

Some other interesting results were reported by Fenxiang/Grzybek/Altmann (2010b), where word length distributions are seen through the prism of word positions within a sentence and the accompanied length of the sentence. The question was whether word length remains stable in the course of a sentence, from its beginning to its end, or if there is a particular change or development of word length, depending on the length of the sentence. There are several empirical studies about this question. For example, in Finnish the longest words tend to occur at the end of a sentence or a clause. This observation has also been empirically confirmed by the above-mentioned authors for Russian, Slovak, Hungarian, Latin, English and Indonesian.

One further important contribution to the "intermediate-level discussion" is Grzybek (2013c). Using chapters from the famous Russian novel *Anna Karenina* by L. Tolstoj, a systematic interrelation between the chapter (paragraph) length and the sentence length ("the longer the chapter, the shorter the sentences") was found. Peter's contribution to word and sentence length studies could be summarized as being oriented towards systems theory, in particular focussing on the question of hierarchically organized relations in language and text systems. Seen from this point of view, Peter grew up in a structuralistic paradigm. In fact, during his academic career he followed this way of scientific thinking, although he was perfectly aware of all the pitfalls and limits of this "-ism", in particular regarding the ontogenetic status of linguistic and textual entities, being the bread and butter of every linguistic and text scientist. Moreover, it is not difficult to discover in Peter a "double face" on epistemological or methodological levels, since he combined and used both inductive as well as deductive approaches and methods. He knew perfectly well what it meant to work deductively, what it meant formulating empirically verifiable hypotheses, but at the same time he was from time to

time delighted by the mass of empirical material, even painfully aware of being in danger of being flooded by it. The deeper he dug, the greater was the danger of being buried by it. The result is a whole host of unpublished works, and many of his works will remain unpublished and unfinished. But to be honest, so many studies had been started and so many ideas had been born, that it couldn't be handled by one single person.

At least two research areas have to be mentioned which remained more or less unfinished, somewhere between a beginning and finalization. There are in fact very few publications on rhythm and rhythmical structure in his bibliography, but Peter was deeply engaged with repetitive structures in text based on the accent. Already many years ago he started to collect systematically available written resources about rhythm and rhythmic structures in general. Grzybek (2013c) showed what kind of interest he had in rhythm studies. As can be seen from this paper he was mostly interested in prose rhythm, i.e. rhythmical structures in literary text, and in particular the scientific discussion about this question in Russian (Soviet) philological circles in the twenties and early thirties of the last century. This period is somehow a first highlight of rhythm studies in general, where, however, in a broad context even a pan-European rhythm analysis tradition can be observed. In particular the ideas of the German philologists like Karl Marbe (cf. also the biographical sketch by Best 2005), who dominated the field, were intensively discussed in Russia. Peter's interest in these works was not only a methodological one, but he was also interested in the reconstruction of the transfer of knowledge and ideas between the different academic traditions during those times. Moreover, it is symptomatic that Peter was not interested in "history" per se, but mainly in the genesis of ideas and methodological approaches to rhythm studies. Of course, being a Slavist scholar, he was mainly interested in the "Slavic world" (cf. the two texts about the history of Russian quantitative text analysis Grzybek/Kelih 2005l and Kelih/Grzybek 2005f), but he always had a much broader perspective and the ability to see beyond the end of his nose. We have to mention his detailed study of the early history of stylometry (Grzybek 2012k) in European and Anglo-American academic communities. Similarly profound is his "discovery" of *phonological stoichiometry* (cf. Grzybek 2013a), a discipline working with phoneme frequencies and the distribution of vowels and consonants, used for language typological purposes.

Peter was usually not interested in "typical" or "classical" linguistic problems; however, his passion was dedicated to the frequency and modelling of "low-level units" like graphemes, letters, allophones etc. in the languages of the world. Again, a slight preference for analysing Slavic languages (he published studies on grapheme frequencies in Russian, Ukrainian and Slovene, among others) is noticeable, but it is obvious that he was mainly interested in one single question: which statistical models and which mechanisms are responsible for the regulation of grapheme frequencies in texts? He took both continuous and discrete statistical models into consideration, and he collaborated with many colleagues to crack this tough nut. Together with Ján Mačutek and the writer of these lines a monography on grapheme frequencies (cf. Grzybek/Kelih/Mačutek 2020) was already in preparation. It contains reflections on many central problems of modelling grapheme frequencies, including the linguistic definition of the measured units (letter, graphemes), appropriate sample size and different methods to approximate the sample size, appropriate corpus design and generally the homogeneity and heterogeneity problem in data modelling. From a statistical point of view the monograph contains an overview of the various possibilities of a descriptive statistical analysis of rank frequencies, including the calculation of various kinds of entropy and the related repeat rate, Ord's criteria etc. Without any doubt the most valuable part of the monography is the systematic discussion of a host of mathematical models which have been discussed in the past (geometric, Estoup, Zeta, Zipf-Mandelbrot, Good, negative hypergeometric, Whitworth, discrete Tuldava, Yule and Waring distribution), along with an overview of different methods of parameter estimation.

Coming to an end seems to be quite difficult, bearing in mind Peter's considerable contribution to quantitative text analysis and quantitative linguistics. In our short sketch only some key aspects could be highlighted. But to summarize properly, the predicates "highly concentrated", "obsessed with details" and "deep-going" are in any case not wrong. Moreover, Peter was far away from accepting any borders and limits, and his interest of going "beyond" can be seen in many of his interdisciplinary works. Among others, his refreshing studies on quantitative film studies and quantitative painting analysis have to be mentioned. Together with Veronika Koch he published a study (Grzybek/Koch 2012h) on Menzerath's law in film studies, based on the measured shot lengths. Grzybek (2001a) gives, embedded in general problems of cultural economy, an overview and re-analysis of available data (mostly rank frequencies) of musical constituents and the frequency of different colours of paints and mosaics. Not to forget to mention his (unpublished) studies about the population size of Russian cities

Postface – and finally a personal note

Science is made by people, and behind the "knowledge-miner" Peter (his ancestors came in the 19th century from Poland to the German Ruhr valley, formerly known as an important coalmining and heavy industry area in western Germany) there was more than only a DOI reference or an ORCID number. He had many good faces. Those who knew him well understand what is meant by this. It was his superhuman ambition, his even unearthly power and energy he could activate for a given project, which however could sometimes quickly turn into some kind of inflexibility in thinking or even stubbornness. In particular situations he had the admirable competence to make unbelievably simple facts and issues unbelievably complicated, and this holds true for the scientist Grzybek and the person Peter. On the other hand, we of course will remember Peter foremost because of his openness and heartiness. Many of his friends and colleagues will for sure remember their visits to his home, his "residence" in Sankt Marein, Holzmannsdorfberg 143, in south-east Styria near Graz. As a tendency, these visits could never be short, since it was a perfect occasion and one of the rare moments where he indeed found time for gathering, discussing, eating and drinking. The conversation with him was always on a high level, with esprit, wit, farsightedness and a healthy portion of humour. This is what will be really missed.

We are completely aware of the fact that in our short contribution not everything could be said and written about Peter, or even adequately presented and summarized. Time will help to sharpen the perspective on his contributions, his impact and his role in quantitative linguistics and text analysis. It seems only fair to give him the final word on how he saw his position in quantitative linguistics and exact text analysis. On his website <http://www.peter-grzybek.eu/>, which includes full texts of all his articles, he himself mentions his favourite areas of research: authorship identification, sound structures, micropoetics, rhythm, stylistics, complexity and compensation in texts, text laws, text typology, word frequencies, grapheme frequencies, word length, sentence length, verse structures and text typology.

References

- Best, K.-H.** (2005): Karl Marbe (1869 – 1953). *Glottometrics* 9, 74-76.
- Eismann, W.** (2019a). In memoriam Peter Grzybek (22.11.1957 – 29.05.2019). *Anzeiger für Slavische Philologie* 46, 7-10.
- Eismann, W.** (2019b). Nachruf auf Peter Grzybek (1957 – 2019). *Yearbook of Phraseology* 10, 200-202.
- Eismann, W., Deutschmann, P.** (2019). Nachruf. In memoriam Peter Grzybek (1957 – 2019). *Bulletin der deutschen Slavistik* 25, 51-52.
- Grzybek, P., Kelih, E., Mačutek, J.** (2020). *Alphabet Analyses. Quantitative Studies on Slavic Letter and Grapheme Frequencies*. Unpublished manuscript.
- Lotman, M., Pilshchikov, I., Lotman, M.-K.** (2019). Peter Grzybek (22.11.1957 – 29.05.2019). *Studia Metrica et Poetica* 6, 1, 119-122. DOI: 10.12697/smp.2019.6.1.05.
- Kelih, E., Köhler, R., Altmann, G.** (2020). Obituary. Peter Grzybek (1957 – 2019). *Glottometrics* 48, 1-2.
- Kelih, E., Köhler, R., Altmann, G.** (2019). Obituary. Peter Grzybek (1957 – 2019). *Journal of Quantitative Linguistics* 26, 4, 356-357. DOI: <https://doi.org/10.1080/09296174.2019.1651514>
- Kõiva, M.** (2019). In Memoriam. Peter Grzybek 22. november 1957 – 29. mai 2019. *Mäetagused* 74, 215-216.
- Mieder, W.** (2019a). In Memoriam: Peter Grzybek (November 22, 1957 – May 29, 2019). In: Soares, R. J. B., Lauhakangas, O. (eds.), *13th Interdisciplinary Colloquium on Proverbs, 3rd to 10th November 2019, at Tavira: 115-128. Portugal*. Tavira: Tipografia Tavirense.
- Mieder, W.** (2019b). Peter Grzybek (November 22, 1957 – May 29, 2019). In Memoriam. *Slavia Centralis* 12, 2, 121-132.
- Wimmer, G., Altmann, G.** (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 791-801*. Berlin, New York: de Gruyter. (= Handbücher zur Sprach- und Kommunikationswissenschaft 27)

I. Monographies

1983 m1

Grzybek, Peter. *Neurolinguistik und Fremdsprachenerwerb. Argumente für eine Aufwertung der rechten Gehirnhälfte des Lerners im Fremdsprachenunterricht*. Wiesbaden: Vieweg. (= LB-Papier 70)

1984 m2

Grzybek, Peter. *Lechts und Rinks kann man nicht velwechsern?!? Zur Neurosemiotik sprachlicher Kommunikation*. Trier: Linguistic Agency University of Trier. (= Series B 106)

1989 m3

Grzybek, Peter. *Studien zum Zeichenbegriff der sowjetischen Semiotik (Moskauer und Tartuer Schule)*, Bochum: Brockmeyer. (= Bochumer Beiträge zur Semiotik 23)

2000 m4

Permjakov, Grigorij L. and Peter Grzybek. *Die Grammatik der sprichwörtlichen Weisheit. Mit einer Analyse allgemein bekannter deutscher Sprichwörter*. Baltmannsweiler: Schneider-Verlag Hohengehren. (= Phraseologie und Parömiologie 4)

2009 m5

Popescu, Ioan-Iovitz; Gabriel Altmann, Peter Grzybek, Bijapur D. Jayaram, Reinhard Köhler, Viktor Krupa, Ján Mačutek, Regina Pustet, Ludmila Uhlířová, and Matummal N. Vidya. *Word Frequency Studies*. Berlin: Mouton de Gruyter. (= Quantitative Linguistics 64)

II. Omnibus volumes (edited)

1984 ov1

Grzybek, Peter and Wolfgang Eismann. *Semiotische Studien zum Sprichwort. Simple Forms Reconsidered I*. Tübingen: Narr. (= Special Issue of: Kodikas, Code – Ars Semeiotica. An International Journal of Semiotics 3/4)

1987 ov2

Eismann, Wolfgang and Peter Grzybek. *Semiotische Studien zum Rätsel. Simple Forms Reconsidered II*. Bochum: Brockmeyer. (= Bochumer Beiträge zur Semiotik 7)

1989 ov3

Eimermacher, Karl, Peter Grzybek and Georg Witte. *Issues in Slavic Literary and Cultural Theory. Studien zur Literatur- und Kulturtheorie in Osteuropa*. Bochum: Brockmeyer. (= Bochum Publications in Evolutionary Cultural Semiotics 21)

1991 ov4

Eimermacher, Karl and Peter Grzybek. *Zeichen – Text – Kultur. Studien zu den sprach- und kultursemiotischen Arbeiten von Vjač. Vs. Ivanov und V.N. Toporov*. Bochum: Brockmeyer. (= Bochum Publications in Evolutionary Cultural Semiotics 8)

1991 ov5

Grzybek, Peter. *Cultural Semiotics: Facts and Facets. Fakten und Facetten der Kultursemiotik.* Bochum: Brockmeyer. (= Bochumer Beiträge zur Semiotik 26)

1993 ov6

Grzybek, Peter. *Psychosemiotik – Neurosemiotik. Psychosemiotics – Neurosemiotics.* Bochum: Brockmeyer. (= Bochumer Beiträge zur Semiotik 41)

1994 ov7

Chlosta, Christoph, Peter Grzybek and Elisabeth Piirainen. *Sprachbilder zwischen Theorie und Praxis. Akten des Westfälischen Arbeitskreises «Phraseologie / Parömiologie» (1991/92).* Bochum: Brockmeyer. (= Studien zur Phraseologie und Parömiologie 2)

1997 ov8

Bernard, Jeff, Peter Grzybek and Gloria Withalm. *Geschichte, Kultur, Kulturgeschichte. Semiotische Aspekte.* Wien: ISSS. (= Special Issue of Semiotische Berichte 21, 2)

1998 ov9

Bernard, Jeff, Peter Grzybek, Vilmos Voigt and Gloria Withalm. *Peter Pázmány. Fokus Gemeinsamer Traditionen.* Wien: ISSS. (= Semiotische Berichte 22; 1-2)

2000 ov9

Bernard, Jeff, Peter Grzybek and Gloria Withalm. *Modellierungen von Geschichte und Kultur. Band I / Vol. I: Theoretische Grundlagen und 5. Österreichisch-Ungarisches Semiotik-Kolloquium.* Wien: ISSS.

2000 ov10

Bernard, Jeff, Peter Grzybek and Gloria Withalm. *Modellierungen von Geschichte und Kultur. Band II / Vol. II: Zeichen, Texte, Identitäten.* Wien: ISSS.

2000 ov11

Bernard, Jeff, Lada Čale Feldman, Peter Grzybek and Gloria Withalm. *Borders, Signs, Transitions.* Wien: ISSS. (= Special Issue of: S: European Journal for Semiotic Studies 12,2)

2001 ov11

Bernard, Jeff, Peter Grzybek and Gloria Withalm. *Form – Struktur – Komposition. Pragmatik und Rezeption. Akten des 3. internationalen bilateralen Symposiums "Offene Grenzen", 7.-8. Dezember 2001, Universität Graz.* Wien: ISSS. (= Special Issue of: Semiotische Berichte 26, 1-4)

2005 ov12

Bernard, Jeff, Jurij Fikfak and Peter Grzybek. *Text & Reality. Text & Wirklichkeit.* Ljubljana: ZRC.

2006 ov13

Grzybek, Peter. *Contributions to the Science of Text and Language. Word Length Studies and Related Issues.* Dordrecht, NL: Springer. (= Text, Speech and Language Technology 31)

2007 ov14

Deutschmann, Peter, Peter Grzybek, Ludwig Karničar and Heinrich Pfandl. *Kritik und Phrase. Festschrift für Wolfgang Eismann zum 65. Geburtstag*, Wien: Praesens.

2007 ov15

Grzybek, Peter., *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. 2, rev. paperback ed., Dordrecht: Kluwer. (= Text, Speech and Language Technology 31)

2007 ov16

Grzybek, Peter and Reinhard Köhler. *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of His 75th Birthday*. Berlin: Mouton de Gruyter. (= Quantitative Linguistics 62)

2010 ov17

Grzybek, Peter, Emmerich Kelih and Ján Mačutek. *Text and Language. Structures · Functions · Interrelations. Quantitative Perspectives*. Wien: Praesens.

2012 ov18

Naumann, Sven, Peter Grzybek, Relja Vulcanović and Gabriel Altmann. *Synergetic Linguistics. Text and Language as Dynamic Systems*. Wien: Praesens.

2014 ov19

Jesenšek, Vida and Peter Grzybek. *Phraseologie im Wörterbuch und Korpus / Phraseology in Dictionaries and Corpora*. Maribor: Filozofska fakulteta.

1989–1996 ov 20 – ov26

Fleischer, Michael and Peter Grzybek. *Znakolog. An International Yearbook of Slavic Semiotics* 1989, 1.

Fleischer, Michael and Peter Grzybek. *Znakolog. An International Yearbook of Slavic Semiotics* 1990, 2.

Fleischer, Michael and Peter Grzybek. *Znakolog. An International Yearbook of Slavic Semiotics* 1991, 3.

Fleischer, Michael and Peter Grzybek. *Znakolog. An International Yearbook of Slavic Semiotics* 1992, 4.

Fleischer, Michael and Peter Grzybek. *Znakolog. An International Yearbook of Slavic Semiotics* 1995, 5.

Fleischer, Michael and Peter Grzybek. *Znakolog. An International Yearbook of Slavic Semiotics* 1996, 6-7.

III. Papers in journals and articles

1983

- a Grzybek, Peter. “Die Komposition der Detektiverzählung – Psychosemiotische Überlegungen zur Struktur 'ihrer' Spannung.” *Kodikas, Code – Ars Semeiotica. An International Journal of Semiotics*, 3-4, pp. 219–35.

1984

- a Grzybek, Peter. "Grigorij L'vovič Permjakov (1919–1983)." *Proverbium. Yearbook of International Proverb Scholarship*, vol. 1, pp. 175–82.
- b Baur, Rupprecht and Peter Grzybek. "Zur (Re-)Integration natürlicher Verhaltensformen in den Fremdsprachenunterricht. Nonverbale Kommunikationsmittel im (fremdsprachlichen) Erwerbungsprozeß." *Zielsprache Deutsch*, vol. 15, no. 2, pp. 24–33.
- c Grzybek, Peter. "Bibliographie der Arbeiten G.L. Permjakovs." *Semiotische Studien zum Sprichwort. Simple Forms Reconsidered I*, edited by Peter Grzybek and Wolfgang Eismann. Tübingen: Narr, pp. 203–14. (= Special Issue of: *Kodikas, Code – Ars Semeiotica. An International Journal of Semiotics* 3/4)
- d Grzybek, Peter. "Grigorij L'vovič Permjakov (1919-1983)." *Semiotische Studien zum Sprichwort. Simple Forms Reconsidered I*, edited by Peter Grzybek and Wolfgang Eismann. Tübingen: Narr, pp. 199–202. (= Special Issue of: *Kodikas, Code – Ars Semeiotica. An International Journal of Semiotics* 3/4)
- e Grzybek, Peter. "How to Do Things with Some Proverbs. Zur Frage eines parömisches Minimums." *Semiotische Studien zum Sprichwort. Simple Forms Reconsidered I*, edited by Peter Grzybek and Wolfgang Eismann. Tübingen: Narr, pp. 351–58. (= Special Issue of: *Kodikas, Code – Ars Semeiotica. An International Journal of Semiotics* 3-4)
- f Grzybek, Peter. "Überlegungen zur semiotischen Sprichwortforschung." *Semiotische Studien zum Sprichwort. Simple Forms Reconsidered I*, edited by Peter Grzybek and Wolfgang Eismann. Tübingen: Narr, pp. 215–50. (Special Issue of: *Kodikas, Code – Ars Semeiotica. An International Journal of Semiotics* 3/4)
- g Grzybek, Peter. "Vorwort." *Semiotische Studien zum Sprichwort. Simple Forms Reconsidered I*, edited by Peter Grzybek and Wolfgang Eismann. Tübingen: Narr, pp. 197–98. (= Special Issue of: *Kodikas, Code – Ars Semeiotica. An International Journal of Semiotics* 3/4)
- h Grzybek, Peter. "Zur lexikographischen Erfassung von Sprichwörtern." *Semiotische Studien zum Sprichwort. Simple Forms Reconsidered I*, edited by Peter Grzybek and Wolfgang Eismann. Tübingen: Narr, pp. 345–50. (= Special Issue of: *Kodikas, Code – Ars Semeiotica. An International Journal of Semiotics* 3/4)
- i Grzybek, Peter. "Zur Psychosemiotik des Sprichworts." *Semiotische Studien zum Sprichwort. Simple Forms Reconsidered I*, edited by Peter Grzybek and Wolfgang Eismann. Tübingen: Narr, pp. 409–32. (= Special Issue of: *Kodikas, Code – Ars Semeiotica. An International Journal of Semiotics* 3/4)
- j Baur, Rupprecht S. and Peter Grzybek. "Argumente für die Integration von Gestik in den Fremdsprachenunterricht." *Sprache, Kultur und Gesellschaft*, edited by Wolfgang Kühlwein. Tübingen: Narr, pp. 63–72.

1985

- a Grzybek, Peter. "G.L. Permjakov (1919-1983)." *Scottish Slavonic Review*, vol. 5, 1985, pp. 170–71.
- b Grzybek, Peter. "[Rev.] Paremiologičeskie issledovanija. Sbornik statej. Sostavlenie i redakcija G.L. Permjakova. Predislovie T.V. Civ'jan. Moskva: Nauka, 1984." *Proverbium. Yearbook of International Proverb Scholarship*, vol. 2, pp. 339–51.
- c Grzybek, Peter. "[Rev.] V.Ja. Propp, Russkaja skazka. Leningrad: Izd. LGU, 1984." *Fabula*, vol. 26, pp. 377–80.

- d Grzybek, Peter. “[Rev.] G.L. Permjakov, 300 allgemeingebrauchliche russische Sprichwörter und sprichwörtliche Redensarten. Ein Illustriertes Nachschlagewerk für deutschsprechende. Russkij jazyk / VEB Verlag Enzyklopädie, Moskau / Leipzig, 1985.” *Zielsprache Russisch*, vol. 7, no. 2, pp. 61–63.
- e Grzybek, Peter and Rupprecht S. Baur. “Motorische Komponenten des Gedächtnisses und Fremdsprachenerwerb.” *Kongreßakten der 10. Arbeitstagung der Fremdsprachendidaktiker*, edited by Jürgen Donnerstag and Annelie Knapp-Potthoff. Tübingen: Narr, pp. 84–93.
- f Grzybek, Peter and Rupprecht S. Baur. “Neuropsychologische Grundlagen des Fremdsprachenerwerbs.” *Kongreßakten der 10. Arbeitstagung der Fremdsprachendidaktiker*, edited by Jürgen Donnerstag and Annelie Knapp-Potthoff. Tübingen: Narr, pp. 173–82.
- 1986**
- a Grzybek, Peter. “[Rev.] N.R. Norrick, How Proverbs Mean. Semantic Studies in English Proverbs. Berlin/New York/Amsterdam: Mouton, 1985.” *Proverbium. Yearbook of International Proverb Scholarship*, vol. 3, pp. 373–80.
- b Grzybek, Peter. “Zur Entwicklung semiotischer Sprichwortforschung in der UdSSR.” *Geschichte und Geschichtsschreibung der Semiotik. Fallstudien*, edited by Klaus D. Dutz and Peter Schmitter. Münster: MAKs Publikationen, pp. 383–409.
- c Grzybek, Peter and Rupprecht S. Baur. “Was ist das – Rätselstrukturen bei der Bedeutungserschließung im Fremdsprachenunterricht.” *Probleme und Perspektiven der Sprachlehrforschung. Bochumer Beiträge zum Fremdsprachenunterricht in Forschung und Lehre*, edited by Karl-Richard Bausch. Frankfurt/Main: Scriptor, pp. 145–62.
- 1987**
- a Grzybek, Peter. “Foundations of Semiotic Proverb Study.” *Proverbium. Yearbook of International Proverb Scholarship*, vol. 4, pp. 39–85.
- b Grzybek, Peter. “Überlegungen zur semiotischen Rätselforschung.” *Semiotische Studien zum Rätsel. Simple Forms Reconsidered II*, edited by Wolfgang Eismann and Peter Grzybek. Bochum: Brockmeyer, pp. 1–37. (= Bochumer Beiträge zur Semiotik 7)
- c Grzybek, Peter. “Zur Ontogenese des Rätselratens.” *Semiotische Studien zum Rätsel. Simple Forms Reconsidered II*, edited by Wolfgang Eismann and Peter Grzybek. Bochum: Brockmeyer, pp. 265–93. (= Bochumer Beiträge zur Semiotik 7)
- d Grzybek, Peter. “Zur Psychosemiotik des Rätsels.” *Semiotische Studien zum Rätsel. Simple Forms Reconsidered II*, edited by Wolfgang Eismann and Peter Grzybek. Bochum: Brockmeyer, pp. 247–64. (= Bochumer Beiträge zur Semiotik 7)
- e Eismann, Wolfgang and Peter Grzybek. “Vorwort der Herausgeber.” *Semiotische Studien zum Rätsel. Simple Forms Reconsidered II*, edited by Wolfgang Eismann and Peter Grzybek. Bochum: Brockmeyer, pp. ix–x. (= Bochumer Beiträge zur Semiotik 7)
- 1988**
- a Grzybek, Peter. “[Rev.] K.-D. Seemann (ed.), Beiträge zur russischen Volksdichtung. Wiesbaden: Harrassowitz, 1987.” *Fabula*, vol. 29, pp. 232–34.

- b Grzybek, Peter. "Sprichwort und Fabel: Überlegungen zur Beschreibung von Sinnstrukturen in Texten." *Proverbium. Yearbook of International Proverb Scholarship*, vol. 5, pp. 39–67.
- 1989**
- a Grzybek, Peter. "[Rev.] E. Kokare, Lettische und deutsche Sprichwortparallelen. Riga, 1988." *Fabula*, vol. 30, pp. 331–33.
- b Grzybek, Peter. "Some Remarks on the Notion of Sign in Jakobson's Semiotics and in Czech Structuralism." *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 1, pp. 113–28.
- c Grzybek, Peter. "Two Recent Publications in Soviet Structural Paremiology." *Proverbium. Yearbook of International Proverb Scholarship*, vol. 6, pp. 181–86.
- d Grzybek, Peter and Michael Fleischer. "Bibliographie zur slavischen Semiotik (1988)." *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 1, pp. 201–19.
- e Grzybek, Peter and Michael Fleischer. "Zum Geleit. Znakolog – Projekt und Programm." *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 1, pp. 9–13.
- f Grzybek, Peter. "Invariant Meaning Structures in Texts (Proverb and Fable)." *Issues in Slavic Literary and Cultural Theory. Studien zur Literatur- und Kulturtheorie in Osteuropa*, edited by Karl Eimermacher et al. Bochum: Brockmeyer, pp. 349–89.
- 1990**
- a Grzybek, Peter. "[Rev.] H. Rölleke (Hg.), «Redensarten des Volks, auf die ich immer horche.» Das Sprichwort in den KHM der Brüder Grimm. Bern: Lang, 1988." *Fabula*, vol. 31, pp. 174–75.
- b Grzybek, Peter and Michael Fleischer. "Bibliographie zur slavischen Semiotik (1989)." *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 2, pp. 291–306.
- c Grzybek, Peter and Michael Fleischer. "Bibliographie zur slavischen Semiotik (Nachträge 1988)." *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 2, pp. 283–90.
- d Grzybek, Peter and Michael Fleischer. "Geleitwort." *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 2, pp. 9–11.
- e Grzybek, Peter. "Kulturelle Stereotype und stereotype Texte." *Natürlichkeit der Sprache und der Kultur. Acta Colloquii*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 300–27. (= Bochumer Beiträge zur Semiotik 18)
- f Grzybek, Peter. "Rechts und Links im Alten Rußland." *Arbeitstreffen des Seminars für Slavistik der Ruhr-Universität Bochum anlässlich des Christianisierungsmillenniums Rußlands 18.11.1988 und 25.11.1988*, edited by Helmut Jachnow. Hagen: Rottmann, pp. 32–55. (= Bochumer Beiträge zur Semiotik 15)
- g Grzybek, Peter and Rupprecht S. Baur. "Untersuchungen zu einem parömisches Minimum im Deutschen." *Interkulturelle Kommunikation. Kongreßbeiträge zur 20. Jahrestagung der Gesellschaft für Angewandte Linguistik GAL E.V.*, edited by Bernd Spillner. Frankfurt/Main: Lang, pp. 220–23.
- h Baur, Rupprecht S. and Peter Grzybek. "Semiotics and Second Language Instruction Research." *Semiotics in the Individual Sciences*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 178–213.
- i Baur, Rupprecht S. and Peter Grzybek. "Sprachlehrforschung und Semiotik." *Semiotics in the Individual Sciences*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 249–86.

1991

- a Grzybek, Peter. "Rechts und Links von Mann und Frau – Jenseits des Geschlechterstreits. [Rev. J. van Leeuwen-Turnovcova, Rechts und Links in Europa. Wiesbaden, 1990]." *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 3, pp. 287–304.
- b Grzybek, Peter. "Sinkendes Kulturgut? Eine empirische Pilotstudie zur Bekanntheit deutscher Sprichwörter." *Wirkendes Wort*, vol. 41, no. 2, pp. 239–64.
- c Grzybek, Peter and Michael Fleischer. "Bibliographie zur slavistischen Semiotik (1990)." *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 3, pp. 381–90.
- d Grzybek, Peter and Michael Fleischer. "Bibliographie zur slavistischen Semiotik (Nachträge 1988-89)." *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 3, pp. 375–80.
- e Grzybek, Peter and Michael Fleischer. "Zum Geleit." *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 3, pp. 9–10.
- f Grzybek, Peter. "Das Drama um den Dorn im Fuß. Anmerkungen zu einem Motiv in Božena Němcová's Roman "Babička.""
Zur Poetik und Rezeption von Božena Němcová's "Babička", edited by Andreas Guski. Wiesbaden: Harrassowitz, pp. 184–88. (= Veröffentlichungen der Abteilung für Slavische Sprachen und Literaturen des Osteuropa-Instituts (Slavisches Seminar) an der Freien Universität Berlin 75)
- g Grzybek, Peter. "Das Sprichwort im literarischen Text." *Sprichwörter und Redensarten im interkulturellen Vergleich*, edited by Annette Sabban. Opladen: Westdeutscher Verlag, pp. 187–205.
- h Grzybek, Peter. "Der Körper im Rätsel. Zum Verhältnis von Mikrokosmos, Mesokosmos und Makrokosmos am Beispiel serbokroatischer Volksrätsel." *Körper, Essen und Trinken im Kulturverständnis der Balkanvölker. Beiträge zur Tagung vom 19.-24. November 1989 in Hamburg*, edited by Dagmar Burkhart. Wiesbaden: Harrassowitz, 1991, pp. 195–216.
- i Grzybek, Peter. "Neurosemiotik – Kultursemiotik. Farbwahrnehmung und Farbbezeichnung als Beispiel eines integrativen Konzepts." *Zeichen – Text – Kultur. Studien zu den Sprach- und Kultursemiotischen Arbeiten von Vjač.Vs. Ivanov und V.N. Toporov*, edited by Karl Eimermacher and Peter Grzybek. Bochum: Brockmeyer, pp. 97–186. (= Bochum Publications in Evolutionary Cultural Semiotics 8)
- j Grzybek, Peter. "Textsemiotik: Semiotik des Textes?" *Problemy lingvistiki teksta. Probleme der Textlinguistik II*, edited by Adam E. Suprun and Helmut Jachnow. Minsk, pp. 5–35.
- k Grzybek, Peter. "Zur semantischen Funktion der sprichwörtlichen Wendungen in Božena Němcová's "Babička.""
Zur Poetik und Rezeption von Božena Němcová's "Babička", edited by Andreas Guski. Wiesbaden: Harrassowitz, pp. 184–88. (= Veröffentlichungen der Abteilung für Slavische Sprachen und Literaturen des Osteuropa-Instituts (Slavisches Seminar) an der Freien Universität Berlin 75)
- l Grzybek, Peter and Karl Eimermacher. "Bibliographie der Arbeiten Vjač.Vs. Ivanovs." *Zeichen – Text – Kultur. Studien zu den Sprach- und Kultursemiotischen Arbeiten von Vjač.Vs. Ivanov und V.N. Toporov*, edited by Karl Eimermacher and Peter Grzybek. Bochum: Brockmeyer, pp. 285–350. (= Bochum Publications in Evolutionary Cultural Semiotics 8)
- m Grzybek, Peter and Karl Eimermacher. "Bibliographie der Arbeiten V.N. Toporovs." *Zeichen – Text – Kultur. Studien zu den sprach- und kultursemiotischen Arbeiten von Vjač.Vs. Ivanov und V.N. Toporov*, edited by Karl Eimermacher and Peter Grzybek.

Bochum: Brockmeyer, pp. 353–413. (= Bochum Publications in Evolutionary Cultural Semiotics 8)

1992

- a Grzybek, Peter. “Mikrokosmos, mezoskosmos, makrokosmos. Model na sveta i poetika vā folklorā (Po primeri ot bālgarskite nardoni gatanki).” *Bālgarski folklor*, vol. 18, no. 1, pp. 5–23.
- b Grzybek, Peter and Michael Fleischer. “Zum Geleit.” *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 4, p. 9.
- c Fleischer, Michael and Peter Grzybek. “Bibliographie zur slavischen Semiotik (Nachträge 1988-1991).” *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 4, pp. 325–28.
- d Grzybek, Peter. “Ein Baustein zur Geschichte der serbokroatischen Rätsel.” *Studia Phraseologica et alia. Festschrift für Josip Matešić zum 65. Geburtstag*, edited by Wolfgang Eismann and Jürgen Petermann. München: Sagner, pp. 187–202.
- e Grzybek, Peter. “Probleme der Sprichwort-Lexikographie (Parömiographie): Definition – Klassifikation – Selektion.” *Worte, Wörter, Wörterbücher. Lexikographische Beiträge zum Essener Linguistischen Kolloquium*, edited by Gregor Meder and Andreas Dörner. Tübingen: Niemeyer, pp. 195–223. (= Lexicographica, Series Maior 42)
- f Grotjahn, Rüdiger, Anna Tóthné Litovkina, Peter Grzybek, Chlosta Christoph and Undine Roos. “Statistical Methods in the Study of Proverb Knowledge. An Analysis of the Knowledge of Proverbs in Contemporary Hungarian Culture (Tolna County).” *Zeichen / Kultur. Akten des 3. Österreichisch-Ungarischen Semiotik-Kolloquiums, Szombathely / Velem 1992*, edited by Jeff Bernard et al. Wien: ÖGS, p. 275.

1993

- a Grzybek, Peter and Christoph Chlosta. “Grundlagen der empirischen Sprichwortforschung.” *Proverbium. Yearbook of International Proverb Scholarship*, vol. 10, pp. 89–128.
- b Grzybek, Peter, Christoph Chlosta, Zorica Stanković-Arnold and Andreas Steczka. “Das Sprichwort in der überregionalen Tagespresse. Eine systematische Analyse zum Vorkommen von Sprichwörtern in den Tageszeitungen 'Die Welt', 'Frankfurter Allgemeine Zeitung' und 'Süddeutsche Zeitung'.” *Wirkendes Wort*, vol. 43, no. 3, pp. 671–95.
- c Grzybek, Peter, Christoph Chlosta, Zorica Stanković-Arnold and Andreas Steczka. “Der Weisheit der Gasse auf der Spur. Eine empirische Pilotstudie zur Bekanntheit kroatischer Sprichwörter.” *Zeitschrift für Balkanologie*, vol. 29, 1993, pp. 85–98.
- d Grzybek, Peter. “Neurosemiotic Remarks on Text Processing.” *Narrative Discourse in Normal Aging and Neurologically Impaired Adults*, edited by Hiram H. Brownell and Yves Joannette. San Diego, CA: Singular Publishing, pp. 47–74.

1994

- a Grzybek, Peter. “Semiotics of History – Historical Cultural Semiotics? [Rev.] B.A. Uspenskij, Semiotik der Geschichte. Wien, 1991.” *Semiotica*, vol. 98, 3-4, pp. 341–56.
- b Grzybek, Peter. “The Concept of 'Model' in Soviet Semiotics.” *Russian Literature*, vol. 36, no. 3, pp. 285–300.

- c Grzybek, Peter, Rupprecht S. Baur and Christoph Closta. "Perspektiven einer empirischen Parömiologie (Sprichwortforschung)." *Zet. Zeitschrift für Empirische Textforschung*, vol. 1, pp. 94–98.
- d Eismann, Wolfgang and Peter Grzybek. "In memoriam Jurij Michajlovič Lotman (1922–1993)." *Zeitschrift für Semiotik*, vol. 16, pp. 105–16.
- e Grzybek, Peter. "Adage." *Simple Forms. An Encyclopaedia of Simple Text-Types in Lore and Literature*, edited by Walter A. Koch. Bochum: Brockmeyer, p. 1. (= Bochum Publications in Evolutionary Cultural Semiotics 4)
- f Grzybek, Peter. "Bemerkungen zum Modellbegriff in der Semiotik (unter besonderer Berücksichtigung der Moskauer / Tartuer Schule)." *Zeichen, Sprache, Bewußtsein*, edited by Jeff Bernard and Katalin Neumer. Wien: ÖGS/ISSS, pp. 117–38. (= Österreichisch-Ungarische Dokumente zur Semiotik und Philosophie 2)
- g Grzybek, Peter. "Blason Populaire." *Simple Forms. An Encyclopaedia of Simple Text-Types in Lore and Literature*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 19–25. (= Bochum Publications in Evolutionary Cultural Semiotics 4)
- h Grzybek, Peter. "Častuška." *Simple Forms. An Encyclopaedia of Simple Text-Types in Lore and Literature*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 51–56. (= Bochum Publications in Evolutionary Cultural Semiotics 4)
- i Grzybek, Peter. "Comparison." *Simple Forms. An Encyclopaedia of Simple Text-Types in Lore and Literature*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 68–74. (= Bochum Publications in Evolutionary Cultural Semiotics 4)
- j Grzybek, Peter. "Ėmpiričeskaja semiotika kul'tury na primere issledovanija poslovic i ispol'zovanjem rezul'tatov probnogo opytnogo izučenija izvestnosti chorvackich poslovic." *Znaki Balkana*, edited by Nikkolaj P. Grincer. Moskva: Radiks, pp. 312–38. (= Balkanskije Čtenija 2)
- k Grzybek, Peter. "Bemerkungen zum Modellbegriff in der Semiotik (unter besonderer Berücksichtigung der Moskauer / Tartuer Schule)." *Zeichen, Sprache, Bewußtsein*, edited by Jeff Bernard and Katalin Neumer. Wien: ÖGS/ISSS, pp. 117–38. (= Österreichisch-Ungarische Dokumente zur Semiotik und Philosophie 2)
- l Grzybek, Peter. "Foundations of Semiotic Proverb Study." *Wise Words. Essays on the Proverb*, edited by Wolfgang Mieder. New York: Garland, pp. 31–71. (= Garland Folklore Casebooks 6)
- m Grzybek, Peter. "Märchen." *Simple Forms. An Encyclopaedia of Simple Text-Types in Lore and Literature*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 144–57. (= Bochum Publications in Evolutionary Cultural Semiotics 4)
- n Grzybek, Peter. "Poetik und Weltmodell. Mikro-, Meso-, Makrokosmos und V.N. Toporovs Anagrammtheorie der indoeuropäischen Poetik." *Die Welt der Lyrik. 15. Bochumer Kolloquium zur Evolution der Kultur (IFIKS 15), im Juni 1991*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 68–107. (= Bochumer Beiträge zur Semiotik 39)
- o Grzybek, Peter. "Proverb." *Simple Forms. An Encyclopaedia of Simple Text-Types in Lore and Literature*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 227–41. (= Bochum Publications in Evolutionary Cultural Semiotics 4)
- p Grzybek, Peter and Wolfgang Eismann. "Riddle." *Simple Forms. An Encyclopaedia of Simple Text-Types in Lore and Literature*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 242–54. (= Bochum Publications in Evolutionary Cultural Semiotics 4)
- q Grzybek, Peter. "The Culture of Nature. The Semiotic Dimensions of Microcosm, Mesocosm, and Macrocosm." *Origins of Semiosis. Sign Evolution in Nature and*

- Culture*, edited by Winfried Nöth. Berlin: Mouton de Gruyter, pp. 121–38. (= Approaches to Semiotics 116)
- r Grzybek, Peter. “Wellerism.” *Simple Forms. An Encyclopaedia of Simple Text-Types in Lore and Literature*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 286–92. (= Bochum Publications in Evolutionary Cultural Semiotics 4)
- s Grzybek, Peter. “Winged Word.” *Simple Forms. An Encyclopaedia of Simple Text-Types in Lore and Literature*, edited by Walter A. Koch. Bochum: Brockmeyer, pp. 293–98. (= Bochum Publications in Evolutionary Cultural Semiotics 4)
- t Grzybek, Peter, Christoph Closta and Undine Roos. “Ein Vorschlag zur Klassifikation von Sprichwortvarianten in der empirischen Sprichwortforschung.” *Europhras 92. Tendenzen der Phraseologieforschung*, edited by Barbara Sandig. Bochum: Brockmeyer, pp. 221–56. (= Studien zur Phraseologie und Parömiologie 1)
- u Grzybek, Peter and Wolfgang Eismann. “Phraseologie und sprichwörtliche Redensart. Vom Mythos der Untrennbarkeit.” *Sprachbilder zwischen Theorie und Praxis. Akten des Westfälischen Arbeitskreises «Phraseologie / Parömiologie» (1991/92)*, edited by Christoph Chlosta, Peter Grzybek and Elisabeth Piirainen. Bochum: Brockmeyer, pp. 89–132.
- v Chlosta, Christoph, Peter Grzybek and Undine Roos. “Wer kennt denn heute noch den Simrock? Ergebnisse einer empirischen Untersuchung zur Bekanntheit traditioneller deutscher Sprichwörter.” *Sprachbilder zwischen Theorie und Praxis. Akten des Westfälischen Arbeitskreises «Phraseologie / Parömiologie» (1991/92)*, edited by Christoph Chlosta, Peter Grzybek and Elisabeth Piirainen. Bochum: Brockmeyer, pp. 31–60.

1995

- a Grzybek, Peter. “Foundations of Semiotic Proverb Study.” *De Proverbio. An Electronic Journal of International Paremiology*, vol. 1, no. 1, <https://deproverbio.com/foundations-of-semiotic-proverb-study/>. (last access 07/25/2020)
- b Grzybek, Peter. “[Rev.] K. Bjuler, “Teorija jazyka.” Moskva, 1993.” *Znakolog. An International Yearbook of Slavic Semiotics*, 6-7, pp. 277–84.
- c Grzybek, Peter. “[Rev.] U. Grabmüller; M. Katz (eds.), Zwischen Anpassung und Widerspruch. Beiträge zur Frauenforschung am Osteuropa-Institut der Freien Universität Berlin. Berlin/Wiesbaden, 1993.” *Information Interuniversitäre Koordinationsstelle für Frauenforschung und Frauenstudien Graz*, vol. 2, no. 2, pp. 81–83.
- d Grzybek, Peter. “S.I. Karcevskijs Thesen «Vom asymmetrischen Dualismus des sprachlichen Zeichens».” *Znakolog. An International Yearbook of Slavic Semiotics*, 6-7, pp. 11–17.
- e Grzybek, Peter. “Semiotički pristup narodnoj zagoneci.” *Radovi. Razdio filoloških znanosti*, 22-23, pp. 99–126.
- f Fleischer, Michael and Peter Grzybek. “Zum Geleit.” *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 5, p. 9.
- g Chlosta, Christoph and Peter Grzybek. “Empirical and Folkloristic Paremiology: Two to Quarrel or to Tango?” *Proverbium. Yearbook of International Proverb Scholarship*, vol. 12, 1995, pp. 67–85.
- h Civ'jan, Tat'jana V. and Peter Grzybek. “Ein Hühnergedächtnis im Flugzeug. Spielerische, surrealistische und absurde Textkonstruktion.” *Znakolog. An International Yearbook of Slavic Semiotics*, vol. 5, 1995, pp. 33–72.
- i Grzybek, Peter. “Bachtinskaja semiotika i moskovsko-tartuskaja škola.” *Lotmanovskij sbornik I*, edited by Evgenij V. Permjakov. Moskva: IC-Garant, pp. 240–59.

- j Grzybek, Peter. "Zum Aufkommen des Kulturbegriffs in Russland." *Kulturauffassungen in der literarischen Welt Rußlands. Kontinuitäten und Wandlungen im 20. Jahrhundert*, edited by Christa Ebert. Berlin: Berlin-Verlag Spitz, pp. 47–75.
- k Grzybek, Peter. "Zur Frage der Satzlänge von Sprichwörtern (unter besonderer Berücksichtigung deutscher Sprichwörter)." *Von der Einwortmetapher zur Satzmetapher. Akten des Westfälischen Arbeitskreises «Phraseologie / Parömiologie»*, edited by Rupprecht S. Baur and Christoph Chlosta. Bochum: Brockmeyer, pp. 203–17. (= Studien zur Phraseologie und Parömiologie 6)
- l Baur, Rupprecht S., Christoph Chlosta and Peter Grzybek. "Verbale und nonverbale Phraseologie." *Well Schrift – De Bliff! Festgabe für Irmgard Simon zum 80. Geburtstag am 6. Oktober 1995*, edited by Robert Damme. Münster: Aschendorff, 1995, pp. 3–29.
- 1996**
- a Grzybek, Peter, Rupprecht S. Baur and Christoph Closta. "Das Projekt »Sprichwörter-Minima im Deutschen und Kroatischen«: What Is Worth Doing – Do It Well!". *Muttersprache*, no. 2, pp. 162–79.
- b Grzybek, Peter. "Anmerkungen zur modernen slowenischen Prosa, mit besonderem Hinblick auf das sog. weibliche Schreiben." *Österreich in neuer Nachbarschaft*, edited by Gerhard Fink. Wien, Linz: Gessellschaft für Ostkooperation, pp. 392–96.
- c Grzybek, Peter. "Podstawy semiotyki ogólnej. (Ju.K. Lekomcev, 1929-1984)." *W świecie znaków. Księga pamiątkowa ku czci Profesora Jerzego Pelca*, edited by J.J. Jadacki and W. Strawiński. Warszawa: Polskie Towarzystwo Filozoficzne, pp. 71–80.
- d Grzybek, Peter. "Ranko Marinković's "Hände" und die Rechts – Links – Problematik." *Diskurs der Schwelle. Aspekte der kroatischen Gegenwartsliteratur*, edited by Dagmar Burkhart and Vladimir Biti. Frankfurt/Main: Lang, pp. 43–69. (= Heidelberger Publikationen zur Slavistik B, Literaturwissenschaftliche Reihe 4)
- 1997**
- a Grzybek, Peter. "Remarks on Obsolescence and Familiarity with Traditional Croatian Proverbs. III: Mijat Stojanović's "Sbirka narodnih posloviceah, riečih i izrazah" (1866)." *Narodna umjetnost*, vol. 34, no. 1, pp. 201–23.
- b Grzybek, Peter. "Anmerkungen zur Obsoletheit und Bekanntheit traditioneller kroatischer Sprichwörter. I: Die «Poslovice» von Gjuro Daničić (1871)." *Prvi Hrvatski Slavistički Kongres. Zbornik radova*, edited by Stjepan Damjanović, vol. 2. Zagreb: Hrvatsko Filološko Društvo, pp. 149–63.
- c Grzybek, Peter. "Text(e) und Welt(en): Textwelt(en)?" *Welt der Zeichen, Welt der Dinge. World of Signs, World of Things. Akten des 8. Symposiums der österreichischen Gesellschaft für Semiotik, Innsbruck 1993*, edited by Jeff Bernard et al. Wien: ÖGS, ISSS, pp. 39–72. (= Angewandte Semiotik 15)
- d Chlosta, Christoph and Peter Grzybek. "Sprichwortkenntnis in Deutschland und Österreich. Empirische Ergebnisse zu einigen mehr oder weniger gewagten Hypothesen." *Österreichisches Deutsch und andere nationale Varietäten plurizentrischer Sprachen in Europa. Empirische Analysen*, edited by Rudolf Muhr and Richard Schrod. Wien: Holder-Pichler-Tempsky, pp. 243–61.

1998

- a Grzybek, Peter. "Anmerkungen zur Obsoletheit und Bekanntheit traditioneller kroatischer Sprichwörter. II: Die «Hrvatske narodne poslovice» von Viktor Juraj Skarpa (1990)." *Suvremena lingvistika*, 41/42, pp. 183–98.
- b Grzybek, Peter. "Ivo Andrić's «Put Alije Đerzeleza»: The Dethronement of Heroism?" *Essays in Poetics. The Journal of the British Neo-Formalist Circle*, vol. 23, pp. 180–205.
- c Grzybek, Peter. "Bogatyrëv, P.G. (1893-1971)." *Encyclopedia of Semiotics*, edited by Paul Bouissac. New York: Oxford University Press, pp. 88–89.
- d Grzybek, Peter. "Explorative Untersuchungen zur Wort- und Satzlänge kroatischer Sprichwörter. (Am Beispiel der 'Poslovice' von Đruo Daničić, 1871)." *Polytropa. K 70-letju Vladimira Nikolaeviča Toporova*, edited by Tat'jana M. Nikolaeva. Moskva: Indrik, pp. 447–65.
- e Grzybek, Peter. "Karcevskij, S.I. (1884–1955)." *Encyclopedia of Semiotics*, edited by Paul Bouissac. New York: Oxford University Press, pp. 335–36.
- f Grzybek, Peter. "Komparative und interkulturelle Parömiologie. Methodologische Bemerkungen und empirische Befunde." *EUROPHRAS 95. Europäische Phraseologie im Vergleich*, edited by Wolfgang Eismann. Bochum: Brockmeyer, pp. 263–82. (= Studien zur Phraseologie und Parömiologie 15)
- g Grzybek, Peter. "Lotman, Ju.M. (1922–1993)." *Encyclopedia of Semiotics*, edited by Paul Bouissac. New York: Oxford University Press, pp. 375–77.
- h Grzybek, Peter. "Moscow-Tartu School." *Encyclopedia of Semiotics*, edited by Paul Bouissac. New York: Oxford University Press, pp. 422–25.
- i Grzybek, Peter. "Mukařovský, Jan (1891–1975)." *Encyclopedia of Semiotics*, edited by Paul Bouissac. New York: Oxford University Press, pp. 425–27.
- j Grzybek, Peter. "Paroemiology." *Encyclopedia of Semiotics*, edited by Paul Bouissac. New York: Oxford University Press, pp. 470–74.
- k Grzybek, Peter. "Peter Pázmány – der kulturgeschichtliche Kontext seiner Grazer Zeit(en). Einführende Bemerkungen unter besonderer Berücksichtigung der kroatisch-ungarischen Adelsbeziehungen." *Peter Pázmány. Fokus gemeinsamer Traditionen*, edited by Jeff Bernard, Peter Grzybek, Vilmos Voigt and Gloria Withalm. Wien: ISSS, 1998, pp. 15–40. *Semiotische Berichte* 22, 1/2.
- l Grzybek, Peter. "Prague School." *Encyclopedia of Semiotics*, edited by Paul Bouissac. New York: Oxford University Press, pp. 517–21.
- m Grzybek, Peter. "Prolegomena zur Bildlichkeit in Sprichwörtern." *Im Zeichen-Raum. Festschrift für Karl Eimermacher zum 60. Geburtstag*, edited by Anne Hartmann. Dortmund: Projekt-Verlag, pp. 133–52. (= Dokumente und Analysen zur russischen und sowjetischen Kultur 11)
- n Grzybek, Peter. "Russian Formalism." *Encyclopedia of Semiotics*, edited by Paul Bouissac. New York: Oxford University Press, pp. 550–53.
- o Grzybek, Peter. "Sprichwort – Wahrwort. Das Sprichwort zwischen Norm und Denkmodell." *Kultur und Lebenswelt als Zeichenphänomene. Akten eines Internationalen Kolloquiums zum 70. Geburtstag von Ivan Bystrina und Ladislav Tondl, Wien, Dezember 1994*, edited by Jeff Bernard and Gloria Withalm. Wien: ISSS, pp. 127–48.
- p Grzybek, Peter. "Trubeckoj, N.S. (1891–1938)." *Encyclopedia of Semiotics*, edited by Paul Bouissac, New York: Oxford University Press, pp. 617–18.

1999

- a Grzybek, Peter. “[Rev.] P. Đurčo (Ed.), Europhras ’97. Phraseology and Paremiology. Bratislava, 1998.” *Anzeiger für Slavische Philologie*, vol. 27, pp. 219–22.
- b Grzybek, Peter. “Wie lang sind slowenische Sprichwörter? Zur Häufigkeitsverteilung von (in Worten berechneten) Satzlängen slowenischer Sprichwörter.” *Anzeiger für Slavische Philologie*, vol. 27, pp. 87–108.
- c Grzybek, Peter. “Empirische Befunde zur Theorie stereotyper Vergleiche. Bosnische Vergleiche auf dem Prüfstand.” *Wörter in Bildern – Bilder in Wörtern. Beiträge zur Phraseologie und Sprichwortforschung aus dem Westfälischen Arbeitskreis*, edited by Rupprecht S. Baur, Christopf Closta and Elisabeth Piirinen. Baltmannsweiler: Schneider-Verlag Hohengehren, 1999, pp. 177–98. (= Phraseologie und Parömiologie 1)
- d Grzybek, Peter. “Randbemerkungen zur Korrelation von Wort- und Silbenlänge im Kroatischen.” *Die grammatischen Korrelationen. GraLiS-1999*, edited by Branko Tošović. Graz: Institut für Slawistik der Karl-Franzens-Universität, pp. 57–67.
- e Grzybek, Peter. “South Slavic Erotic Folklore. Traditional Erotic Phraseology from Dalmatia.” *Sex and the Meaning of Life / Life and the Meaning of Sex. Akten des "Wiener Semiotischen Ateliers" über "Sex and the Meaning of Life / Life and the Meaning of Sex" (Wien, 26.-29. März 1998)*, edited by Jeff Bernard et al. Wien: ISSS, 1999, pp. 131–54. (= Semiotische Berichte 23, 1-4)
- f Grzybek, Peter. “Sowjetische und russische Konzepte der Semiotik.” *Handbuch der sprachwissenschaftlichen Russistik und ihrer Grenzdisziplinen*, edited by Helmut Jachnow. Wiesbaden: Harrassowitz, 1999, pp. 1274–305. (= Slavistische Studienbücher NF 8)

2000

- a Grzybek, Peter. “Pogostnostna analiza besed iz elektronskego korpusa slovenskih besedel.” *Slavistična revija*, vol. 48, no. 2, pp. 141–57.
- b Grotjahn, Rüdiger and Peter Grzybek. “Methodological Remarks on Statistical Analyses in Empirical Paremiology.” *Proverbium. Yearbook of International Proverb Scholarship*, vol. 17, pp. 121–32.
- c Grzybek, Peter. “Apophthegma.” *Modellierungen von Geschichte und Kultur. Band II / Vol. II: Zeichen, Texte, Identitäten*, edited by Jeff Bernard, Peter Grzybek and Gloria Withalm. Wien: ISSS, 2000, pp. 13–14.
- d Grzybek, Peter. “G.L. Permajakovs Grammatik der sprichwörtlichen Weisheit.” *Die Grammatik der sprichwörtlichen Weisheit. Mit einer Analyse allgemein bekannter deutscher Sprichwörter*, edited by Grigorij L. Permjakov and Peter Grzybek. Baltmannsweiler: Schneider-Verlag Hohengehren, pp. 1–41. (= Phraseologie und Parömiologie 4)
- e Grzybek, Peter. “Južnoslovenski erotski folklor. Zapažnja o narodnoj erotskoj frazeologiji iz Dalmacije.” *Erotsko u folkloru Slovena. Zbornik radova Priredio Dejan Ajdačić*, edited by Predrag Marković, Stubovi Kulture, pp. 295–325. (= Biblioteka Lazulit 8)
- f Grzybek, Peter. “Slawistik und Kulturwissenschaft(en).” *Kultur – Wissenschaft – Russland. Beiträge zum Verhältnis von Kultur und Wissenschaft aus slawistischer Sicht*, edited by Wolfgang Eismann and Peter Deutschmann. Frankfurt/Main: Lang, pp. 93–133.
- g Grzybek, Peter. “Zum Status der Untersuchung von Satzlängen in der Sprichwortforschung. methodologische Vor-Bemerkungen.” *Slovo vo vremeni i*

prostranstve. K 60-letiju professora V.M. Mokienko, edited by G.A. Lilič et al. Moskva: Folio-Press, pp. 430–57.

- h Grzybek, Peter and Christoph Chlosta. “Versuch macht klug! Logisch-semiotische Klassifikation allgemein bekannter deutscher Sprichwörter.” *Die Grammatik der sprichwörtlichen Weisheit. Mit einer Analyse allgemein bekannter deutscher Sprichwörter*, edited by Grigorij L. Permjakov and Peter Grzybek. Baltmannsweiler: Schneider-Verlag Hohengehren, pp. 169–99. (= Phraseologie und Parömiologie 4)
- 2001
- a Grzybek, Peter. “Kultur–Ökonomie. Zur Häufigkeit text-konstitutiver Elemente.” *Wiener Slawistischer Almanach, Sonderband*, vol. 54, pp. 485–509.
- b Grzybek, Peter. “[Rev.] S. Schmidt-Knaebel, Textlinguistik der einfachen Form. Die Abgrenzung von Märchen, Sage und Legende zur literarischen Kunstform der Novelle. Frankfurt: Lang, 1999.” *Fabula*, vol. 41, pp. 350–54.
- c Grzybek, Peter. “Mythos und Ritual, Natur und Kultur. Synoptische Einleitungsbemerkungen.” *Mythen, Rituale, Simulacra. Semiotische Perspektiven*, edited by Jeff Bernard and Gloria Wiltham. Wien: ISSS, 2001, pp. 25–38.
- d Grzybek, Peter. “Zur Satz- und Teilsatzlänge formelhafter zweigliedriger Sprichwörter.” *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift für Luděk Hřebíček*, edited by Ludmila Uhlířová, Gejza Wimmer, Gabriel Altmann and Reinhard Köhler. Trier: wvt, pp. 64–75.
- e Grzybek, Peter. “Versuchen wir einmal, die Kräfte aus dem Gleichgewicht zu bringen... Quantitative Aspekte von Puškins 'Evgenij Onegin' und 'Domik v Kolomne'.” *Form – Struktur – Komposition. Pragmatik und Rezeption. Akten des 3. Internationalen Bilateralen Symposiums "Offene Grenzen", 7.-8. Dezember 2001, Universität Graz*, edited by Jeff Bernard, Peter Grzybek, Anton Pokrivčák and Gloria Withalm. Wien: ISSS, pp. 305–35. (= Special Issue of: Semiotische Berichte 26, 1-4)
- 2002
- a Grzybek, Peter. “Quantitative Aspekte slawischer Texte (am Beispiel von Puškins 'Evgenij Onegin').” *Wiener Slawistisches Jahrbuch*, vol. 48, pp. 21–36.
- b Grzybek, Peter. “[Rev.] Best, K.-H. (Ed.), Häufigkeitsverteilungen in Texten. Göttingen: Peust & Gutschmidt, 2001.” *Journal of Quantitative Linguistics*, vol. 9, no. 1, pp. 86–97.
- c Grzybek, Peter and Gabriel Altmann. “Oscillation in the Frequency-Length Relationship.” *Glottometrics*, vol. 6, pp. 97–107.
- d Grzybek, Peter and Ernst Stadlober. “The Graz Project on Word Length (Frequencies). Project Report.” *Journal of Quantitative Linguistics*, vol. 9, no. 2, pp. 187–92.
- e Kelih, Emmerich and Peter Grzybek. “Wortlängen in Texten. Internationales Symposium zur quantitativen Textanalyse.” Tagungsbericht. *etc. Empirische Text- und Kulturforschung*, vol. 2, pp. 89–91.
- f Antić, Gordana, Emmerich Kelih and Peter Grzybek. “Word Length in Texts. An International Symposium on Quantitative Text Analysis (Conference Report).” *Journal of Quantitative Linguistics*, vol. 9, no. 3, pp. 275–79.
- g Grzybek, Peter and Rudi Schlatter. “Zur Satzlänge deutscher Sprichwörter. Ein Neu-Ansatz.” *Phraseologie in Raum und Zeit. Akten der 10. Tagung des Westfälischen Arbeitskreises 'Phraseologie/Parömiologie' (Münster 2001)*, edited by Elisabeth Piirainen and Ilpo Piirainen. Baltmannsweiler: Schneider-Verlag Hohengehren, pp. 287–305.
- h Eismann, Wolfgang and Peter Grzybek. “Kultur.” *Lexikon der Russischen Kultur*, edited by Norbert Franz, Darmstadt: Primus, pp. 248–57.

2003

- a Grzybek, Peter. "Viktor Jakovlevič Bunjakovskij" *Glottometrics*, vol. 6, 2003, pp. 103–06.
- b Grzybek, Peter and Emmerich Kelih. "Graphemhäufigkeiten (am Beispiel des Russischen). Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen." *Anzeiger für Slavische Philologie*, vol. 31, pp. 131–62.
- c Kelih, Emmerich, Peter Grzybek and Ernst Stadlober. "Das Grazer Projekt zu Wortlängen(häufigkeiten)." *Glottometrics*, vol. 6, pp. 94–102.
- d Grzybek, Peter. "Zur lexikalischen Struktur von Sprichwörtern." *Flut von Texten – Vielfalt der Kulturen (Ascona 2001). Zur Methodologie und Kulturspezifität der Phraseologie*, edited by Harald Burger et al. Baltmannsweiler: Schneider-Verlag Hohengehren, pp. 99–116. (Phraseologie und Parömiologie 14)
- e Grzybek, Peter and Ernst Stadlober. "Zur Prosa Karel Čapeks. Einige quantitative Bemerkungen." *Rusistika - Slavistika - Lingvistika. Festschrift für Werner Lehfeldt zum 60. Geburtstag*, edited by Sebastian Kempgen et al. München: Otto Sanger, pp. 474–88.

2004

- a Grzybek, Peter. "A Quantitative Approach to Lexical Structure of Proverbs." *Journal of Quantitative Linguistics [= Special issue: Festschrift in honour of Professor Raimund Piotrowski]*, vol. 11, 1-2, pp. 79–92.
- b Grzybek, Peter. "Nikolaj Gavrilovič Černyševskij. A Forerunner of Quantitative Stylistics in Russia." *Glottometrics*, vol. 7, pp. 91–93.
- c Grzybek, Peter and Emmerich Kelih. "Anton Seměnovič Budilovič (1846-1908) – A Forerunner of Quantitative Linguistics in Russia?" *Glottometrics*, vol. 7, pp. 94–96.
- d Grzybek, Peter, Emmerich Kelih and Gabriel Altmann. "Graphemhäufigkeiten (Am Beispiel des Russischen). Teil II: Modelle der Häufigkeitsverteilung." *Anzeiger für Slavische Philologie*, vol. 32, pp. 25–54.
- e Kelih, Emmerich and Peter Grzybek. "Häufigkeiten von Satzlängen. Zum Faktor der Intervallgröße als Einflussvariable (am Beispiel slowenischer Texte)." *Glottometrics*, vol. 8, pp. 23–41.
- f Grzybek, Peter. "Worthäufigkeit und Wortlänge in Sprichwörtern (Am Beispiel slowenischer Sprichwörter)." *Phraseologismen als Gegenstand sprach- und kulturwissenschaftlicher Forschung*, edited by Csaba Földes and J. Wիրrer: Baltmannsweiler: Schneider-Verlag Hohengehren, pp. 47–58.
- g Grzybek, Peter. "Zur Wortlänge und ihrer Häufigkeitsverteilung in Sprichwörtern (Am Beispiel slowenischer Sprichwörter, mit einer Re-Analyse estnischer Sprichwörter)." *Europhras 2000. Internationale Tagung zur Phraseologie Vom 15.-18. Juni 2000 in Aske / Schweden*, edited by Christine Palm-Meister. Tübingen: Stauffenburg, pp. 161–71.
- h Chlosta, Christoph and Peter Grzybek. "Was heißt eigentlich 'Bekanntheit' von Sprichwörtern? Methodologische Bemerkungen anhand einer Fallstudie zur Bekanntheit anglo-amerikanischer Sprichwörter in Kanada und in den USA." *Res humanae proverbiorum et sententiarum. Ad honorem Wolfgangi Mieder*, edited by Csaba Földes. Tübingen: Narr, 2004, pp. 37–57.

2005

- a Grzybek, Peter and Emmerich Kelih. "Häufigkeiten von Buchstaben / Graphemen / Phonemen: Konvergenzen des Rangierungsverhaltens." *Glottometrics*, vol. 9, pp. 62–73.
- b Grzybek, Peter, Emmerich Kelih and Gabriel Altmann. "Graphemhäufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfangs – Eine Nebenbemerkung zur Diskussion um das 'ë'." *Anzeiger für Slavische Philologie*, vol. 33, pp. 117–40.
- c Kelih, Emmerich and Peter Grzybek. "Satzlänge: Definitionen, Häufigkeiten, Modelle. (Am Beispiel slowenischer Prosatexte)." *Quantitative Methoden in Computerlinguistik und Sprachtechnologie*. [= *Special Issue of: LDV-Forum. Zeitschrift für Computerlinguistik und Sprachtechnologie / Journal for Computational Linguistics and Language Technology*], vol. 20, no. 2, pp. 31–51.
- d Grzybek, Peter. "A Study on Russian Graphemes." *Jazyk. Ličnost'. Tekst. Sbornik statej k 70-letiju T.M. Nikolaevoj*, edited by Vladimir N. Toporov. Moskva: Jazyki slavjanskich kul'tur, pp. 237–63.
- e Kelih, Emmerich, Gordana Antić, Peter Grzybek and Ernst Stadlober. "Classification of Author and/or Genre? The Impact of Word Length." *Classification: The Ubiquitous Challenge*, edited by Claus Weihs and Wolfgang Gaul. Heidelberg, New York: Springer, pp. 498–505.
- f Kelih, Emmerich and Peter Grzybek. "Neuanfang und Etablierung quantitativer Verfahren in der sowjetischen Sprach- und Literaturwissenschaft (1956-1962)." *Quantitative Linguistics / Quantitative Linguistik. An International Handbook / Ein Internationales Handbuch*, edited by Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski. Berlin: de Gruyter, pp. 65–82. (= Handbücher zur Sprach- und Kommunikationswissenschaft 27)
- g Grzybek, Peter, Ernst Stadlober, Emmerich Kelih and Gordana Antić. "Quantitative Text Typology: The Impact of Word Length." *Classification: The Ubiquitous Challenge*, edited by Claus Weihs and Wolfgang Gaul. Heidelberg, New York: Springer, pp. 53–64.
- h Grzybek, Peter and Emmerich Kelih. "Empirische Textsemiotik und quantitative Text-Typologie." *Text & Reality. Text & Wirklichkeit*, edited by Jeff Bernard, Jurij Fikfak and Peter Grzybek. Ljubljana: ZRC, 2005, pp. 95–120.
- i Grzybek, Peter and Emmerich Kelih. "Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph." *Problemi kvantitativnoi lingvistiki. Problems of Quantitative Linguistics*, edited by Gabriel Altmann, Viktor Levickij and Valentina Perebejnisi. Černivci: Ruta, 2005, pp. 159–79.
- j Grzybek, Peter and Emmerich Kelih. "Textforschung: Empirisch!". *Die Leipziger Text-Tage*, edited by Julia K. Banke et al., Leipzig: FSR, pp. 13–34.
- k Grzybek, Peter and Emmerich Kelih. "Towards a General Model of Grapheme Frequencies in Slavic Languages." *Computer Treatment of Slavic and East European Languages*, edited by Radovan Garabík, Bratislava: Veda, pp. 73–87.
- l Grzybek, Peter and Emmerich Kelih. "Zur Vorgeschichte quantitativer Ansätze in der russischen Sprach- und Literaturwissenschaft." *Quantitative Linguistics / Quantitative Linguistik. An International Handbook / Ein Internationales Handbuch*, edited by Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski. Berlin: de Gruyter, pp. 23–64. (= Handbücher zur Sprach- und Kommunikationswissenschaft 27)
- m Antić, Gordana, Peter Grzybek and Ernst Stadlober. "Mathematical Aspects and Modifications of Fucks' Generalized Poisson Distribution." *Quantitative Linguistics*

/ *Quantitative Linguistik. An International Handbook / Ein Internationales Handbuch*, edited by Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski. Berlin: de Gruyter, pp. 158–80. (= Handbücher zur Sprach- und Kommunikationswissenschaft 27)

2006

- a Grzybek, Peter. “A Very Early Slavic Letter Statistic and the Czech Journal 'Krok' (1841). Jan Svatopluk Presl (1791–1849).” *Glottometrics*, vol. 12, pp. 88–91.
- b Grzybek, Peter. “Tomo Maretić’s First Croatian and/or Serbian Sound Statistics (1899).” *Glottometrics*, vol. 12, pp. 92–96.
- c Grzybek, Peter, Emmerich Kelih and Ernst Stadlober. “Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik.” *Anzeiger für Slavische Philologie*, vol. 34, pp. 41–74.
- d Grzybek, Peter. “History and Methodology of Word Length Studies. The State of the Art.” *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*, edited by Peter Grzybek. Dordrecht, NL: Springer, 2006, pp. 15–90. (= Text, Speech and Language Technology 31)
- e Grzybek, Peter. “Introductory Remarks: On the Science of Language in Light of the Language of Science.” *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*, edited by Peter Grzybek. Dordrecht, NL: Springer, 2006, pp. 1–14. (= Text, Speech and Language Technology 31)
- f Grzybek, Peter. “Semiotik und Phraseologie.” *Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung*, edited by Harald Burger et al. Berlin: de Gruyter, pp. 188–208.
- g Kelih, Emmerich, Peter Grzybek, Gordana Antić and Ernst Stadlober. “Quantitative Text Typology. The Impact of Sentence Length.” *From Data and Information Analysis to Knowledge Engineering*, edited by Myra Spiliopoulou et al. Heidelberg, Berlin: Springer, pp. 382–89.
- h Strauss, Udo, Peter Grzybek and Gabriel Altmann. “Word Length and Word Frequency.” *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*, edited by Peter Grzybek. Dordrecht, NL: Springer, 2006, pp. 277–95. (= Text, Speech and Language Technology 31)
- i Grzybek, Peter, Ernst Stadlober and Emmerich Kelih. “The Relationship of Word Length and Sentence Length. The Inter-Textual Perspective.” *Advances in Data Analysis. Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8-10, 2006*, edited by Reinhold Decker and Hans-Jörg Lenz. Heidelberg, Berlin: Springer, 2007, pp. 611–18.
- j Grzybek, Peter and Emmerich Kelih. “Häufigkeiten von Wortlängen und Wortlängenpaaren. Untersuchungen am Beispiel russischer Texte von Viktor Pelevin.” *Zeit – Ort – Erinnerung. Slawistische Erkundungen aus sprach-, literatur- und kulturwissenschaftlicher Perspektive. Festschrift für Ingeborg Ohnheiser und Christine Engel zum 60. Geburtstag*, edited by Eva Binder et al. Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck, Abt. Sprachwissenschaft, pp. 395–407.
- k Grzybek, Peter, Emmerich Kelih and Gabriel Altmann. “Graphemhäufigkeiten im Slowakischen. Teil II: Mit Digraphen.” *Sprache und Sprachen im mitteleuropäischen Raum*, edited by Ružena Kozmová. Trnava: Univerzita sv. Cyrila a Metoda, Filozofická fakulta, pp. 661–64.

- l Antić, Gordana, Kelih, Emmerich and Peter Grzybek. “Zero-syllable Words in Determining Word Length.” *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*, edited by Peter Grzybek. Dordrecht, NL: Springer, pp. 117–56. (= Text, Speech and Language Technology 31)
- m Antić, Gordana, Ernst Stadlober, Peter Grzybek and Emmerich Kelih. “Word Length and Frequency Distributions in Different Text Genres.” *From Data and Information Analysis to Knowledge Engineering*, edited by Myra Spiliopoulou et al. Heidelberg, Berlin: Springer, pp. 310–17.
- 2007**
- a Grzybek, Peter. “On the Systematic and System-Based Study of Grapheme Frequencies. A Re-Analysis of German Letter Frequencies.” *Glottometrics*, vol. 15, pp. 82–91.
- b Grzybek, Peter. “What a Difference an «E» Makes. Die erleichterte Interpretation von Graphemhäufigkeiten unter erschwerten Bedingungen.” *Kritik und Phrase. Festschrift für Wolfgang Eismann zum 65. Geburtstag*, edited by Peter Deutschmann, Peter Grzybek, Ludwig Karničar and Heinrich Pfandl. Wien: Praesens, p. 205.
- c Grzybek, Peter. “Wolfgang Eismann – Leben in Bewegung.” *Kritik und Phrase. Festschrift für Wolfgang Eismann zum 65. Geburtstag*, edited by Peter Deutschmann, Peter Grzybek, Ludwig Karničar and Heinrich Pfandl. Praesens, pp. 11–21.
- d Grzybek, Peter and Ernst Stadlober. “Do we Have Problems with Arens' Law? A New Look at the Sentence-Word Relation.” *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*, edited by Peter Grzybek and Reinhard Köhler. Berlin: Mouton de Gruyter, pp. 205–17. (= Quantitative Linguistics 62)
- 2008**
- a Grzybek, Peter. “Fundamentals of Slovenian Paremiology.” *Traditiones*, vol. 37, no. 1, pp. 23–46.
- b Grzybek, Peter. “Slawistik in Graz / Slavistika v Gracu.” *Signal*, 2007/2008, pp. 183–2007.
- c Grzybek, Peter, Emmerich Kelih and Ernst Stadlober. “The Relation between Word Length and Sentence Length. An Intra-Systemic Perspective in the Core Data Structure.” *Glottometrics*, vol. 16, pp. 111–21.
- 2009**
- a Grzybek, Peter. “Some Essentials on the Popularity of (American) Proverbs.” *The Proverbial 'Pied Piper'. A Festschrift Volume of Essays in Honor of Wolfgang Mieder on the Occasion of his 65th Birthday*, edited by Kevin J. McKenna. Frankfurt/Main: Lang, pp. 95–110.
- b Grzybek, Peter. “The Popularity of Proverbs. A Case Study of the Frequency-Familiarity Relation for German.” *Proceedings of the Second Interdisciplinary Colloquium on Proverbs*, edited by Rui J. B. Soares and Outi Lauhakangas. Tavira: IAP, pp. 214–29.
- c Grzybek, Peter, Emmerich Kelih and Ernst Stadlober. “Slavic Letter Frequencies: A Common Discrete Model and Regular Parameter Behavior?” *Issues in Quantitative Linguistics*, edited by Reinhard Köhler. Lüdenschied: RAM, pp. 17–33. (= Studies in Quantitative Linguistics 5)

2010

- a Grzybek, Peter. "Poslovica i ee situacii." *Živaja starina*, no. 4, pp. 51–54.
- b Fenxiang, Fan, Peter Grzybek and Gabriel Altmann. "Dynamics of Word Length in Sentence." *Glottometrics*, no. 20, pp. 70–109.
- c Grzybek, Peter. "Die Grenze(:) zwischen Märchen und Schwank." *Märchen in den südslawischen Literaturen*, edited by Vladimir Biti and Bernarda Katušić. Frankfurt: Peter Lang, pp. 111–41.
- d Grzybek, Peter. "Text Difficulty and the Arens-Altman Law." *Text and Language. Structures · Functions · Interrelations. Quantitative Perspectives*, edited by Peter Grzybek, Emmerich Kelih and Ján Mačutek. Wien: Praesens, pp. 57–70.
- e Grzybek, Peter and Christoph Chlost. "Überlegungen zur empirischen Validierung von Sprichwörter-Dummies." *Sprachlehrforschung: Theorie und Empirie. Festschrift für Rüdiger Grotjahn*, edited by Annette Berndt and Karin Kleppin. Frankfurt/Main: Lang, pp. 197–209.

2011

- a Grzybek, Peter. "Der Satz und seine Beziehungen. I: Satzlänge und Wortlänge im Russischen (Am Beispiel von L.N. Tolstojs «Анна Каренина»)." *Anzeiger für Slavische Philologie*, 39, pp. 39–74.

2012

- a Grzybek, Peter. "Harry Dexter Kitson (1886–1959)." *Glottometrics*, no. 24, 2012, pp. 88–94.
- b Grzybek, Peter. "Michail Lopatto: Attempt at an Introduction into the Theory of Prose (1918)." *Glottometrics*, no. 25, pp. 70–80.
- c Grzybek, Peter and Emmerich Kelih. "[Rev.] Popescu, Ioan-Iovitz; Čech, Radek; Altmann, Gabriel (2011): The Lambda-Structure of Texts. Lüdenscheid: Ram-Verlag (= Studies in Quantitative Linguistics 10). 181 pp." *Literary and Linguistic Computing*, vol. 27, no. 1, pp. 104–06.
- d Grzybek, Peter. "Facetten des parömiologischen Rubik-Würfels. Kenntnis ≡ Bekanntheit [↔ Verwendung ≈ Frequenz] !?" *Sprichwörter Multilingual. Theoretische, empirische und angewandte Aspekte der modernen Parömiologie*, edited by Kathrin Steyer. Tübingen: Narr, pp. 99–138.
- e Grzybek, Peter. "Proverb Variants and Variations: A New Old Problem?" *Proceedings of the Fifth Interdisciplinary Colloquium on Proverbs*, edited by Outi Lauhakangas and Rui J. B. Soares. Tavira: IAP, pp. 136–52.
- f Grzybek, Peter. "The Köhler Knick: Prolegomena on the Synergetics of Academic and Bio-Demographic Politics." *Synergetic Linguistics. Text and Language as Dynamic Systems*, edited by Sven Naumann and Petr Grzybek. Wien: Praesens, pp. 47–66.
- g Grzybek, Peter. "Close and Distant Relatives of the Sentence: Some Results from Russian." *Methods and Applications of Quantitative Linguistics*, edited by Ivan Obradović, Emmerich Kelih and Reinhard Köhler. Belgrade: Academic Mind, 2013, pp. 44–58.
- h Grzybek, Peter and Veronika Koch. "Shot Length: Random or Rigid, Choice or Chance? An Analysis of Lev Kulešov's *Po Zakonu* [By the Law]." *Sign Culture. Zeichen Kultur*, edited by Ernest W.B. Hess-Lüttich. Würzburg: Königshausen & Neumann, 2012, pp. 169–88.

2013

- a Grzybek, Peter. "Historical Remarks on the Consonant-Vowel Proportion – From Cryptoanalysis to Linguistic Typology. The Concept of Phonological Stoichiometry (Francis Lieber, 1800-1872)." *Glottometrics*, no. 26, pp. 96–103.
- b Popescu, Ioan-Iovitz, Peter Grzybek, Sven Naumann and Gabriel Altmann. "Some Statistics for Sequential Text Properties." *Glottometrics*, no. 26, pp. 50–95.
- c Grzybek, Peter. "Empirische Textwissenschaft. Prosarhythmus im ersten Drittel des 20. Jahrhunderts als historisch-systematische Fallstudie." *Form und Wirkung. Phänomenologische und empirische Kunstwissenschaft in der Sowjetunion der 1920er Jahre*, edited by Aage Hansen-Löve et al. München: Fink, pp. 427–55.
- d Grzybek, Peter. "Homogeneity and Heterogeneity within Language(s) and Text(s): Theory and Practice of Word Length Modeling." *Issues in Quantitative Linguistics 3. Dedicated to Karl-Heinz Best on the Occasion of his 70th Birthday*, edited by Reinhard Köhler and Gabriel Altmann. Lüdenscheid: RAM, pp. 66–99.
- e Grzybek, Peter. "Samoreguljacija v tekste (na primere ritmičeskich processov v proze)." *Slučajnost' i nepredskazuemost' v istorii kultury*, edited by Igor' A. Pilščikov. Tallinn, pp. 78–115.
- f Duraš, Gordana, Ernst Stadlober, Emmerich Kelih and Peter Grzybek. "Komplexität sprachlicher Formen. Die Singh-Poisson-Verteilung: ein Modell in der Wortlängenforschung?" *Issues in Quantitative Linguistics 3. Dedicated to Karl-Heinz Best on the Occasion of his 70th Birthday*, edited by Reinhard Köhler and Gabriel Altmann. Lüdenscheid: RAM, pp. 291–308.

2014

- a Grzybek, Peter. "Mosaic or Jigsaw? Publishing an Article from Estonia in the 'West', 30 Years Ago, when Circumstances were Quite Different from Today." *Proverbium. Yearbook of International Proverb Scholarship*, vol. 31, pp. 11–34.
- b Grzybek, Peter. "G.L. Permjakov (1919–1983). A Biographical Mini-Sketch." 8 *Colóquio Interdisciplinar Sobre Provérbios*, IAP, pp. 107–09.
- c Grzybek, Peter. "Regularities of Estonian Proverb Word Length: Frequencies, Sequencies, Dependendies." *Scala Naturae. Festschrift in Honour of Arvo Krikmann*, edited by Anneli Baran et al. Tartu: ELM Scholarly Press, pp. 121–48.
- d Grzybek, Peter. "Semiotic and Semantic Aspects of the Proverb." *Introduction to Paremiology: A Comprehensive Guide to Proverb Studies*, edited by Hrisztalina Hrisztova-Gotthardt and Melita Aleksa Varga. Berlin: de Gruyter, pp. 68–111.
- e Grzybek, Peter. "Simple Form." *Encyclopedia of Humor Studies*, edited by Salvatore Attardo. Los Angeles: Sage, pp. 693–95.
- f Grzybek, Peter. "The Emergence of Stylometry: Prolegomena to the History of Term and Concept." *Text within Text – Culture within Culture*, edited by Katalin Kroó and Peeter Torop. Budapest/Tartu: L'Harmattan, 2014, pp. 58–75.
- g Grzybek, Peter. "Word Length." *The Oxford Handbook of the Word*, edited by John R. Taylor. Oxford: Oxford University Press, 2014, pp. 1–25.
- h Grzybek, Peter and Darinka Verdonik. "General Extenders: From Interaction to Model." *Phraseologie im Wörterbuch und Korpus / Phraseology in Dictionaries and Corpora*, edited by Vida Jesenšek and Peter Grzybek. Maribor: Mednarodna založba oddelka za slovanske jezike in književnosti, Filozofska fakulteta, pp. 113–30.
- i Grzybek, Peter and Vida Jesenšek. "Phraseologie im Wörterbuch und Korpus. Einführende Bemerkungen." *Phraseologie im Wörterbuch und Korpus / Phraseology in Dictionaries and Corpora*, edited by Vida Jesenšek and Peter

Grzybek. Maribor: Mednarodna založba oddelka za slovanske jezike in književnosti, Filozofska fakulteta, pp. 9–17.

- j Grzybek, Peter and Vida Jesenšek. “Phraseology in Dictionaries and Corpora. Introductory Remarks.” *Phraseologie im Wörterbuch und Korpus / Phraseology in Dictionaries and Corpora*, edited by Vida Jesenšek and Peter Grzybek. Maribor: Mednarodna založba oddelka za slovanske jezike in književnosti, Filozofska fakulteta, pp. 19–26.

2015

- a Grzybek, Peter. “«Opyt vvedenija v teoriju prozy». Zabytoe nasledie M.O. Lopatto s točki zrenija kvantitativnoj lingvistiki.” *Antropologija kul'tury, vyp. 5: K 85-letija akademika RAN Vjač. Vs. Ivanova*. Moskva: MGU, pp. 367–83.
- b Chlosta, Christoph and Peter Grzybek. “Zum Teufel mit dem ...: Anfang und Ende in der experimentellen Parömiologie.” *Bis dat, qui cito dat. Gegengabe in Paremiology, Folklore, Language, and Literature. Honoring Wolfgang Mieder on His Seventieth Birthday*, edited by Christian Grandl and Kevin J. McKenna. Frankfurt/Main: Lang, pp. 109–20.

2016

- a Grzybek, Peter. “Verse Diversification: Frequencies and Variations of Verse Types in Vana Kannel and Kalevipoeg.” *Studia metrica i poetica*, vol. 3, no. 2, pp. 50–98.
- b Grzybek, Peter. “Word Length in Estonian Prose.” *Trames. A Journal of the Humanities and Social Sciences*, vol. 20, no. 2, pp. 145–75.
- c Grzybek, Peter. “Několik poznámek k pojetí znaku v Jakobsonově sémiotice a v českém strukturalismu.” *Český strukturalismus v diskusi*, edited by Ondřej Sládek. Brno: Host, pp. 247–60.
- d Grzybek, Peter. “On Whether Weather Proverbs Are Weather Proverbs. Towards a Fresh Look at Weather Lore and Meteo-Prognostic Paroemias.” *9th Interdisciplinary Colloquium on Proverbs, ACTAS ICP15 Proceedings*, edited by Rui Soares and Outi Lauhakangas. Tavira: AIP-IAP, pp. 273–90.