

Glottometrics 41 2018

RAM-Verlag

ISSN 2625-8226

Glottometrics

Indexed in ESCI by Thomson Reuters and SCOPUS by Elsevier

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.

Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.

Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druckversion** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).

Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.

Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies**.

Herausgeber – Editors

G. Altmann	Univ. Bochum (Germany)	ram-verlag@t-online.de
K.-H. Best	Univ. Göttingen (Germany)	kbest@gwdg.de
R. Čech	Univ. Ostrava (Czech Republic)	cechradek@gmail.com
F. Fan	Univ. Dalian (China)	Fanfengxiang@yahoo.com
E. Kelih	Univ. Vienna (Austria)	emmerich.kelih@univie.ac.at
R. Köhler	Univ. Trier (Germany)	koehler@uni-trier.de
H. Liu	Univ. Zhejiang (China)	lhtzju@gmail.com
J. Mačutek	Univ. Bratislava (Slovakia)	jmacutek@yahoo.com
A. Mehler	Univ. Frankfurt (Germany)	amehler@em.uni-frankfurt.de
M. Místecký	Univ. Ostrava (Czech Republic)	MMistecky@seznam.cz
G. Wimmer	Univ. Bratislava (Slovakia)	wimmer@mat.savba.sk
P. Zörnig	Univ. Brasilia (Brasilia)	peter@unb.br

External academic peers for Glottometrics

Prof. Dr. Haruko Sanada

Rissho University, Tokyo, Japan (<http://www.ris.ac.jp/en/>);

Link to Prof. Dr. Sanada: [http://researchmap.jp/read0128740/?lang=english](http://researchmap.jp/read0128740/?lang=english;);

<mailto:hsanada@ris.ac.jp>

Prof. Dr. Thorsten Roelcke

TU Berlin, Berlin, Germany (<http://www.tu-berlin.de/>)

Link to Prof. Dr. Roelcke: [http://www.daf.tu-](http://www.daf.tu-berlin.de/menue/deutsch_als_fremd_und_fachsprache/mitarbeiter/professoren_und_pds/prof_dr_thorsten_roelcke)

[berlin.de/menue/deutsch_als_fremd_und_fachsprache/mitarbeiter/professoren_und_pds/prof_dr_thorsten_roelcke](http://www.daf.tu-berlin.de/menue/deutsch_als_fremd_und_fachsprache/mitarbeiter/professoren_und_pds/prof_dr_thorsten_roelcke)

<mailto:Thosten.Roelcke@tu-berlin.de>

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an

Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de

Herunterladen/ Downloading: <https://www.ram-verlag.eu/journals-e-journals/glottometrics/>

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 41 (2018), Lüdenscheid: RAM-Verlag, 2018. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access). unter der Adresse
<https://www.ram-verlag.eu/journals-e-journals/glottometrics/> verfügbar
Bibliographische Deskription nach 41 (2018)

ISSN 2625-8226

Contents

Michal Místecký

- Counting Stylometric Properties of Sonnets: A Case Study of Machar's *Letní sonety* 1 - 12

Sergey Andreev

- Distribution of Syllables in Russian Sonnets 13 - 23

Jieqiang Zhu, Haitao Liu

- The Distribution of Synonymous Variants in Wenzhounese 24 - 39

Haruko Sanada, Gabriel Altmann

- Word Length and Polysemy in Japanese 40 - 45

Michal Místecký

- Belza Chains in Machar's *Letní sonety* 46 - 56

Andrij Rovenchak, Olha Rovenchak

- Quantifying Comprehensibility of Christmas and Easter Addresses from the Ukrainian Greek Catholic Church Hierarchs 57 - 66

Gabriel Altmann

- Some Properties of Adjectives in Texts 67 - 79

History

- Antoni Hernández-Fernández, Ramon Ferrer-i-Cancho** 80 - 86
José María de Oleza Arredondo, S.J. (1887-1975)

Book Reviews

Haitao Liu, Junying Liang (eds.) (2017), *Motifs in Language and Text*. 87 - 90
Berlin/ Boston: De Gruyter Mouton, pp. 271. (*Quantitative Linguistics*
Vol. 71). Reviewed by **Hanna Gnatchuk**

Mikhail Kopotev, Olga Lyashevskaya, & Arto Mustajoki (Eds.) 91 - 95
(2017). *Quantitative Approaches to the Russian Language*. New York:
Routledge. ISBN:978-1-138-09715-5, 220 pp. Reviewed by **Heng Chen**

Counting Stylometric Properties of Sonnets: A Case Study of Machar's *Letní sonety*

Michal Místecký¹

Abstract. The genre of sonnet has been a traditional must-write of all lyric poets since Renaissance; it has shaped itself into many a form, usually following the possibilities of individual national languages. Here, a sample of the Czech sonnet production – *Letní sonety* (1890–91) by Josef Svatopluk Machar, a poet of the 1890s generation – will be analysed. First, the distribution of parts-of-speech in the rhyme words will be investigated; next, various stylistic indicators (activity counts, repeat rate, entropy, h-point, thematic concentration, and the curve-length index of vocabulary richness) will be calculated. The workings of the indicators will be explained in suitable places.

Keywords: Parts of speech, Czech, sonnet, Busemann's coefficient, h-point, vocabulary richness, repeat rate, entropy.

1. Rhyme-Word POS Distribution

A longer study concerning the membership of rhyme words in the classes of POS has already been published (cf. Lupea, Rukk, Popescu, Altmann 2017); here, the research will consider merely sonnets. Sonnet has a prescribed form, and any comparison may be made without any transformations. The parts-of-speech used are those that were proposed already in the Antiquity and hold true still today – in the majority of languages. They are: N = nouns, V = verbs, A = adjectives, Av = adverbs, Pn = pronouns, Nu = numerals, I = interjections, C = conjunctions. In Czech, there are several bordering cases that need to be accounted for: there are verbal adjectives, which are here considered as adjectives, and verbal nouns, which are taken as nouns. It makes sense, according to both the morphological and syntactical properties of the two. Besides, we introduced the class *Ab*, meaning abbreviations.

In Table 1.1, the sequences of rhyme word-POS in Machar's *Letní sonety* are presented. If one ranks them according to decreasing frequency, one obtains a sequence which can be satisfactorily captured by the Zipf-Alekseev function defined as

$$(1.1) \quad y = 1 + cx^{a+b \cdot \ln(x)};$$

it can be derived from the unified theory of language laws (cf. Wimmer, Altmann 2005) by means of a differential equation in which both the requirements of the writer, the reader and the necessity of equilibrium are taken into consideration. It may be remarked here that the formula can be simplified, e.g. by considering $b = 0$ yielding the usual exponential function. If the exponential function is the background of this phenomenon – it can be perhaps shown

¹ Univ. Ostrava (Czech Republic); mail: MMistecky@seznam.cz

analyzing other languages –, then the parameter *b* expresses a boundary condition whose nature must be studied separately.

Table 1.1
Frequency of POS in rhyme words in *Letní sonety* by Machar

Sonnet	POS of rhyme words
E. Zolovi	Pn, Av, A, N, A, Av, A, Av, Av, V, N, N, N, N
Matce	N, A, V, A, N, A, V, Av, V, A, A, V, Pn, V
Sonnet cynický	N, V, Av, Av, Av, N, Av, Av, V, A, V, A, N, N
Sonnet de vanitate	V, A, V, A, V, A, V, A, V, N, N, N, N, N
Sonnet elegický	V, V, N, V, N, V, N, V, A, Av, Av, N, N, N
Sonnet ironický	N, N, N, V, Pn, N, V, N, V, V, V, V, A
Sonnet k sociální otázce	N, V, N, V, N, V, N, V, N, A, V, V, Pn, Nu
Sonnet k teorii: Boj o život	A, N, A, A, N, Av, A, A, N, N, V, V, V, V
Sonnet materialistický	N, N, N, V, N, V, N, N, N, A, Pn, N, A, N
Sonnet mystický	N, V, N, N, N, N, V, Nu, V, Av, V, N, V, Av
Sonnet na Chopinovu melodii	A, N, A, N, V, N, V, V, N, N, Av, N, V, Pn
Sonnet na sentenci z Goetha	N, Av, N, N, N, N, V, N, V, N, V, V, V, V
Sonnet na sklonku století	N, N, N, N, A, N, A, N, N, N, Av, A, N, N
Sonnet nad verši z mládí	N, V, V, V, V, N, V, N, N, N, N, N, N, Av
Sonnet noční	N, V, N, N, V, N, N, V, N, N, N, Av, N, Av
Sonnet o antice a vlasech	N, V, N, A, A, A, N, V, A, N, N, N, N, Pn
Sonnet o bídě	N, V, V, V, N, V, N, N, A, A, V, N, A, A
Sonnet o hodinách	N, V, A, V, A, V, N, V, N, N, N, N, Pn, Av
Sonnet o lásce	V, A, V, A, N, V, V, A, A, A, N, V, V, A
Sonnet o minulosti	N, Av, A, Av, Av, Av, Av, N, Av, V, Av, N, Av, V
Sonnet o Panně Marii	N, Av, V, Nu, N, N, N, N, A, N, N, Pn, N, N
Sonnet o rokoku	Av, N, N, N, N, N, Pn, N, V, N, N, N, N, N
Sonnet o staré metafoře	N, Av, Av, N, Av, Av, Av, N, N, N, N, N, N, V
Sonnet o starém líci a rubu	A, N, N, Pn, N, A, N, N, Ab, N, V, Av, Av, V
Sonnet o třech metaforách	V, N, N, A, N, N, Av, A, N, A, A, Av, N, A
Sonnet o třetí hodině v červenci	N, V, N, N, N, N, V, N, A, A, A, A, N, V
Sonnet o vídeňských kosech	A, N, V, V, V, N, V, V, N, A, A, V, V, A
Sonnet o západu slunce	V, N, V, N, V, N, V, N, N, N, V, V, N, V
Sonnet o zlatém věku naší poezie	Av, N, N, A, Av, N, V, Av, A, N, V, N, A, N
Sonnet o životě	N, N, N, V, V, V, Av, V, Av, V, A, A, V, V
Sonnet patologický	N, N, N, N, N, C, N, N, A, V, N, Av, Av, N
Sonnet polední	N, N, A, N, V, N, N, N, N, N, N, V, N, N
Sonnet sarkastický	N, V, N, N, N, N, N, V, Av, Av, N, N, V, Av
Sonnet svatební	Av, V, A, Pn, N, Pn, Av, A, V, Av, V, N, N, A
Sonnet úvodní	A, Av, N, A, A, N, A, N, N, Av, V, V, N, A
Sonnet večerní	A, V, A, Nu, N, N, N, N, Av, N, A, N, A, A
Sonnet z dvacátého září	V, N, V, V, A, V, V, V, N, A, A, V, V, N
Sonnet-apostrofa	V, N, V, N, V, N, V, A, N, N, N, N, Pn, N
Sonnet-epilog čtenáři	V, V, A, Av, N, V, Av, V, N, V, V, N, V, N

Sonet-intermezzo	V, A, N, V, A, V, N, V, Av, V, N, A, Pn, V
Sonet-intermezzo	Pn, V, A, A, V, V, N, V, A, A, V, Av, A, V
Sonety-causerie I.	V, A, V, V, V, A, V, A, N, Pn, V, Pn, N, Pn
Sonety-causerie II.	V, Pn, N, N, Pn, V, N, N, N, Pn, V, A, Pn, V
Sonety-causerie III.	C, N, A, N, N, A, N, V, V, V, N, V, N, V
Sonety-causerie IV.	V, V, V, N, A, V, V, V, N, N, V, V, Pn, A
Sonety-causerie V.	N, V, N, V, V, N, N, N, N, Pn, N, N, V, N
Své ženě s předešlým sonetem	A, V, N, N, V, A, N, V, N, N, N, Av, V, Av
Frequencies	272 186 105 63 25 4 2 1 1
Classes	N V A Av Pn Nu I C Ab
Zipf-Alekseev ft. +1	271.06 190.97 99.92 51.97 28.18 16.10 9.72 6.22 4.22 a = 0.1878, b = -1.0031, c = 270.0607, R ² = 0.9944

The Zipf-Alekseev function shows an outstanding fit, confirming thus the tendency of POS to follow the given distribution/function. The numbers presented here may be of use when sonnet collections of various authors are to be compared; this research can demonstrate whether a subconscious law is observed, or whether there are differences as to authors or national literatures.

2. Activity Counts

Activity of a text is usually calculated via Busemann's coefficient, which is expressed as the number of verbs divided by the sum of adjectives and verbs. In the present research, the decision as to the POS of a word is a matter of the QUITA software, which is used for text processing; i.e., static verbs (such as "be", "have", and the modals) are not included in the computation. Busemann's coefficient is defined as

$$(2.1) \quad B = \frac{V}{A + V}.$$

In the first sonnet (*Sonet úvodní* – "The Introductory Sonnet") by Machar, we obtain the following sequence:

A-A-A-A-V-V-A-A-A-A-A-A-A-V-A-A-A-V-V-V-V-A-A-A,

in which there are 18 adjectives and 8 verbs. One thus acquires

$$B = \frac{8}{26} = 0.3.$$

In the given case, the descriptiveness is greater than the activity; however, a question needs to be asked whether the difference is significant. To this end, several statistical tests may be performed, the simplest of which is the chi-square test. This was defined (Zörnig 2015) as

$$(2.2) \quad \chi^2 = \frac{(V - A)^2}{V + A},$$

with 1 degree of freedom. The respective probabilities can be found in the usual chi-square tables.

For example, the first sonnet by Machar yields

$$\chi^2 = \frac{(8 - 18)^2}{8 + 18} = 3.846,$$

with 1 degree of freedom, which means a slightly significant result. The text may therefore be declared significantly descriptive (SD). The interpretation of the results distinguishes the following categories:

SA = significantly active ($V > A$, $X^2 > 3.84$);

AC = active ($V > A$, $X^2 < 3.84$);

N = neutral ($B \approx 0.5$);

DE = descriptive ($V < A$, $X^2 < 3.84$);

SD = significantly descriptive ($V < A$, $X^2 > 3.84$).

The summary of the results is to be found in Table 2.1.

Table 2.1
Busemann's coefficient in individual sonnets

Sonnet	V	A	B	Type
E. Zolovi	7	8	0.47	DE
Matce	12	8	0.60	AC
Sonet cynický	12	8	0.60	AC
Sonet de vanitate	8	13	0.38	DE
Sonet elegický	9	9	0.50	N
Sonet ironický	8	11	0.42	DE
Sonet k sociální otázce	12	6	0.67	AC
Sonet k teorii: Boj o život	13	8	0.62	AC
Sonet materialistický	8	12	0.40	DE
Sonet mystický	11	10	0.52	AC
Sonet na Chopinovu melodii	11	6	0.65	AC
Sonet na sentenci z Goetha	9	6	0.60	AC
Sonet na sklonku století	7	8	0.47	DE
Sonet nad verši z mládí	8	7	0.53	AC
Sonet noční	11	7	0.61	AC
Sonet o antice a vlasech	7	16	0.30	DE
Sonet o bídě	11	5	0.69	AC
Sonet o hodinách	13	4	0.76	SA
Sonet o lásce	12	8	0.60	AC
Sonet o minulosti	10	2	0.83	SA
Sonet o Panně Marii	7	13	0.35	DE
Sonet o rokoku	10	10	0.50	N
Sonet o staré metafoře	10	9	0.53	AC
Sonet o starém líci a rubu	12	7	0.63	AC
Sonet o třech metaforách	6	13	0.32	DE
Sonet o třetí hodině v červenci	5	10	0.33	DE
Sonet o vídeňských kosech	14	6	0.70	AC

Sonet o západu slunce	15	7	0.68	AC
Sonet o zlatém věku naší poezie	8	4	0.67	AC
Sonet o životě	19	7	0.73	SA
Sonet patologický	15	5	0.75	SA
Sonet polední	10	11	0.48	DE
Sonet sarkastický	9	7	0.56	AC
Sonet svatební	16	5	0.76	SA
Sonet úvodní	8	18	0.31	SD
Sonet večerní	6	10	0.38	DE
Sonet z dvacátého září	10	10	0.50	N
Sonet-apostrofa	12	6	0.67	AC
Sonet-epilog čtenáři	12	9	0.57	AC
Sonet-intermezzo ₂	8	9	0.47	DE
Sonet-intermezzo	8	10	0.44	DE
Sonety-Causerie I.	14	5	0.74	SA
Sonety-Causerie II.	14	5	0.74	SA
Sonety-Causerie III.	5	9	0.36	DE
Sonety-Causerie IV.	9	5	0.64	AC
Sonety-Causerie V.	11	8	0.58	AC
Své ženě s předešlým sonetem	11	9	0.55	AC

If one computes the individual test results, one obtains their ranking as presented in Table 2.2.

Table 2.2
Ranking of tested Busemann's coefficient from Table 2.1

Rank	Type	Number	Exp. + 1
1	AC	22	22.43
2	DE	13	11.95
3	SA	7	6.60
4	N	3	3.86
5	SD	1	2.46
a = 41.9121, b = 0.6721, R ² = 0.9850			

Though in individual cases, the whole of *Letní sonety* prefers various structuring of sonnets, the general trend resulting from testing can be captured by the exponential function defined as

$$(2.3) \quad y = 1 + a^{-bx} .$$

3. Repeat Rate and Entropy

The two present indicators show a degree of vocabulary richness. The counts work with types and tokens, the characteristics of which are those that are predefined in the QUITA software,

the main processor of the texts. As to repeat rate (RR), Čech et al. (2014) defines it as a mark of lexical concentration of a text: the higher the RR value is, the more concentrated a text lexically is, which means the fewer types it contains. The RR formula reads

$$(3.1) \quad RR = \frac{1}{N^2} \sum_{r=1}^V f_r^2,$$

where N stands for the number of tokens, V for types, and f_r for a frequency of a particular token r .

On the other hand, the notion of entropy is broader – it has been introduced to cybernetics by Shannon (1948), and made use for linguistics in the mid of the 20th century. Generally, it is defined as the measure of unpredictability of a system – within the sphere of lexical statistics, it means that the higher an entropy of a text is, the richer it may be considered. The entropy is calculated as follows:

$$(3.2) \quad H = \log_2 N - \frac{1}{N} * \sum_{r=1}^N f_r * \log_2 f_r,$$

the meaning of the abbreviations being the same as in RR.

Let an example be presented. The first sonnet by Machar includes 84 tokens and 69 types; if the frequency data are provided, the formula gives the result

$$RR = \frac{1}{84^2} \sum_{r=1}^{69} (5^2 + 3^2 + 3^2 + 3^2 + \dots) = 0.0187.$$

The same data will be employed in the entropy count; i.e. –

$$H = \log_2 84 - \frac{1}{84} * \sum_{r=1}^{84} (5 + 3 + 3 + 3 + \dots) * \log_2(5 + 3 + 3 + 3 + \dots) = 5.9652.$$

The results of RR and H for Machar's *Letní sonety* are presented in Table 3.1. The results may be employed in comparisons of sonnet books from different languages on the basis of statistical tests.

Table 3.1
Repeat rate and entropy in Machar's *Letní sonety*

Sonnet	Tokens	Types	RR	H
E. Zolovi	78	66	0.0187	5.9230
Matce	87	72	0.0168	6.0578
Sonnet cynický	97	69	0.0275	5.7489
Sonnet de vanitate	85	65	0.0206	5.8447
Sonnet elegický	90	69	0.0212	5.8903
Sonnet ironický	84	69	0.0179	5.9844
Sonnet k sociální otázce	94	77	0.0168	6.1210

Sonet k teorii: Boj o život	97	71	0.0216	5.9140
Sonet materialistický	92	73	0.0189	6.0103
Sonet mystický	76	67	0.0166	6.0012
Sonet na Chopinovu melodii	108	67	0.0233	5.7827
Sonet na sentenci z Goetha	86	70	0.0176	6.0015
Sonet na sklonku století	89	71	0.0193	5.9784
Sonet nad verši z mládí	94	72	0.0197	5.9810
Sonet noční	72	60	0.0193	5.8156
Sonet o antice a vlasech	91	67	0.0190	5.9115
Sonet o bídě	97	79	0.0165	6.1514
Sonet o hodinách	88	68	0.0201	5.9080
Sonet o lásce	90	68	0.0247	5.8248
Sonet o minulosti	74	50	0.0285	5.4185
Sonet o Panně Marii	81	64	0.0184	5.8921
Sonet o rokoku	91	73	0.0199	5.9806
Sonet o staré metafoře	81	55	0.0261	5.5468
Sonet o starém líci a rubu	102	73	0.0254	5.8510
Sonet o třech metaforách	77	63	0.0204	5.8316
Sonet o třetí hodině v červenci	91	69	0.0228	5.8675
Sonet o vídeňských kosech	93	74	0.0195	6.0081
Sonet o západu slunce	103	86	0.0171	6.2290
Sonet o zlatém věku naší poezie	70	59	0.0204	5.7756
Sonet o životě	89	67	0.0203	5.8855
Sonet patologický	89	67	0.0211	5.8704
Sonet polední	111	89	0.0166	6.2682
Sonet sarkastický	89	67	0.0203	5.8789
Sonet svatební	92	75	0.0168	6.0984
Sonet úvodní	84	69	0.0187	5.9652
Sonet večerní	78	71	0.0155	6.0962
Sonet z dvacátého září	70	60	0.0208	5.7826
Sonet-apostrofa	78	59	0.0227	5.7179
Sonet-epilog čtenáři	94	75	0.0213	5.9934
Sonet-intermezzo ₂	90	67	0.0227	5.8277
Sonet-intermezzo	71	54	0.0264	5.5424
Sonety-Causerie I.	98	72	0.0194	5.9629
Sonety-Causerie II.	94	69	0.0231	5.8453
Sonety-Causerie III.	86	70	0.0206	5.9394
Sonety-Causerie IV.	90	62	0.0279	5.6298
Sonety-Causerie V.	91	68	0.0245	5.8153
Své ženě s předešlým sonetem	89	70	0.0183	5.9828

It is to be noted that RR and H can be both mutually transformable and also expressed in many other forms known from statistics (cf. Altmann, Lehfeldt 1980: 181).

4. The h-point

The h-point is usually considered a border between autosemantic and synsemantic words (Čech et al. 2014); it is mathematically defined as the position where the rank of a given word matches its frequency. If this is not to be found, the h-point is counted as an “average” of the bordering positions:

$$(4.1) \quad h = \frac{f(i) * r_j - f(j) * r_i}{r_i - r_j + f(i) - f(j)}.$$

In the formula, r_i is the highest rank for which $r_i < f(i)$ is valid, whereas r_j is the lowest rank for which $r_j < f(j)$ is valid.

The calculation of the h-point will be exemplified upon Machar’s *Sonet k sociální otázce* (“A Sonnet on the Social Question”). In Table 4.1, the rank-frequency distribution of types in the poem is demonstrated; it is to be seen that the h-point will fit in between ranks 3 and 4, for which the aforementioned conditions hold well. The count will thus yields

$$h = \frac{4 * 4 - 3 * 3}{4 - 3 + 4 - 3} = 3.5 .$$

Table 4.1

The rank-frequency distribution of types in Machar’s *Sonet o sociální otázce*

Rank	Type	Frequency	26	vždyť	1	52	vlažný	1
1	z	4	27	mít	1	53	všechno	1
2	a	4	28	hrom	1	54	denní	1
3	se	4	29	do	1	55	být	1
4	v	3	30	už	1	56	půlnoc	1
5	jenž	2	31	jeden	1	57	Mila	1
6	chudák	2	32	by	1	58	tma	1
7	jak	2	33	dost	1	59	námaha	1
8	po	2	34	oko	1	60	zřít	1
9	ten	2	35	chudý	1	61	muž	1
10	ráz	2	36	trochu	1	62	krok	1
11	žádný	1	37	i	1	63	klít	1
12	pán	1	38	pánbůh	1	64	pár	1
13	tady	1	39	on	1	65	silueta	1
14	bohatý	1	40	dva	1	66	sbor	1
15	protest	1	41	splést	1	67	činit	1
16	zalétat	1	42	dát	1	68	nesnáz	1
17	k	1	43	ulice	1	69	kašel	1
18	spět	1	44	vyleštěný	1	70	znít	1
19	ted’	1	45	srpek	1	71	hlas	1
20	já	1	46	van	1	72	skvít	1
21	můj	1	47	lít	1	73	dálka	1
22	dřít	1	48	modravý	1	74	dlažba	1
23	trosky	1	49	temno	1	75	prát	1
24	věta	1	50	mosaz	1	76	kdos	1
25	platit	1	51	měsíc	1	77	hůl	1

Table 4.2
The h-points in Machar's *Letní sonety*

	H-point
E. Zolovi	3
Matce	3
Sonet cynický	4
Sonet de vanitate	4
Sonet elegický	3
Sonet ironický	3
Sonet k sociální otázce	3.5
Sonet k teorii: Boj o život	3
Sonet materialistický	3.33
Sonet mystický	2
Sonet na Chopinovu melodii	4
Sonet na sentenci z Goetha	3
Sonet na sklonku století	3
Sonet nad verši z mládí	3
Sonet noční	2.5
Sonet o antice a vlasech	3
Sonet o bídě	3.5
Sonet o hodinách	3
Sonet o lásce	3
Sonet o minulosti	3
Sonet o Panně Marii	3
Sonet o rokoku	4
Sonet o staré metafoře	3.5
Sonet o starém líci a rubu	4
Sonet o třech metaforách	3
Sonet o třetí hodině v červenci	3
Sonet o vídeňských kosech	3
Sonet o západu slunce	4
Sonet o zlatém věku naší poezie	2.5
Sonet o životě	3
Sonet patologický	3
Sonet polední	4
Sonet sarkastický	3.5
Sonet svatební	3
Sonet úvodní	3
Sonet večerní	2
Sonet z dvacátého září	3
Sonet-apostrofa	2.5
Sonet-epilog čtenáři	3.5
Sonet-intermezzo ₂	3.5
Sonet-intermezzo	3.5
Sonety-Causerie I.	4
Sonety-Causerie II.	3.5

Sonety-Causerie III.	3
Sonety-Causerie IV.	4.33
Sonety-Causerie V.	3
Své ženě s předešlým sonetem	3

As can be seen, the values of the h-point lie in the interval $\langle 2, 4.33 \rangle$. Since sonnets have approximately the same length, one can consider this interval as a characteristic property of Czech sonnets; however, a comparison with other languages may show that it is approximately equal in all languages. The concentration of these numbers can, perhaps, be expressed by their average, and the averages may be compared statistically.

5. Curve-length index of vocabulary richness

Another way to measure lexical richness of a text is an index based on a proportion of the length of the whole curve representing the distribution of tokens and its part below the h-point (R Index; see Čech et al. 2014). The idea behind the calculation is that what counts in measuring vocabulary richness are the expressions which are usually to be found in the below-h-point part of the curve ($L - L_h$). The R Index is then formally expressed as

$$(5.1) \quad R = 1 - \frac{L_h}{L},$$

where L stands for the length of the whole curve and L_h for its part above the h-point.²

The count will be illustrated upon an example. The curve length of the first sonnet by Machar is 70.07, out of which 4.87 lies above the h-point; the count thus yields

$$R = 1 - \frac{4.87}{70.07} = 0.93.$$

The results are listed in Table 5.1. They may be made use of in comparisons with other sonnet collections, if the attention is paid to authors' styles, literary schools, or language differences.

Table 5.1

Curve-length, above-h-point curve-length, and R Index values in Machar's *Letní sonety*

Sonnet	L	L _h	R Index
E. Zolovi	66.24	3.41	0.95
Matce	72.24	3.83	0.95
Sonet cynický	72.13	7.30	0.90
Sonet de vanitate	65.65	5.24	0.92
Sonet elegický	70.48	4.65	0.93
Sonet ironický	69.24	3.41	0.95
Sonet k sociální otázce	77.24	4.83	0.94
Sonet k teorii: Boj o život	74.93	7.10	0.91
Sonet materialistický	74.06	4.65	0.94
Sonet mystický	66.83	2.41	0.96
Sonet na Chopinovu melodii	69.82	6.99	0.90
Sonet na sentenci z Goetha	69.83	3.00	0.96

² For details of the calculation of L and L_h , see Čech et al. (2014).

Sonet na sklonku století	72.48	5.06	0.93
Sonet nad verši z mládí	74.95	6.12	0.92
Sonet noční	59.83	2.41	0.96
Sonet o antice a vlasech	67.24	3.41	0.95
Sonet o bídě	79.24	4.41	0.94
Sonet o hodinách	69.99	5.16	0.93
Sonet o lásce	72.34	7.51	0.90
Sonet o minulosti	51.48	4.65	0.91
Sonet o Panně Marii	63.83	3.41	0.95
Sonet o rokoku	73.66	4.83	0.93
Sonet o staré metafoře	55.66	4.83	0.91
Sonet o starém líci a rubu	76.54	8.12	0.89
Sonet o třech metaforách	63.24	3.41	0.95
Sonet o třetí hodině v červenci	72.37	6.54	0.91
Sonet o vídeňských kosech	75.48	4.65	0.94
Sonet o západu slunce	87.48	6.06	0.93
Sonet o zlatém věku naší poezie	59.24	2.83	0.95
Sonet o životě	68.99	5.16	0.93
Sonet patologický	68.48	5.06	0.93
Sonet polední	91.40	6.58	0.93
Sonet sarkastický	67.66	4.83	0.93
Sonet svatební	76.06	4.65	0.94
Sonet úvodní	70.06	4.24	0.94
Sonet večerní	70.83	2.41	0.97
Sonet z dvacátého září	60.24	3.41	0.94
Sonet-apostrofa	60.99	4.58	0.92
Sonet-epilog čtenáři	78.37	7.95	0.90
Sonet-intermezzo ₂	68.48	6.06	0.91
Sonet-intermezzo	54.66	4.83	0.91
Sonety-Causerie I.	72.66	4.83	0.93
Sonety-Causerie II.	70.89	6.06	0.91
Sonety-Causerie III.	71.48	5.06	0.93
Sonety-Causerie IV.	65.23	7.40	0.89
Sonety-Causerie V.	71.23	6.40	0.91
Své ženě s předešlým sonetem	71.06	4.24	0.94

6. Conclusions

The point of presenting the first results of the analysis of Machar's sonnets is to provide useful data for further research, which will incorporate various manifestations of the genre. The goal of this article was thus only informative, as more in-depth investigations will be carried out once more colourful material has been processed.

An analysis of other aspects will follow.

References

- Altmann, G.- Lehfedt, W.** (1980). *Einführung in die Quantitative Phonologie*. Bochum: Brockmeyer.
- Čech, R. – Popescu, I. I. – Altmann, G.** (2014). *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: Univerzita Palackého v Olomouci.
- Lupea, M. – Rukk, M. – Popescu, I.-I. – Altmann, G.** (2017). *Some Properties of Rhyme*. Lüdenscheid: RAM-Verlag.
- Machar, J. S.** *Letní sonety*. Available at:
<<http://www.rodon.cz/admin/files/ModuleKniha/1100-Ctyri-knihy-sonetu.pdf>>.
- Shannon, C. E.** (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal* 27(3) 1948, 379–423.
- 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: de Gruyter.
- Zörnig, P. et al.** (2015). *Descriptiveness, Activity and Nominality in Formalized Text Sequences*. Lüdenscheid: RAM-Verlag.

Distribution of Syllables in Russian Sonnets

*Sergey Andreev*¹

Abstract. Different types of syllables in 46 sonnets, written by prominent Russian poets during the 18th – 21st centuries which cover four important periods of Russian poetry, were counted and ranked according to frequency. The syllabic types were formed on the basis of vowel–consonant sequences of phonemes in syllables and their number in syllables. The count of identical syllables allowed receiving ranked frequencies. To catch the rank distribution exponential plus 1 and Lorentzian functions were used and brought about good fitting results.

Keywords: Russian, syllable, sonnet.

Syllabic patterns in syllabotonic versification in Russian poetry have been studied numerically since the beginning of the 20th century when poet-symbolist A. Belyj examined the variation and gradation of stress on different syllabic positions (Belyj, 1910). In the present-day works in the field of linguistics of verse attention is mostly focused on the correlation of syllabic word length of different POS classes with their position in the line (Gasparov, 2012). These studies revealed quite a number of interesting facts and tendencies but were focused on the patterns of the verse line.

The present study is devoted to the syllabic organization of the whole poetic works and aims at catching the distribution of different types of syllables with an appropriate function. The data-base for the present study includes sonnets – the genre which imposes strict rules both on its formal features (rhyme, number of lines) and on thematic composition (four-part plot with volta at the end). This makes the genre of sonnets especially convenient for such a study since demanding more or less identical forms of poetry it allows to carry out comparisons of different authors and styles.

For the present study 46 sonnets of Russian prominent poets, written during the period of over 2 hundred years, were chosen. They include the works of the late 18th century (the period when the genre started to develop after it appeared in Russian literature in mid 1800s), the first part of the 19th century (the so-called “Golden Age” of Russian poetry), the beginning of the 20th century (“the Silver Age” of Russian poetry) and the end of the 20th – the beginning of the 21st centuries (modern period).

The list of the sonnets is demonstrated in Table 1.

Table 1
Analyzed sonnets

Number	Author	Title	Period
Text 1	V.Trediakovskij	Sonet	I
Text 2	V. Trediakovskij	Sonet iz seja grecheskija rechi	I

¹ Sergej Andreev, Smolensk State University, 214000 Przhevalskij str. 4, Smolensk, Russia. Email: smol.an@mail.ru

Text 3	M. Heraskov	Sonet i jepitafija	I
Text 4	M. Heraskov	«Kol' budu v zhizni ja nakazan nishhetoju...»	I
Text 5	A. Rzhetskij	Sonet, zakljuchajushhij v sebe tri mysli:	I
Text 6	A. Rzhetskij	Sonet, tri raznye sistemy zakljuchajushhij	I
Text 7	I. Dmitriev	Sonet	I
Text 8	V. Zhukovskij	Sonet	II
Text 9	A. Del'vig	N. M. Jazykovu	II
Text 10	A. Del'vig	Vdohnovenie	II
Text 11	A. Del'vig	"Ja plyn odin s prekrasnoju v gondole..."	II
Text 12	E. Baratynskij	"My p'jom v ljubvi otravu sladkiju..."	II
Text 13	E. Baratynskij	"Hotja ty malyj molodoj..."	II
Text 14	N. Jazykov	K.K. Janish	II
Text 15	N. Jazykov	"Na prazdnik vash prines ja dva priveta..."	II
Text 16	A. Pushkin	Sonet	II
Text 17	A. Pushkin	Pojetu	II
Text 18	A. Pushkin	Madona	II
Text 19	V Benediktov	Priroda	II
Text 20	V Benediktov	Kometa	II
Text 21	V Benediktov	Vulkan	II
Text 22	V Benediktov	Groza	II
Text 23	V Benediktov	Cvetok	II
Text 24	V Benediktov	"Krasavica, kak rajskoe viden'e..."	II
Text 25	V Benediktov	"Kogda vdali ot suety vseмирnoj..."	II
Text 26	F. Sologub	Sonet	III
Text 27	V. Brjusov	Sonet	III
Text 28	V. Brjusov	Egipetskij rab	III
Text 29	A. Blok	"Ne ty l' v moih mechtah, pevuchaja, proshla..."	III
Text 30	V. Ivanov	Pritcha o devah	III
Text 31	V. Ivanov	Hramina chuda	III
Text 32	M. Voloshin	Venok sonetov. Sonet 1	III
Text 33	M. Voloshin	Venok sonetov. Sonet 2	III
Text 34	I. Severjanin	Sonet	III
Text 35	A. Belyj	Prosti	III
Text 36	N. Gumilev	Popugaj	III
Text 37	N. Gumilev	Roza	III

Distribution of Syllables in Russian Sonnets

Text 38	S. Esenin	Moej carevne	III
Text 39	K. Bal'mont	Mikel' Andzhelo	III
Text 40	K. Bal'mont	Leonardo da Vinci	III
Text 41	K. Bal'mont	Marlo	III
Text 42	S. Gorodeckij	Mudrost'	IV
Text 43	I. Sel'vinskij	Sonet	IV
Text 44	V. Prokoshin	Deti RA	IV
Text 45	T. Averina	"Ochnjosh'sja – pogruzhjon po grud' v boloto..."	IV
Text 46	N. Beljaeva,	"Schitaju vnov' chasy do nashej vstrechi..."	IV

Syllabic types are singled out according to their phonemic composition, reflecting two aspects. The qualitative-quantitative aspect is vowel-consonant syllabic structure; the second – the number of phonemes in a syllable, irrespective of the type of phonemes. In the first case such syllabic types are singled out: V, CV, CVC, etc. where V is a vowel and C – a consonant. These syllabic types hereinafter are referred to as “phonemic” types. In the second case syllables are classified according to the number of phonemes they contain, and thus they are subdivided into 1-phoneme (V; C), 2-phonemes (CV; VC), etc. and are referred to as “length-types” syllables.

Syllable separation was carried out on the basis of sonorant theory which for the Russian language was worked out by A.A. Reformatskij (1950) and R.I. Avanesov (1954) and fully coincides with our empirical measurement the validity of which was underlined in Köhler, Altmann (2014: 136).

Table 2 contains phonemic types of syllables and their frequency in all 46 sonnets.

Table 2
Phonemic syllable types and their frequency

Syllable	Period I	Period II	Period III	Period IV	Sum
V	64	134	115	27	340
CV	570	1288	1065	321	3244
CCV	137	271	235	62	705
CVC	320	797	638	195	1950
CCVC	72	142	122	49	385
CVCC	15	35	30	5	85
CCCV	15	23	17	6	61
VC	28	110	88	13	239
CCVCC	4	9	7	0	20
CCCVCC	3	1	1	0	5
CCVC	1	8	22	5	36
VCC	0	2	4	0	6
CCCVV	0	1	1	0	2
CVCCCC	1	0	1	4	6

CCCCVC	0	1	1	0	2
VCCC	0	1	0	0	1

The syllables were ranked according to the decreasing frequencies, forming a new sequence of rank frequencies.

To capture the given data of the ranks the exponential function with added 1 was used (Andreev, Popescu, Altmann, 2017, 34-35), defined as:

$$f_x = 1 + a * \exp^{-bx}$$

where a and b are parameters, $x \geq 1$. Parameter b shows the decrease of the function.

The results of fitting are presented in Table 3. Rows show the ranks, and columns the frequencies, observed in each individual sonnet, and the frequencies expected.

Table 3
Fitting of ranks of phonemic syllable types (Exponential + 1 function)

Rank	Nr 1		Nr 2		Nr 3		Nr 4		Nr 5	
	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1
1	94	93.30	69	70.96	85	85.73	97	96.25	89	87.82
2	41	43.80	49	40.99	47	42.77	37	39.68	36	40.35
3	24	20.85	16	23.86	14	21.59	17	16.70	22	18.84
4	8	10.21	12	14.06	14	11.15	10	7.38	9	9.09
5	8	5.27	9	8.47	9	6.00	8	3.59	7	4.67
6	4	2.98	8	5.27	3	3.47	2	2.05	5	2.66
7	2	1.92	6	3.44	2	2.22	1	1.43	3	1.75
8	1	1.43	4	2.39	1	1.60	1	1.17	2	1.34
9			1	1.80	1	1.30	1	1.07	1	1.15
10			1	1.46						
11			1	1.26						
	a = 199.0425, b = 0.7684, R2 = 0.9955		a = 122.3924, b = 0.5593, R2 = 0.9693		a = 171.8836, b = 0.7073, R2 = 0.9855		a = 234.5943 , b = 0.9013, R2 = 0.9956		a = 191.5495, b = 0.7913, R2 = 0.9933	

Rank	Nr 6		Nr 7		Nr 8		Nr 9		Nr 10	
	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1
1	60	67.20	76	78.62	82	85.20	60	62.31	67	67.63
2	58	42.74	52	43.41	56	44.52	43	36.59	39	35.17
3	26	27.31	18	24.17	14	23.50	19	21.66	12	18.52
4	17	17.59	13	13.66	13	12.63	11	13.00	11	9.98
5	5	11.46	8	7.92	4	7.01	8	7.97	8	5.61
6	3	7.59	3	4.78	3	4.11	5	5.04	6	3.36
7	2	5.16	2	3.06	3	2.61	1	3.35	2	2.21
8	2	3.62	1	2.13	2	1.83	1	2.36	1	1.62
9	2	2.65							1	1.32
	a = 105.0002, b = 0.4613, R2 = 0.9207		a = 142.9791, b = 0.6045, R2 = 0.9765		a = 162.8791, b = 0.6598, R2 = 0.9617		a = 105.6059, b = 0.5438, R2 = 0.9803		a = 129.9497, b = 0.6679, R2 = 0.9819	

Distribution of Syllables in Russian Sonnets

	Nr 11		Nr 12		Nr 13		Nr 14		Nr 15	
Rank	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1
1	70	69.79	53	52.23	53	55.76	64	65.68	72	71.98
2	36	35.13	30	30.27	38	29.94	42	36.25	35	35.52
3	13	17.93	13	17.72	14	16.29	16	20.21	20	17.79
4	12	9.40	13	10.56	5	9.08	10	11.47	6	9.16
5	9	5.17	10	6.46	3	5.27	7	6.71	6	4.97
6	4	3.07	7	4.12	2	3.26	5	4.11	3	2.93
7	2	2.03	1	2.78	1	2.19	2	2.70	3	1.94
8	1	1.51	1	2.02	1	1.63	1	1.92	1	1.46
9							1	1.50	1	1.22
10									1	1.11
	a = 138.6755, b = 0.7010, R2 = 0.9878		a = 89.6588, b = 0.5597 R2 = 0.9749		a = 103.6152, b = 0.6378, R2 = 0.9629		a = 118.6674, b = 0.6069, R2 = 0.9849		a = 145.9485, b = 0.7209, R2 = 0.9963	

	Nr 16		Nr 17		Nr 18		Nr 19		Nr 20	
Rank	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1
1	72	72.78	67	72.34	77	81.13	87	86.59	90	91.47
2	39	35.30	60	42.75	58	44.47	40	41.60	48	41.84
3	13	17.39	13	25.44	16	24.58	22	20.26	14	19.43
4	8	8.83	12	15.30	10	13.79	9	10.13	7	9.32
5	6	4.74	11	9.37	8	7.94	6	5.33	4	4.76
6	4	2.79	8	5.90	4	4.77	4	3.05	4	2.70
7	3	1.86	1	3.87	1	3.04	3	1.97	3	1.77
8	2	1.41	1	2.68			2	1.46	2	1.35
9	1	1.20	1	1.98			1	1.22	1	1.16
10									1	1.07
	a = 150.2100 b = 0.7384, R2 = 0.9915		a = 121.8946, b = 0.5357, R2 = 0.9023		a = 147.6958, b = 0.6116, R2 = 0.9458		a = 180.4356, b = 0.7459, R2 = 0.9985		a = 200.4053, b = 0.7954, R2 = 0.9897	

	Nr 21		Nr 22		Nr 23		Nr 24		Nr 25	
Rank	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1
1	74	75.28	77	79.49	77	80.69	75	75.14	71	72.51
2	47	41.67	51	43.03	56	42.68	37	35.86	42	36.24
3	18	23.27	19	23.50	13	22.80	15	17.39	14	18.37
4	11	13.20	10	13.05	8	2.40	9	8.71	7	9.56
5	10	7.68	9	7.45	6	6.96	6	4.62	5	5.22
6	6	4.66	2	4.46	6	4.12	4	2.70	5	3.08
7	4	3.00	2	2.85	4	2.63	2	1.80	2	2.02
8	2	2.10	1	1.99	2	1.85			1	1.50
9	2	1.60	1	1.53	2	1.44				
10	1	1.33	1	1.28	1	1.23				
11			1	1.15						
	a = 135.6361 b = 0.6022, R2 = 0.9864		a = 146.5836, b = 0.6246, R2 = 0.9826		a = 152.3845, b = 0.6482, R2 = 0.9505		a = 157.6637, b = 0.7546, R2 = 0.9975		a = 145.1118, b = 0.7077, R2 = 0.9852	

	Nr 26		Nr 27		Nr 28		Nr 29		Nr 30	
Rank	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1
1	52	51.37	83	81.72	81	81.65	86	85.50	65	65.57
2	23	25.41	37	41.93	46	42.95	40	40.01	40	36.35
3	15	12.83	27	21.75	19	22.82	14	19.01	14	20.36
4	6	6.73	10	11.52	12	12.35	12	9.32	13	11.60
5	6	3.78	6	6.34	8	6.90	10	4.84	8	6.80
6	2	2.35	6	3.71	7	4.07	5	2.77	7	4.18
7	1	1.65	3	2.37	1	2.60	4	1.82	3	2.74
8			2	1.70	1	1.83	3	1.38	2	1.95
9			2	1.35					1	1.52
10			1	1.18					1	1.29
	a = 1033.9461, b = 0.7245, R2 = 0.9914		a = 159.1986, b = 0.6791 R2 = 0.9897		a = 155.0642, b = 0.6537 R2 = 0.9931		a = 183.0284, b = 0.7729 R2 = 0.9874		a = 117.9236, b = 0.6023 R2 = 0.9833	

	Nr 31		Nr 32		Nr 33		Nr 34		Nr 35	
Rank	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1
1	64	65.07	55	60.51	56	61.27	50	53.30	45	45.24
2	40	34.75	51	35.97	52	35.53	40	30.12	23	23.62
3	12	18.78	15	21.55	11	20.78	12	17.22	17	12.57
4	10	10.37	12	13.08	8	12.34	9	10.03	4	6.92
5	7	5.93	3	8.10	6	7.50	3	6.03	1	4.03
6	6	3.60	3	5.17	5	4.72	2	3.80		
7	5	2.37	2	3.45	3	3.13	2	2.56		
8	2	1.72	2	2.44	3	2.22	2	1.87		
9	1	1.38	2	1.85	1	1.70				
10			1	1.50	1	1.40				
	a = 121.6213, b = 0.6410, R2 = 0.9757		a = 101.2733, b = 0.5316 R2 = 0.9143		a = 105.1804, b = 0.5569, R2 = 0.8951		a = 93.9283, b = 0.5855, R2 = 0.9413		a = 86.5155, b = 0.6706, R2 = 0.9696	

	Nr 36		Nr 37		Nr 38		Nr 39		Nr 40	
Rank	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1	Frequ	Exp+1
1	66	66.49	73	73.85	93	92.69	64	64.33	70	71.27
2	38	35.09	39	36.28	41	41.54	38	35.07	41	34.74
3	14	18.74	18	18.08	18	18.92	13	19.33	9	17.20
4	11	10.23	5	9.27	8	8.92	13	10.86	8	8.78
5	7	5.81	5	5.01	8	4.50	8	6.31	8	4.74
6	4	3.50	5	2.94	6	2.55	7	3.85	5	2.79
7	4	2.30	1	1.94	2	1.68	2	2.54	3	1.86
8	3	1.68	1	1.45			1	1.83	1	1.41
9	1	1.35	1	1.22			1	1.44	1	1.20
10							1	1.24		
	a = 125.9305, b = 0.6530, R2 = 0.9899		a = 150.4521, b = 0.7252, R2 = 0.9934		a = 207.3951, b = 0.8162, R2 = 0.9959		a = 117.7195, b = 0.6199, R2 = 0.9823		a = 146.3284, b = 0.7335 R2 = 0.9718	

Distribution of Syllables in Russian Sonnets

Rank	Nr 41		Nr 42		Nr 43		Nr 44		Nr 45		Nr 46	
	Fr	Exp+1	Fr	Exp+1	Fr	Exp+1	Fr	Exp+1	Fr	Exp+1	Fr	Exp+1
1	62	66.04	53	54.53	62	61.33	59	63.03	77	77.66	70	72.26
2	49	36.34	35	29.23	22	25.02	49	36.49	44	40.52	45	36.45
3	14	20.20	10	15.89	13	10.57	13	21.30	16	21.38	11	18.64
4	6	11.43	10	8.85	7	4.81	12	12.62	13	11.50	10	9.77
5	5	6.67	5	5.14	3	2.52	6	7.65	7	6.42	4	5.36
6	5	4.08	3	3.18	1	1.60	4	4.80	4	3.79	2	3.17
7	2	2.67	1	2.15			2	3.18	4	2.44	2	2.08
8	1	1.91	1	1.61			1	2.24	3	1.74	1	1.54
9	1	1.49							1	1.38	1	1.27
10	1	1.27										
11	1	1.15										
	a =119.6948 b = 0.6010, R ² = 0.9453		a =101.5021 b = 0.6399 R ² = 0.9710		a =151.4751 b = 0.9207 R ² = 0.9920		a =108.4224 b = 0.5564 R ² = 0.9308		a =148.6928 b = 0.6625 R ² = 0.9907		a =143.2549 b = 0.6982 R ² = 0.9708	

As seen from Table 3, the results are quite satisfactory with R² in all cases reaching the values greater than or equal to 0.9. The only exception in Nr. 33 can be rounded to 0.9.

Some sonnets have specific peculiarities. T.6, *Sonet, tri raznye sistemy zakljuchajushhij* by A. Rzhevskij (18th c.), was written in the experimental form. The author transferred thematic composition of a canonic sonnet from its vertical direction to a horizontal one. Thus instead of developing the plot from stanza to stanza he did it within the lines, opposing in the meaning of their beginnings and ends.

Sonnets T.17 (*Pojetu*) and T.18 (*Madona*) belong to A. Pushkin, the poet who made a revolution in Russian language of poetry and style. A. Pushkin wrote only three sonnets – all in 1830. If in his first sonnet (T.16 – *Sonnet*) he tried to observe the canonic sonnet pattern, two others were written in a rather free form.

Sonnets T.32 and T.33 are the first and the second sonnets in a crown of sonnets (*Crown of sonnets Corona Astralis*) written by M. Voloshin in 1909. Being part of a larger poetic work they acquire a strophic status which makes them less independent and leads to certain changes in the plot when the last tercet (especially its last line) instead of making a strong break in the general plot weakens the plot composition.

These facts might be taken into account as possible causes for the slight deviations from the common tendency in fitting (R²), but of course such possible influence of the above-mentioned factors needs further investigation.

The next step of the analysis consists in the study of syllabic length-types, when syllables are classified according to their length in phonemes.

It was found out that the exponential plus 1 function judging by the determination coefficients fitted the given data cannot be used here because the frequencies have a bell-shape hence the Lorentzian function (Wimmer, Altmann 2005) was chosen for fitting.

The Lorentzian function is defined as:

$$y = a / (1 + ((x - b) / c)^2)$$

where *a*, *b* and *c* are parameters. The parameter *b* shows approximately the turning point of the function. As can be seen, it is always between *x* = 2 and *x* = 3.

The results of such counts and fitting of the Lorentzian function are demonstrated in Table 4.

Table 4
Fitting of ranks of length syllable types (Lorentzian function)

L	T 1		T 2		T 3		T 4		T 5	
	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor
1	8	9.85	12	14.41	9	12.99	10	10.48	7	9.94
2	98	97.94	78	77.73	87	86.72	99	98.98	94	93.90
3	65	65.13	65	65.43	61	61.58	54	54.08	58	58.27
4	11	8.55	18	13.05	18	10.97	10	8.29	13	8.32
5			1	5.08	1	4.23	1	3.16	2	3.13
6			2	2.66						
	a = 1710.3397 b = 2.4467 c = 0.1101 R ² = 0.9984		a = 150.2663 b = 2.4590 c = -0.4751 R ² = 0.9959		a = 201.8410 b = 2.4329 c = 0.3757 R ² = 0.9862		a = 452.8212 b = 2.4105 c = -0.2171 R ² = 0.9989		a = 613.3345 b = 2.4325 c = 0.1839 R ² = 0.9950	

L	T 6		T 7		T 8		T 9		T 10	
	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor
1	5	12.82	13	12.40	13	11.60	8	9.96	6	9.58
2	63	61.99	79	79.04	85	85.08	65	64.86	75	74.82
3	84	84.55	70	69.94	70	69.88	62	62.16	51	51.40
4	21	15.36	11	11.68	9	10.66	12	9.76	14	8.15
5	2	2.90					-	-	1	3.10
6							1	1.88		
	a = 151.0164 b = 2.5748 c = 0.4796 R ² = 0.9823		a = 210.6948 b = 2.4764 c = -0.3691 R ² = 0.9998		a = 297.0366 b = 2.4668 c = -0.2957 R ² = 0.9990		a = 195.2480 b = 2.4922 c = -0.3432 R ² = 0.9975		a = 245.1225 b = 2.4373 c = 0.2898 R ² = 0.9875	

L	T 11		T 12		T 13		T 14		T 15	
	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor
1	9	12.72	10	12.96	3	5.98	5	9.47	6	8.03
2	74	73.67	60	59.58	55	54.79	71	70.76	75	74.93
3	48	48.76	43	43.78	52	51.89	58	58.35	55	55.13
4	16	10.06	15	10.49	6	5.87	14	8.73	10	7.20
5					-	-			2	2.65
6					1	1.08				
	a = 129.7294 b = 2.4037 c = 0.4627 R ² = 0.9816		a = 89.0446 b = 2.4088 c = -0.5814 R ² = 0.9822		a = 505769.0 b = 2.4932 c = 0.0051 R ² = 0.9971		a = 264.2540 b = 2.4682 c = -0.2831 R ² = 0.9848		a = 870.5851 b = 2.4587 c = 0.1408 R ² = 0.9972	

L	T 16		T 17		T 18		T 19		T 20	
	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor
1	6	7.62	11	12.33	4	9.10	6	9.36	7	8.95
2	80	79.93	75	74.88	87	86.60	91	90.89	93	92.93
3	52	52.02	72	72.16	74	73.67	62	62.24	62	62.06

Distribution of Syllables in Russian Sonnets

4	8	6.61	15	12.11	8	8.63	13	8.19	10	7.82
5	2	2.44	1	4.55	1	3.14	2	3.04	2	2.89
	a = 174432.7 b = 2.4465 c = 0.0096 R ² = 0.9990		a = 197.0080 b = 2.4926 c = 0.3857 R ² = 0.9955		a = 1056135.1 b = 2.4798 c = 0.0043 R ² = 0.9956		a = 1265.2171 b = 2.4498 c = -0.1251 R ² = 0.9943		a = 155.139.2 b = 2.4497 c = -0.0110 R ² = 0.9986	

L	T 21		T 22		T 23		T 24		T 25	
	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor
1	11	11.82	9	10.77	6	8.94	9	8.24	5	7.70
2	84	83.95	79	78.90	90	89.91	81	81.02	78	77.86
3	65	65.09	71	71.14	65	65.17	52	51.94	56	55.96
4	12	10.53	13	10.31	12	8.03	6	7.04	8	6.90
5	3	3.99	2	3.83	2	2.95				
	a = 247.1871 b = 2.4547 c = 0.3261 R ² = 0.9993		a = 316.2128 b = 2.4830 c = 0.2785 R ² = 0.9975		a = 40689.9 b = 2.4598 c = 0.0216 R ² = 0.9960		a = 906.0401 b = 2.4404 c = 0.1380 R ² = 0.9996		a = 374793.7 b = 2.4588 c = -0.0066 R ² = 0.9978	

L	T 26		T 27		T 28		T 29		T 30	
	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor
1	6	6.15	10	10.70	7	9.72	12	14.08	14	15.57
2	54	53.99	89	88.97	89	88.90	89	88.85	72	71.76
3	38	38.02	65	65.07	65	65.20	54	54.40	48	48.58
4	6	5.38	11	9.45	13	8.68	14	10.82	16	11.88
5	1	2.01	2	3.54	1	3.22	5	4.29	3	4.88
6									1	2.62
	a = 309.3204 b = 2.4488 c = 0.2064 R ² = 0.9994		a = 413.9675 b = 2.4522 c = 0.2366 R ² = 0.9991		a = 808.2327 b = 2.4573 c = 0.1608 R ² = 0.9950		a = 164.5965 b = 2.3935 c = 0.4262 R ² = 0.9971		a = 102.5293 b = 2.3832 c = 0.5853 R ² = 0.9935	

L	T 31		T 32		T 33		T 34		T 35	
	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor
1	6	10.37	3	9.43	6	9.44	2	6.25	4	4.78
2	71	70.72	58	57.47	61	60.74	59	58.67	45	44.83
3	52	52.52	66	66.40	64	64.25	52	51.74	40	39.68
4	15	9.00	16	10.10	14	9.68	5	5.99	1	4.59
5	3	3.44	3	3.71	1	3.58	2	2.17		
	a = 179.9874 b = 2.4438 c = 0.3570 R ² = 0.9849		a = 184.7478 b = 2.5271 c = -0.3542 R ² = 0.9793		a = 202.7425 b = 2.5101 c = 0.3336 R ² = 0.9902		a = 901708.1 b = 2.4843 c = 0.0039 R ² = 0.98842		a = 711295.9 b = 2.4848 c = 0.9938 R ² = 0.9916	

L	T 36		T 37		T 38		T 39		T 40	
	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor
1	11	11.75	5	7.74	8	9.46	8	8.57	8	10.35
2	70	69.94	78	77.81	99	98.96	77	76.98	75	74.86
3	52	52.11	57	56.82	59	59.11	52	52.06	49	49.34

4	11	10.02	7	6.97	10	7.93	9	7.39	13	8.52
5	4	3.90	1	2.56			1	2.77	1	3.28
6							1	1.43		
	a = 140.4537 b = 2.4354 c = -0.4336 R ² = 0.9996		a = 631688.1 b = 2.4608 c = -0.0051 R ² = 0.9980		a = 2089.5338 b = 2.4335 c = -0.0967 R ² = 0.9989		a = 457.4086 b = 2.4434 c = 0.1995 R ² = 0.9988		a = 189.5879 b = 2.4234 c = -0.3420 R ² = 0.9923	

L	T 41		T 42		T 43		T 44		T 45		T 46	
	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor	Fr	Lor
1	5	7.32	5	8.52	7	5.82	6	10.01	7	12.18	2	7.46
2	67	66.85	56	55.75	63	63.03	60	59.66	81	80.65	74	73.83
3	64	63.95	45	45.37	35	34.90	62	62.31	60	60.61	55	55.31
4	8	7.21	12	7.71	3	4.76	14	10.23	17	10.60	14	6.77
5	2	2.60					4	3.82	4	-	1	2.49
6	1	1.33								2.12		
	a = 363937.9 b = 2.4945 c = 0.0067 R ² = 0.9987		a = 143.726 b = 2.4604 c = -0.3665 R ² = 0.9832		a = 14388.396 b = 2.4242 c = -0.0908 R ² = 0.9981		a = 163.862 b = 2.5086 c = 0.3849 R ² = 0.9911		a = 195.543 b = 2.4444 c = -0.3724 R ² = 0.9850		a = 9774.039 b = 2.4637 c = 0.0405 R ² = 0.9810	

The function gives a very good fit and the ranking of syllabic length-types is caught better than for phonetic types. In majority of cases $R^2 = 0.99$ with very few exceptions. Out of the four periods the best fitting is observed for the sonnets of the Silver Age when this genre was very popular among the poets who introduced into it many new forms, especially in the rhyme scheme.

Conclusions

The results demonstrate good fitting of the ranks of both types of syllables, but for length-types they are to some extent better than for phonemic types. This may suggest that the distribution of syllable length positions is determined by the genre itself which does not allow considerable variations but the filling of these positions by concrete syllabic types is to some extent optional and depends on the author's style.

It must be admitted that at this stage of analysis a rather high level of generalization for defining syllable categories was used. One of further possible directions is to introduce distinctions into syllable types using a more detailed classification of phonemes. Further, the above models should be applied also to other texts in the same and other languages. It may be conjectured that synthetic languages strongly differ from analytic ones, e.g. many Polynesian languages have only two types of syllable (V and CV), hence the above results may be used also in typology.

References

- Andreev, S., Popescu, I.-I., Altmann, G. (2017). Some problems of adnominals in Russian texts. *Glottometrics* 38, 77–106.
- Avanesov, R.I. (1954). O slogorazdele i stroenii sloga v russkom jazyke. *Voprosy jazykoznanija*. 6, 88-101.

- Belyj, A.** (1910). *Simvolizm*. Musaget: Moskva.
- Gasparov, M.L.** (2012). Fonetika, morfologija i sintaksis v bor'be za stih. In: *M.L. Gasparov. Izbrnnye trudy*, 4: 325-334.
- Köhler, R., Altmann, G.** (2014). *Problems in Quantitative Linguistics*, 4. Lüdenscheid: RAM-Verlag.
- Reformatskij, A.A.** (1950). *Metodicheskie ukazanija i rukovodstvo po sovremennomu russkomu jazyku dlja studentov-zaochnikov*. Moskva: Moskovskij gosudarstvennyj pedagogičeskij institut.
- 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: de Gruyter.

The Distribution of Synonymous Variants in Wenzhounese

Jieqiang Zhu¹, Haitao Liu^{1,2}

Abstract. This study investigates the diversification phenomenon of synonymy in Wenzhounese, a variety of Chinese spoken around the city of Wenzhou in southern China. Five groups of synonymous variants were extracted from Wenzhou Spoken Corpus (WSC) for finding the regularity of their rank-frequency distribution. The result revealed that the rank-frequency distribution of the investigated synonymous variants abides by the right-truncated modified Zipf-Alekseev distribution, a unified function modelling the diversifying process.

Key words: Diversification, Wenzhounese, Zipf-Alekseev Distribution

1. Introduction

Diversification – also known as variation, dialectal variants, polysemy, word associations, multifunctionality, allophony and allomorphy, classes of style, and spelling errors in different domains of linguistics – is one of the most productive and powerful processes in languages. It can take place within or between different levels of language, such as concept, unit, meaning/function, and category. A great variety of grammatical diversification phenomena can be found in Rothe (1991). For years, much effort was made in figuring out a unified distribution to model diversification process (e.g., Altmann 1991; Köhler 2012). Eventually, Zipf-Alekseev pattern, a well-known Zipf's Law-related distribution, was found (Altmann 2016). It is now widely used in investigating diversification process at or among a variety of linguistic levels, such as at the lexical level (see Chen & Liu 2014; Čech & Uhlířová 2014; Mohanty & Popescu 2014), at the syntactic level (Liu 2009), at the semantic level (Liu 2012), and at the discourse level (Yue & Liu 2012; Zhang & Liu 2015). The fitting results of these investigations revealed that most of the distributions are excellently fitted with a modified right-truncated Zipf-Alekseev distribution.

One of the most common phenomena of diversification progress is synonymy – the case where one concept can be expressed by a variety of units. However, the quantitative study concerning this phenomenon is scanty. Using LAMSAS corpus data, Kretzschmar (2015) has listed several groups of concepts that have many synonymous variants, the number of which ranges from a low of 39 for *dry spell*, to 239 for *cobbler*, and found out that the rank-frequency of variants of each group conform vaguely to Zipf's 80/20 Rule. This study, however, has not fitted a specific distribution to the data.

¹ Department of Linguistics, Zhejiang University, China; ² Centre for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China. Correspondence to: Haitao Liu. Email address: htliu@163.com.

Thus, this paper aims at investigating whether the rank-frequency distribution of synonymous variants abides by a certain law in a specific language. We choose Wenzhounese, a variety of Chinese spoken around the city of Wenzhou in southern China (Newman et al. 2007). There are seven or ten Chinese dialect groups in China; Mandarin Chinese is a group of related varieties of Chinese spoken across most of northern and southeastern China. This group includes the Beijing dialect, which is the basis of the Standard Mandarin. At the phonological and syntactic levels, many other Chinese dialects have significant differences compared to Mandarin Chinese, one of the extreme examples of which is Wenzhounese. Having little to no mutual intelligibility with any other variety of Chinese, Wenzhounese has the nickname “Devil’s Language” for its difficulty and complexity.² Due to its long history and the isolation of the region, Wenzhounese preserves a large amount of lexemes and distinctive phonological and grammatical systems of classical Chinese lost elsewhere (e.g., Scholz & Chen 2014; Sheng 2004). However, at the same time, it has been inevitably influenced by mandarin Chinese (Putonghua) since 1950s, when this national language began to be used in education, in the media, and at formal occasions.

This paper consists of four sections. Section 2 describes the material and methods used. Section 3 presents the results of the distribution investigated. Section 4 concludes this study.

2. Material and Method

The material of this study is selected from Wenzhou Spoken Corpus (WSC)³, an online searchable corpus developed by Jinxia Lin and John Newman, and technically supported by the Text Analysis for Research Portal (TAPoR) team. This corpus consists of six sub-corpora: face-to-face conversation, phone call, Wenzhou news commentary, Internet chat, story, and Wenzhou song. These six genres provide a mix of informal and formal contexts, as well as of private and public uses of language (Newman et al. 2007). Most of the conversational data were collected around the city of Wenzhou from 2004 to the present day. The total word token count of the corpus is 154,710.

It is obvious that the corpus is of modest size. However, Zipf’s Law, as well as other distributional rules, is more clearly manifested in large-scale data, while in small-size language material, contingency may lead to the deviation from the original result to a certain degree. In order to reduce contingency effect, we will choose lexical items that meet the following three requirements: first, they should be of the most commonly used ones in the vocabulary of the intimate everyday life; second, they should come from cultural domains appropriate to women and men, the young and the old, from any socioeconomic group; third, they should have at least 10 synonymous variants available in the corpus.

This online corpus does not offer the whole text, but provides us with utterances of each word or phrase we type in. For example, if we search speeches containing the word 路 (*lov*, “road”), this online corpus will show us all the utterances containing the word 路 (*lov*), but other utterances without this word will not be shown. To tackle this problem, we strive to select words or phrases that meet the three requirements mentioned above from Shen

² <https://en.wikipedia.org/wiki/Wenzhounese>

³ <http://www.artsrn.ualberta.ca/wenzhou/>

Kecheng and Shen Jia's books, the *Wenzhouhua* (2004) series, in which there is an exhaustive list of words or phrases used in everyday Wenzhounese as well as detailed grammatical and syntactical explanations about these lexical items. Eventually, we have managed to select out 5 groups of variants that meet these premises above, and by searching with "concordance" and "keyword + whole utterance", we have got all the utterances containing these variants and copied them into Excel to count their frequencies of occurrence.

The five expressions we have managed to extract are:

- 1) 姆姆 (*mai mai*): child or kid; a young human that is not yet an adult
- 2) 路 (*lov*): road, street; a pathway that allows for pedestrians and vehicles to pass through;
- 3) 覈 (*gau*): here; in, at, or to this position or place
- 4) 老早 (*le ze*): before; a previous occasion
- 5) 显 (*xi*): very; in high degree; used to give emphasis

For convenience of illustration, here we use the variants that have the highest frequency in each group to indicate the whole group of synonymous expressions. Each variant's pronunciation is labelled next to the words in brackets. The phonetic notation system we adopt here was structured by Shen Kecheng and Shen Jia in *Wenzhouhua* (2004), and is by far the most developed and professional one for Wenzhounese. It is based on *Pinyin*, the Chinese phonetic alphabet, while some adjustments were made to fit the Wenzhounese's unique phonetic system.

Supposing that the phenomenon of synonymy is one kind of diversifying process, we assume that the distribution of synonymous variants obeys the Zipf-Alekseev model (Hřebíček 1996, cited from Strauss & Altmann, 2006). Hřebíček used two assumptions:

(i) The logarithm of the ratio of the probabilities P_l and P_x is proportional to the logarithm of the class size, i.e. –

$$\ln(P_1/P_x) \propto \ln x$$

(ii) The proportionality function is given by the logarithm of Menzerath's law (Hierarchy), i.e. –

$$\ln(P_1/P_x) = \ln (cx^b) \ln x$$

yielding the solution –

$$(1) \quad P_x = P_1 x^{-(\ln c + b \ln x)}, \quad x = 1, 2, 3, \dots$$

As $\ln c$ is a constant, one can write

$$P_x = P_1 x^{-(a+b \ln x)}, \quad x = 1, 2, 3, \dots$$

If (1) is considered a probability distribution, the P_l is the normalizing constant; otherwise, it is estimated as the size of the first class, $x = 1$. Very often, diversification distribution displays a diverging frequency in the first class, while the rest of the distribution behaves regularly. In these cases, one usually ascribes the first class a special value α , modifying (1) as

$$(2) \quad P_x = \left\{ \frac{(1 - \alpha)x^a x^{-(a+b \ln x)}}{T} \right.$$

where

$$T = \sum_{j=2}^n j^{-(a+b \ln j)}, a, b \in \mathfrak{R}, 0 < \alpha < 1.$$

Distributions (1) or (2) are called Zipf-Alekseev distributions. If n is finite, (2) is called a modified right-truncated Zipf-Alekseev distribution.

Then, we use the Altmann-Fitter software for fitting the model to the data observed.

3. Results and Discussion

3.1 The Fitting Result by Zipf-Alekseev Distribution

Tab. 1 to 5 show the fitting result of the five groups of expressions. The first column contains the rank of frequency, the second one variants, the third one their frequencies, and the last one the corresponding right-truncated modified Zipf-Alekseev fit. Fig. 1 to 5 illustrate the results in Tab. 1 to 5. The Wenzhou phonetic transcription and explanations in English for each variant are listed in Appendix.

Table 1
Fitting the right-truncated modified Zipf-Alekseev distribution to the data of synonymous variants meaning “mai mai”

X[i]	Variant	F[i]	NP[i]
1	姆姆	162	162.00
2	细儿	29	36.25
3	小孩	28	22.40
4	姆	17	14.04
5	小朋友	13	9.10
6	小孩子	4	6.10
7	孩子	3	4.20
8	姆佬	2	2.97
9	儿童	2	2.15
10	小宝宝	1	1.59
11	姆姆儿	1	1.19

$a = 0.0618, b = 0.6282, n = 11.0000,$ $\alpha = 0.6183, DF = 6, X^2 = 6.7871,$ $P = 0.3410, C = 0.0259$

In this and following similar tables: $X[i]$ – the rank of each variant; $F[i]$ – observed frequency; $NP[i]$ – calculated frequency according to the modified right-truncated Zipf-Alekseev distribution; a, b, n and α – the parameters of the modified right-truncated Zipf-Alekseev distribution; DF – degrees of freedom; X^2 – Chi-square; P – probability of Chi-square; C – discrepancy coefficient.

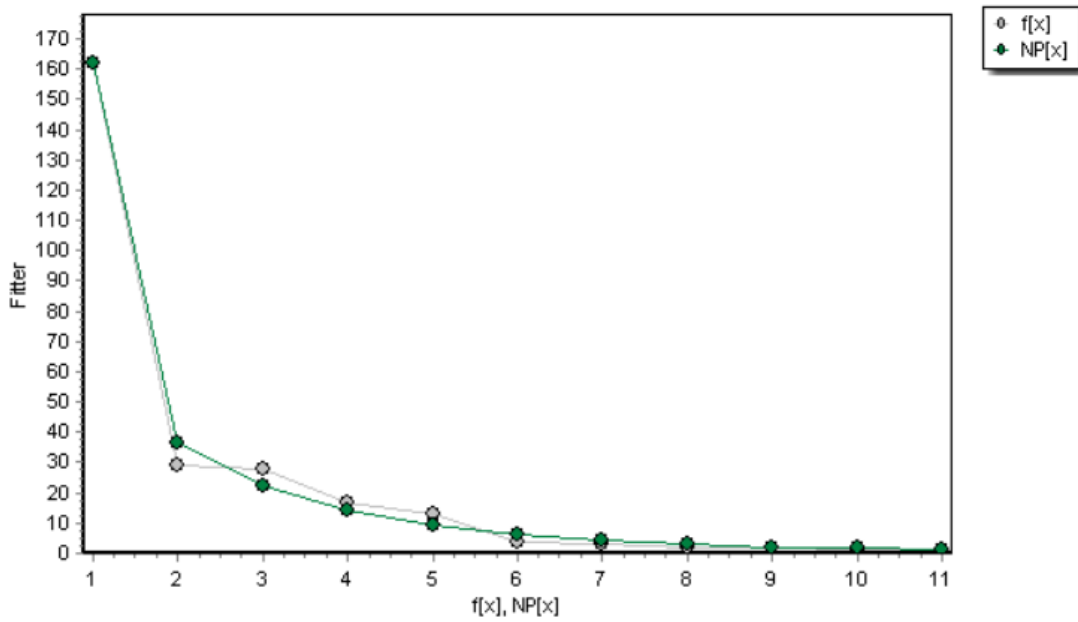


Figure 1. Graphic representation of Table 1

Table 2

Fitting the right-truncated modified Zipf-Alekseev distribution to the data of synonymous variants meaning “lov”

X[i]	Variant	F[i]	NP[i]
1	路	203	203.00
2	街	24	27.90
3	街道	20	21.00
4	街路	19	16.07
5	大道	17	12.58
6	大街	12	10.05
7	公路	12	8.17

8	道路	5	6.75
9	车路	3	5.64
10	马路	3	4.77
11	大路	2	4.07
$a = 0.1096, b = 0.3299, n = 11.0000,$ $\alpha = 0.6344, DF = 6, X^2 = 8.2422,$ $P = 0.2209, C = 0.0258$			

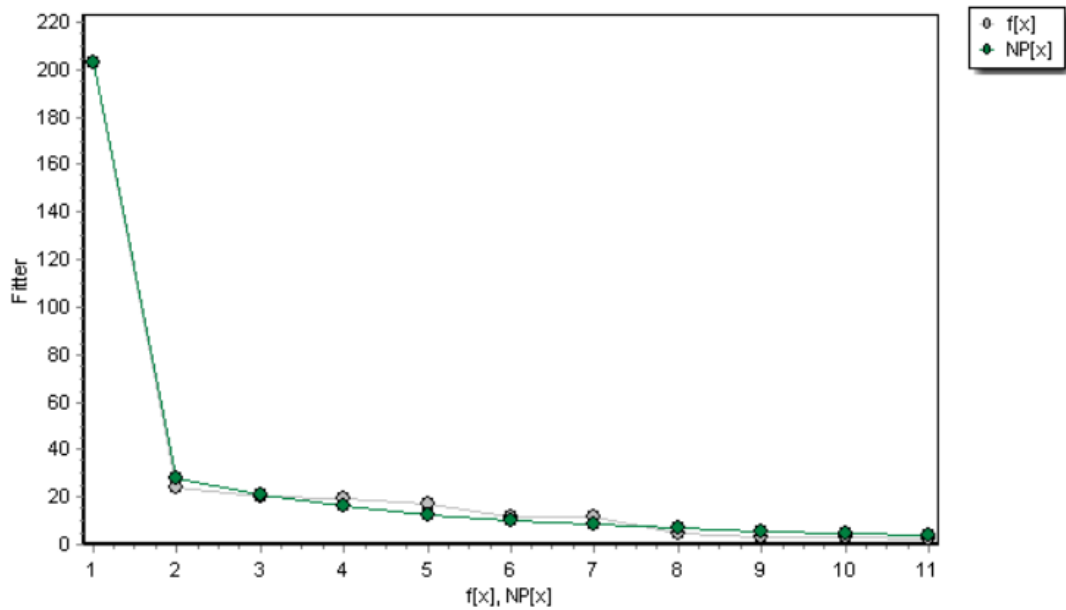


Figure 2. Graphic representation of Table 2

Table 3.

Fitting the right-truncated modified Zipf-Alekseev distribution to the data of synonymous variants meaning “gau”

X[i]	Variant	F[i]	NP[i]
1	穀	142	142.00
2	该里	92	81.51
3	该抵	30	48.92
4	穀穀	29	31.74
5	该角	25	21.80
6	这里	25	15.62
7	这边	14	11.58

8	彀宕	9	8.81
9	这	9	6.85
10	该	7	5.42
11	该遍	3	4.36
12	该个地方	2	3.56
13	该厘儿	2	2.93
14	该境	1	2.44
15	该境里	1	2.05
16	该伢	1	1.74
17	该头	1	1.49
18	彀园	1	1.28
19	这抵	1	1.11
20	这个地方	1	0.96
21	这伢	1	0.84

$a = 0.6276$, $b = 0.3526$, $n = 21.0000$,
 $\alpha = 0.3577$, $DF = 15$, $X^2 = 20.0111$,
 $P = 0.1715$, $C = 0.0504$

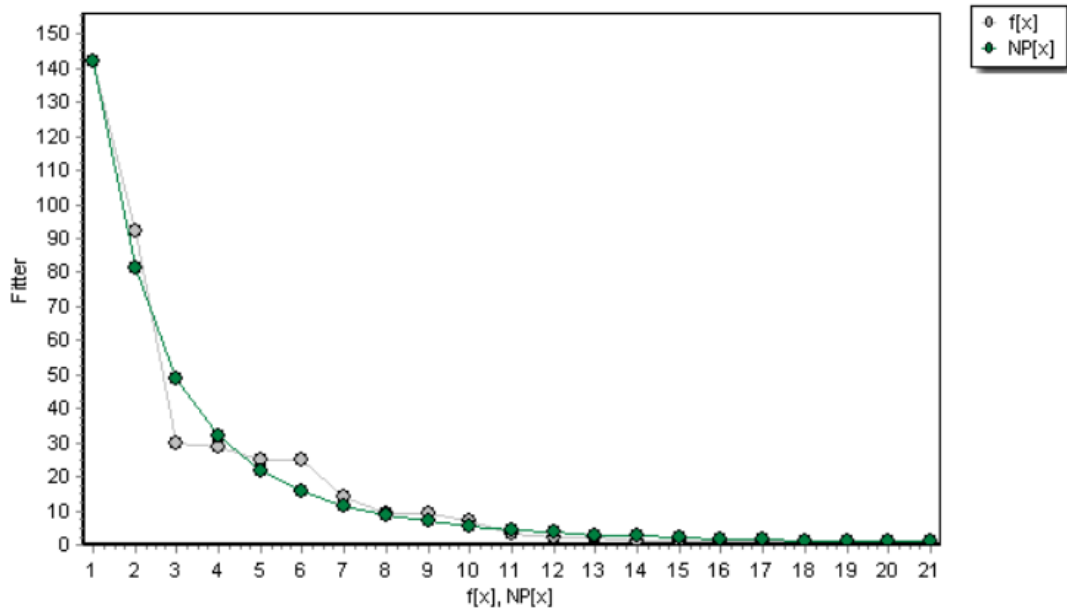


Figure 3. Graphic representation of Table 3

Table 4
Fitting the right-truncated modified Zipf-Alekseev distribution to the data of synonymous variants meaning “le ze”

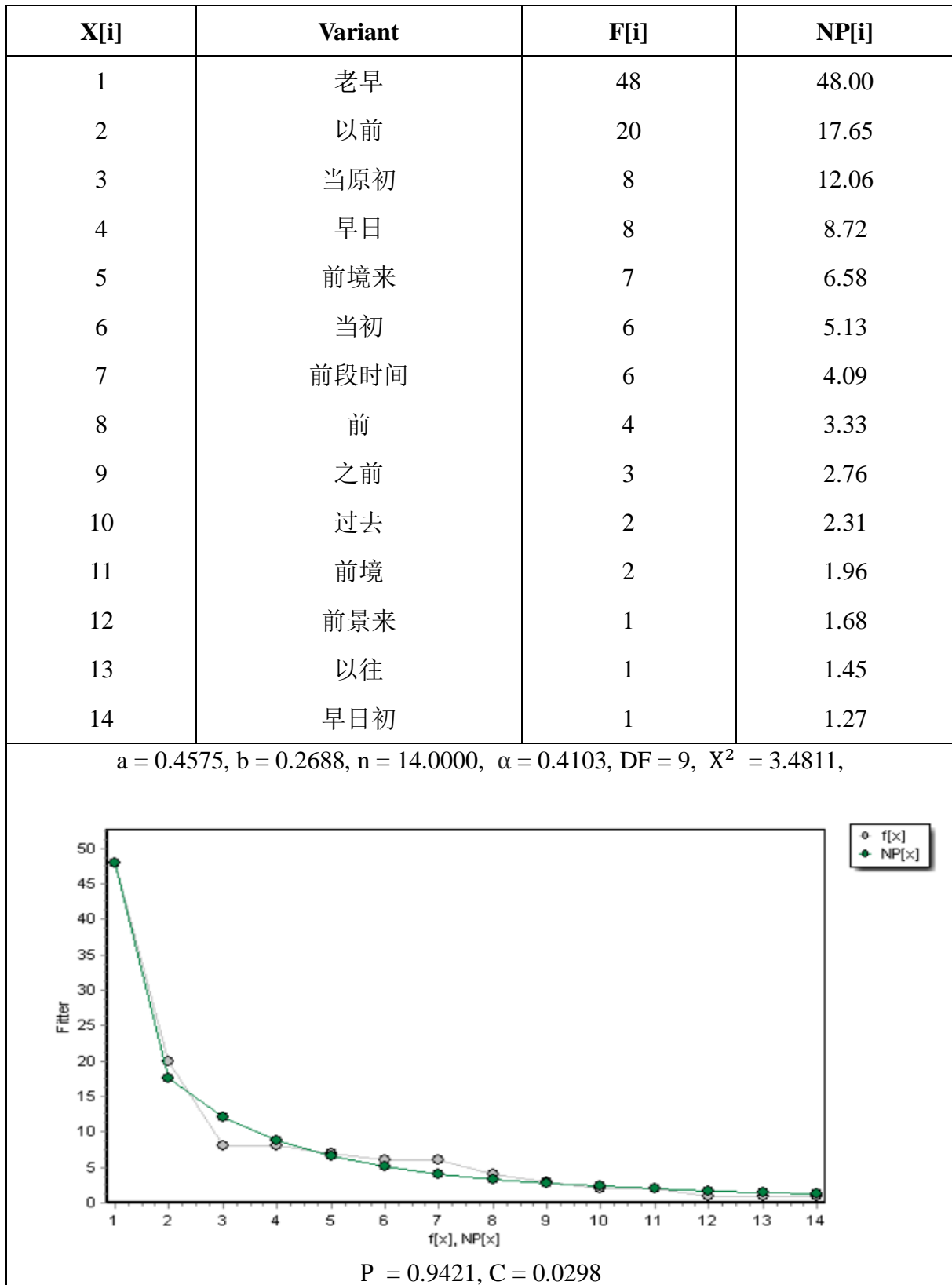


Figure 4. Graphic representation of Table 4

Table 5
Fitting the right-truncated modified Zipf-Alekseev distribution to the data
of synonymous variants meaning “xi”

X[i]	Variant	F[i]	NP[i]
1	显	774	774.00
2	很	97	116.21
3	恁	89	82.65
4	真	65	59.54
5	忒	51	43.94
6	真真	36	33.21
7	特别	35	25.63
8	短命	20	20.13
9	非常	18	16.07
10	大显	14	13.00
11	几俫	8	10.64
12	好	7	8.81
13	爻道	7	7.36
14	特	4	6.20
15	挺	4	5.26
16	分外	1	4.49
17	相当	1	3.86

a = 0.0658, b = 0.4323, n = 17.0000,
 $\alpha = 0.6288$, DF = 12, $X^2 = 16.2357$,
P = 0.1807, C = 0.0132

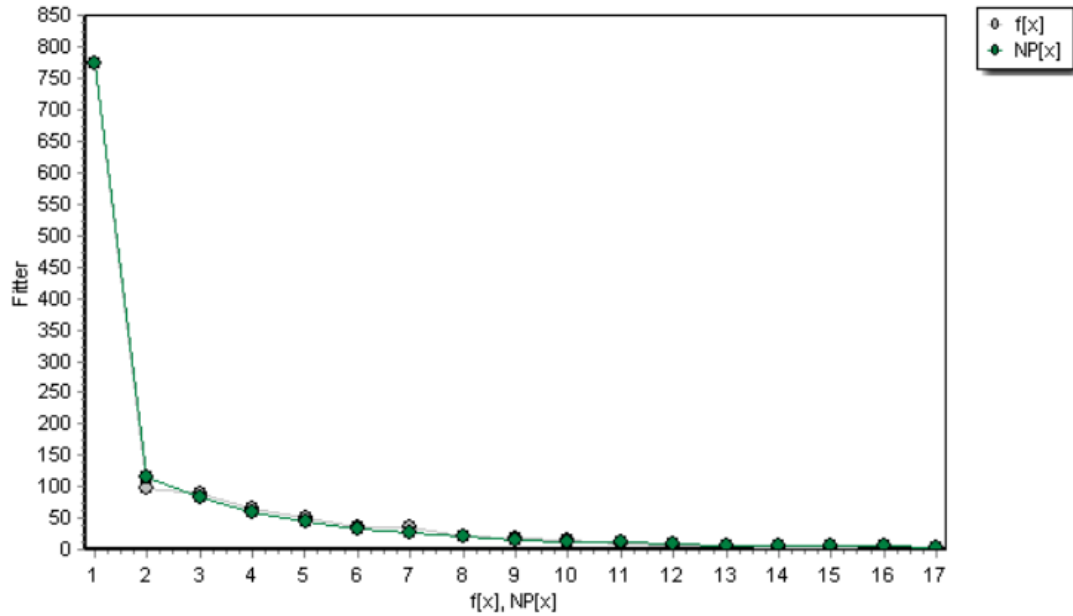


Figure 5. Graphic representation of Table 5

Table 6 shows the results of fitting the modified right-truncated Zipf-Alekseev distribution to the data of five groups of synonymous variants.

Table 6
Fitting the right-truncated modified Zipf-Alekseev distribution to the five groups of synonymous variants

NO.	χ^2	DF	P	C
1	6.7871	6	0.3410	0.0259
2	8.2422	6	0.2209	0.0258
3	20.0111	15	0.1715	0.0504
4	3.4811	9	0.9421	0.0298
5	16.2357	12	0.1807	0.0132

3.2. Discussion

As can be seen, Tables 1 to 6 show that there is a good fitting of the modified right-truncated Zipf-Alekseev distribution to all five groups of synonymous variants. Hence, the above hypothesis is compatible with the data.

Examining the five cases in detail, we have found that though all the collected lexical items are spoken in Wenzhounese, some of them are actually borrowed from Mandarin

Chinese, despite being pronounced according to Wenzhou dialect phonology. Take 姆姆 (*mai mai*, “child”), for example. More than half of the 11 variants – namely 小孩 (*xie hai*, “child”), 小朋友 (*xie bong yau*, “child”), 小孩子 (*xie hai zii*, “child”), 孩子 (*hai zii*, “child”), 儿童 (*ng dong*, “child”), and 小宝宝 (*xie bo bo*, “infant”) – are actually borrowed from Mandarin Chinese. These six variants have the frequency of 50 in total, which is 19% of all 262 entries. In the other four cases, however, the boundary between expressions from Wenzhounese and Mandarin Chinese is not clearly defined. Some expressions, such as 路 (*lov*, “way”), 当初 (*duo ceu*, “several months/years before”; “in the first place”), 这个地方 (*gi ga di fo*, “here”), cannot be said to belong to one particular language variety, but can have sources in both varieties, or are hybrids – the results of these two varieties’ interaction. One conclusion that can be made – in spite of the blurry boundary between these two varieties – is that the dialect system in one particular region is not closed and static, but an open and dynamic one that constantly absorbs in expressions from other language varieties; on the other hand, some of its traditional expressions get gradually lost, in a way a language evolves and interacts with other languages. Moreover, the expressions newly absorbed will not change the rank-frequency pattern of the whole group of variants, which abides by the right-truncated Zipf-Alekseev function.

4. Conclusion

This paper investigated five groups of synonymous variants from Wenzhounese, a unique dialect of Chinese, and finds out that the rank-frequency distributions of these variants fit well with the modified right-truncated Zipf-Alekseev distribution, a unified function modelling the diversifying process. The investigation verifies that synonymy phenomenon develops following the diversification process.

It should also be mentioned that there is a possibility that the figures presented in this investigation are influenced by the size and genre of the corpus, since the WSC is of modest size and not ideally balanced or representative. For example, more than 72 percent of the total word count come from the recorded News Commentary (Newman et al. 2007). Further study should investigate the synonymy phenomenon in other languages or language varieties, so as to check if the linguistic phenomenon found in this paper has a cross-language universality.

Acknowledgements

This work is partly supported by the National Social Science Foundation of China (Grant No. 17AYY021) and the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

References

- Altmann, G. (1991). Modeling diversification phenomena in language. In: Rothe, U. (ed.), *Diversification Processes in Language: Grammar: 33–46*. Hagen: Rottmann.
- Altmann, G. (2016). Types of hierarchies in language. *Glottometrics 34*, 44–55.
- Čech, R., Uhlířová, L. (2014). Adverbials in Czech: Models for their frequency distribution. In: Altmann, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.), *Empirical Approaches to Text and Language Analysis*. Lüdenscheid: RAM-Verlag, 2014, 49–49.
- Chen, H., & Liu, H. (2014). A diachronic study of Chinese word length distribution. *Glottometrics 29*, 81–94.
- Hřebíček, L. (1996). Word associations and text. In P. Schmidt (ed.), *Glottometrika 15*, 12–17. Trier: Wissenschaftlicher Verlag Trier.
- Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin / New York: de Gruyter.
- Kretzschmar, W., Jr. (2015). *Language and complex systems*. Cambridge: Cambridge University Press.
- Liu, H. (2009). Probability Distribution of Dependencies based on Chinese Dependency Treebank. *Journal of Quantitative Linguistics 16* (3): 256–273.
- Liu, H. (2012). Probability distribution of semantic roles in a Chinese treebank annotated with semantic roles. In: Altmann, G., Peter Grzybek, P., Naumann, S., Vulcanovic, R. (eds.), *Synergetic Linguistics. Text and Language as Dynamic Systems: 101–108*. Wien: Praesens Verlag.
- Mohanty, P., Popescu, I.-I. (2014). Word length in Indian languages 1. *Glottometrics 29*, 95–109.
- Newman, J., Lin Jingxia, Butler, T., Zhang, E. (2007). The Wenzhou Spoken Corpus. *Corpora 2*(1), 97–109.
- Rothe, U. (1991). Diversification processes in grammar. An introduction. In: Rothe, U. (Ed.), *Diversification Processes in Language: Grammar: 33–46*. Hagen: Rottmann.
- Scholz, F., Chen, Y. (2014). Sentence planning and f0 scaling in Wenzhou Chinese. *Journal of Phonetics 47*: 81–91.
- Shen, K., Shen, J. (2004). *Wenzhouhua*. Ningbo: Ningbo Press.
- Sheng, A. (2004). Dialect words in placename of Wenzhou and standardization problem. *Applied Linguistics 2*: 55–61.
- Strauss, U., Altmann, G. (2006). Diversification – laws in quantitative linguistics. Retrieved March 12, 2007, from http://www.uni-trier.de/uni/fb2/ldv/lql_wiki/index.php/Diversificatio.
- Yue, M., Liu H. (2011). Probability Distribution of Discourse Relations Based on a Chinese RST-annotated Corpus. *Journal of Quantitative Linguistics, 18*(2): 107–121.
- Zhang, H., Liu, H. (2015). Quantitative Aspects of RST Rhetorical Relations across Individual Levels. *Glottometrics 33*, 8–24.

Appendix

Table 1a

The phonetic transcription and explanations for each variants meaning “mai mai”

Rank	Variant
1	姆姆 mai mai (child)
2	细儿 sei ng (baby)
3	小孩 xie hhai (child)
4	姆 mai (child)
5	小朋友 xie bong yau (child)
6	小孩子 xie hhai zii (child)
7	孩子 hhai zii (child)
8	姆佬 mai le (baby)
9	儿童 ng dong (child)
10	小宝宝 xie bo bo (infant)
11	姆姆儿 mai mai ng (baby)

Table 2a

The phonetic transcription and explanations for each variants meaning “lov”

Rank	Variant
1	路 lov (way)
2	街 ga (street)
3	街道 ga dde (street)
4	街路 ga lov (street)
5	大道 ddeu dde (avenue)
6	大街 ddeu ga (street)
7	公路 gong lov (road)
8	道路 dde lov (way; road)

9	车路 co lov (road; street)
10	马路 mo lov (road; street)
11	大路 ddeu lov (road; street)

Table 3a.

The phonetic transcription and explanations for each variants meaning “gau”

Rank	Variant
1	骸 gau (here)
2	该里 gi lei (here)
3	该抵 gi dei (here)
4	骸骸 gau gau (here)
5	该角 gi go (here)
6	这里 ze lei (here)
7	这边 ze bi (here)
8	骸宕 gau dduo (here)
9	这 ze (here)
10	该 gi (here)
11	该遍 gi bie (here)
12	该个地方 gi gai di (here)
13	该厘儿 gi lei ng (here)
14	该境 gi jang (here)
15	该境里 gi jang lei (here)
16	该伧 gi kuo (here)
17	该头 gi ddeu (here)
18	骸园 gau kuo (here)
19	这抵 ze dei (here)
20	这个地方 ze gai di fo (here)
21	这伧 ze kuo (here)

Table 4a

The phonetic transcription and explanations for each variants meaning “xi”

X[i]	Variant
1	显 xi (very)
2	很 han (very)
3	恁 nang (so)
4	真 zang (very)
5	忒 teu (particularly)
6	真真 zang zang (quite)
7	特别 ddee bbi (particularly)
8	短命 ddov meng (particularly)
9	非常 fi ye (very)
10	大显 ddeu xi (quite)
11	几俵 gi lee (quite)
12	好 he (very)
13	爻道 hhuo dde (quite)
14	特 ddee (particularly)
15	挺 teing (quite)
16	分外 wang wei (particularly)
17	相当 xi duo (particularly)

Table 5a

The phonetic transcription and explanations for each variants meaning “le ze”

Rank	Variant
1	老早 le ze (before)
2	以前 yi yi (before)
3	当原初 duo nyv ceu (in the first place)
4	早日 ze nee (several months/years before; in the first place)
5	前境来 yi jang lee (several weeks before)

The Distribution of Synonymous Variants in Wenzhounese

6	当初 duo ceu (several months/years before; in the first place)
7	前段时间 yi ddoe ssii ga (several weeks/months ago)
8	前 yi (before)
9	之前 zii yi (before)
10	过去 gu qv (before)
11	前境 yi jang (several weeks before)
12	前景来 yi zang lee (several weeks before)
13	以往 yi wang (before)
14	早日初 ze nee ceu (in the first place)

Word Length and Polysemy in Japanese

Haruko Sanada¹, Gabriel Altmann

Abstract. In Japanese, we analyze the relation of word polysemy to word length measured in different ways: in terms of stroke numbers of signs, in terms of syllable numbers, and in terms of mora numbers. The synergetic law holds true for all aspects.

Keywords: Japanese, word length, polysemy

In a previous study (Sanada 2008), it has been stated that the number of strokes in Japanese signs influences the polysemy of the respective words. If one wants to specify the meaning, one must add something to the sign. As a result, one obtains a decreasing relationship: the more complex is a sign, the smaller the number of its meanings is. One usually begins to apply simple functions, as even in theory, they are easy to be treated. Sanada (2008) took into account the mean number of meanings of signs containing x number of strokes. The strokes, even the order of their writing, are codified in Japanese, and it is easy to decipher their number. Here, we consider merely the number of strokes, but not their complexity.

Now, since there are many words represented by a sign with x strokes, Sanada used rather the mean of polysemies and applied the power function to the resulting sequence. Since the determination coefficient yielded $R^2 = 0.75$, one usually looks after functions expressing this relation better. The problem is especially the fact that the smallest complexity ($x = 2$) yields a value (4.00) which is smaller than that of $x = 4$, which is 5.00 (while there is no $x = 3$). This is, of course, a boundary condition which could be solved by making use of a function consisting of two parts. But if one obtains different results for newer texts, or if one wants to consider the evolution or the dictionary, it is better to have a unique formula which can be substantiated synergetically (cf. Köhler 2005).

We start from the conjecture that the increase of polysemy is a controlled process: for the speaker, it is easier to have strongly polysemic words, as thereby, his coding effort decreases. On the other hand, the hearer wants to have monosemic words, as thereby, his decoding effort decreases. We may conjecture that the relative rate of change of polysemy can be expressed as dy/y , i.e. the greater y is, the smaller the rate of change will be. The speaker may change the value holding for the given language (k), but he is braked by the existing complexity of the signs, expressed with $r \cdot \ln(x)$, and by the hearer, who equilibrates the motion by mx . He also represents the language community. We thus obtain the differential equation

$$(1) \quad \frac{dy}{y} = \frac{k + r \cdot \ln(x)}{mx} dx$$

Since we omit all values of y smaller than 1, we may write in the above formula $y - 1$, which says that the rate of change is relativized by $(y-1)$. Solving the above equation and reparametrizing it, we obtain the well-known Zipf-Alekseev formula

$$(2) \quad y = 1 + cx^{a+b \cdot \ln x}.$$

¹ Rissho University, Tokyo, Japan. Correspondence to: Haruko Sanada. Email address: hsanada@ris.ac.jp.

Word Length and Polysemy in Japanese

Eliminating the boundary condition and taking the mean of $x = 2$ and $x = 4$, i.e. 3, and the mean of the two values of

$$y = \frac{4+5}{2} = 4.5 ,$$

we obtain the results presented in Table 1.

Table 1
Mean polysemy as a function of sign complexity in Japanese (cf. Sanada 2008: 119)

Number of strokes (complexity)	Average polysemy	Zipf-Alekseev formula	Number of data
3	4.50	4.40	2
5	2.50	2.71	2
6	2.25	2.36	4
7	1.63	2.13	8
8	2.45	1.97	11
9	1.36	1.85	14
10	2.00	1.76	9
11	2.13	1.68	31
12	1.57	1.63	51
13	1.84	1.58	49
14	1.64	1.54	55
15	1.50	1.50	64
16	1.45	1.47	71
17	1.51	1.45	83
18	1.32	1.42	68
18	1.52	1.42	82
20	1.47	1.38	73
21	1.38	1.37	58
22	1.44	1.35	70
23	1.52	1.34	58
24	1.38	1.33	40
25	1.28	1.32	47
26	1.30	1.31	40
27	1.29	1.30	24
28	1.50	1.29	20
29	1.60	1.28	15
30	1.20	1.28	15
31	1.33	1.27	12
32	1.33	1.26	6
33	1.40	1.26	5
34	1.00	1.25	2
35	1.20	1.25	5
36	1.00	1.24	1
37	1.00	1.24	1
38	1.00	1.23	1
41	1.00	1.22	1
42	1.00	1.22	1
a = -1.7328, b = 0.1429, c = 19.1761, R ² = 0.8777			

The results are graphically presented in Figure 1.

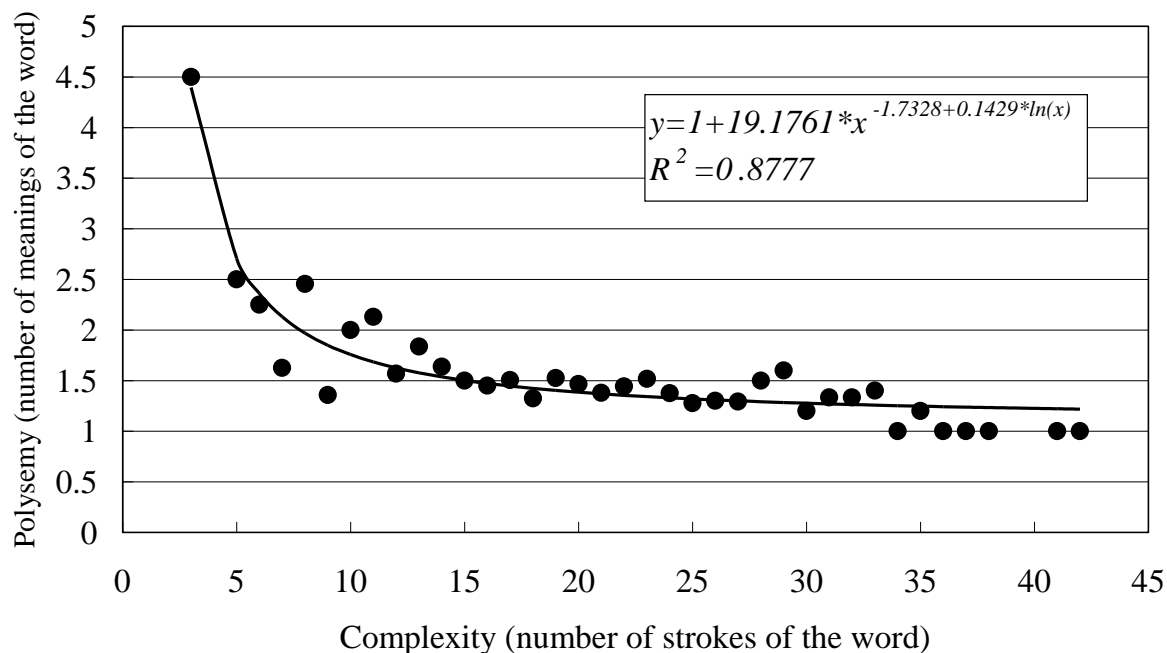


Figure 1. The relation of number of strokes and polysemy

The value of the determination coefficient is satisfactory. As can be seen, the parameters a and b have negative and positive values, respectively.

Since sign complexity computed in terms of stroke numbers is analogous to the length of words in other languages (in terms of, e.g., syllable numbers), the above formula is in concordance with the unified theory of length in language (cf. Popescu, Best, Altmann 2014). It would be very interesting to compare the results with data in other languages using signs, old or new ones. Further, it would be of interest to study the development of Japanese from this point of view.

As a matter of fact, word length is measured always in terms of syllable numbers. The fact that we used stroke numbers is possible only in some languages. In Old Egyptian or Assyrian, one could use quite other entities, but we conjecture that the results would be analogous.

Now, in Japanese, word length is also measured in terms of mora numbers. Morae are analogous to syllables. The results of measurement are presented in Table 2.

Table 2
Mean polysemy as a function of moraic word length in Japanese
(cf. Sanada 2008: 117)

Number of strokes (complexity)	Average polysemy	Zipf-Alekseev formula	Number of data
1	3.00	3.00	5
2	2.12	2.11	68
3	1.49	1.54	437
4	1.44	1.27	588
5	1.00	1.14	1
a = -0.2457, b = -0.8623, c = 1.9996, R ² = 0.9785			

Word Length and Polysemy in Japanese

As can be seen, length and polysemy are perfectly correlated here. The results are graphically presented in Figure 2.

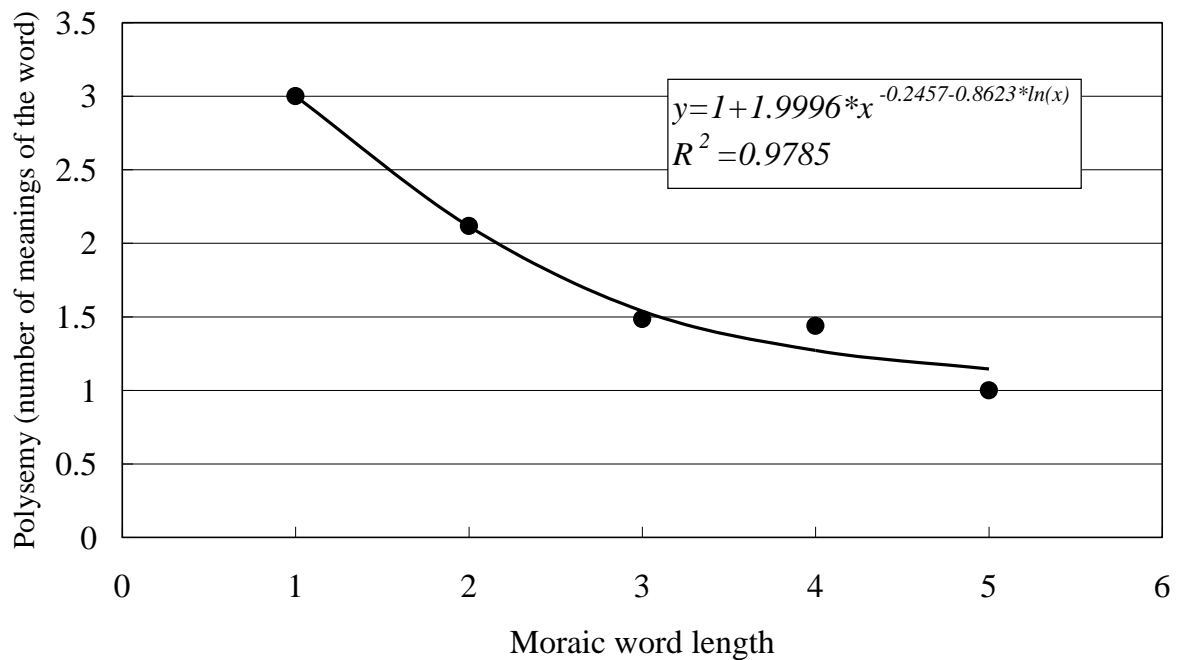


Figure 2. Mean polysemy as a function of moraic word length in Japanese

In the same way, it can be done by measuring word length in number of syllables. We obtain the results presented in Table 3 and Figure 3.

Table 3
Mean polysemy as a function of syllabic word length in Japanese
(cf. Sanada 2008: 118)

Number of strokes	Average polysemy	Zipf-Alekseev formula	Number of data
1	3.59	3.59	17
2	1.46	1.53	649
3	1.51	1.37	385
4	1.31	1.38	48
a = -3.1950, b = 1.3008, c = 2.5858, R ² = 0.9923			

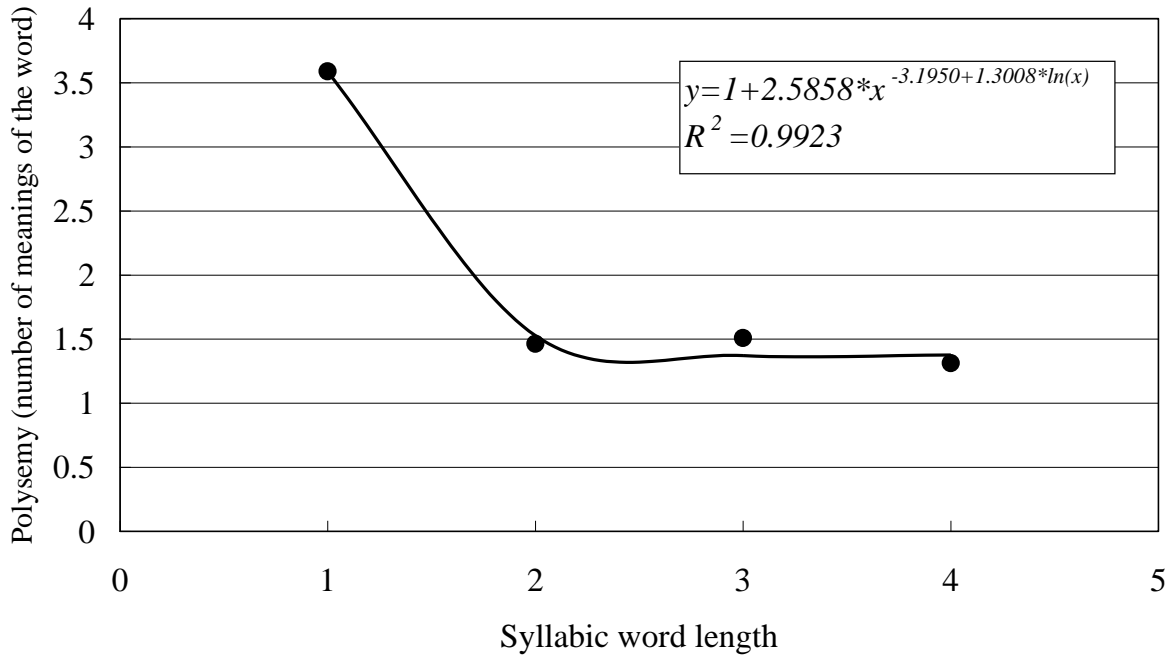


Figure 3. Mean polysemy as a function of syllabic word length in Japanese

In all cases, the hypothesis can be considered corroborated. A comparison with other languages would be, of course, important.

It is to be remarked that the number of data – as compared with the total lexicon – is very small, but it would not be sufficient to consider the complete lexicon; other languages must be analyzed, too, in order to state the validity of the hypothesis. The words we have chosen were taken from a list of 1,847 scholarly terms published in 1881 as “Dictionary of Philosophy”; 1,099 out of 1,847 words survived and are found in a dictionary of the present Japanese “Sanseido Kokugo Jiten”, from which we take the number of meanings of the 1,099 words. We may also study the relation of the polysemy to the number of words having the given weight. The elementary numbers are presented in the fourth column of Table 1. Here, we can see a different trend. The sequence is concave – there are few simple words, and there are few very complex words. Even this relation may be captured by a function. In the last years, one has been using the Lorentzian function defined as

$$(3) \quad y = \frac{a}{1 + \left(\frac{x - b}{c}\right)^2},$$

which can be derived from the same general theory. The results are presented in Table 4.

Table 4
Polysemy and number of words with the given polysemy

Complexity	Number of words	Lorentzian function
3	2	9.50
5	2	12.09
6	4	13.77
7	8	15.79

Word Length and Polysemy in Japanese

8	11	18.24
9	14	21.25
10	9	24.95
11	31	29.51
12	51	35.13
13	49	41.98
14	55	50.12
15	64	59.32
16	71	68.77
17	83	76.91
18	68	81.71
18	82	81.71
20	73	76.69
21	58	68.47
22	70	59.01
23	58	49.84
24	40	41.74
25	47	34.93
26	40	29.34
27	24	24.81
28	20	21.14
29	15	18.16
30	15	15.72
31	12	13.71
32	6	12.04
33	5	10.64
34	2	9.47
35	5	8.47
36	1	7.62
37	1	6.88
38	1	6.25
41	1	4.76
42	1	4.41
a = 82.3240, b = 18.4837, c = -5.593843, R ² = 0.9253		

As can be seen, Japanese is no exception, and some synergetic laws hold true here, too.

References

- Köhler, R. (2005). Synergetic Linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistics. An International Handbook: 760–784*. Berlin / New York: de Gruyter.
- Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified Modeling of Length in Language*. Lüdenscheid: RAM-Verlag (= Studies in Quantitative Linguistics Vol. 16).
- Sanada, H. (2008). *Investigations in Japanese Historical Lexicology* (rev. ed.). Göttingen: Peust & Gutschmidt Verlag.

Belza Chains in Machar's *Letní sonety*

Michal Místecký¹

Abstract. Belza chains, a special method for studying the association or semantic concentration of the lines of Czech sonnets written by the Czech poet Machar are analyzed. It is shown that the classical relations known from quantitative linguistics are valid.

Keywords: Czech, Sonnet, Machar, Belza chains

1. Introduction

A Belza-chain is formed by subsequent sentences containing a common concept. Usually, one studies them in scientific texts, in which the subsequent sentences usually form such a sequence. Nevertheless, they can also be used in poetic texts, in which the poet has the freedom of choice. No line needs to be even logically associated with the previous one. But if it is, one may do it by repeating the same word, or by presenting a word in two different parts-of-speech – e.g., in Czech *lehký – lehce* (“easy” – “easily”), or by some relative pronouns or particles like *which, who, this, that, he*, etc., or – in an inflecting language – simply by a respective verbal person, e.g., *já pracuji, on pracuje* (“I work” – “he works”). Languages have various means to show that the same concepts are concerned.

In “normal” texts composed of sentences demarcated by some signs – e.g., dots, semicolons, question marks, etc. –, it is easy to see the end of a sentence. In spoken language, it is quite different. And in poetry, there is only one unit which is clearly separated from the next one, namely the line (verse). Two or more lines form a Belza-chain if they contain a common semantic, lexical, or grammatical element. In some languages – especially in strongly analytic ones –, grammar can usually be ignored, but in strongly synthetic ones, like Czech, it must be taken into consideration.

Belza chains show the conceptual coherence of the text. It offers a quantitative scope for the issue of text structure, which has been researched for years (cf. Halliday, Hasan, 1976; Van Dijk, 1977; Dressler, Beaugrande, 1981). The poet has an absolute freedom, but even a lyrical poem has an “object”. There are many ways to study the conceptual coherence/concentration of a poem; here, the research will focus on Belza chains. In general, it holds that the longer the chains are, the stronger the thematic concentration is. The measurement can be performed in various ways, and the problem will be shown from several views.

First of all, one is aware of the fact that a Belza chain may (a) contain another Belza-chain, (b) a new Belza-chain can begin within another chain, which is the same as the ending of a chain within the other chain. The same concept may be repeated in the poem, but if it does not occur in subsequent lines, it does not form a chain. The Belza-structure of a poem can be very rich, and it is not even possible to compute the maximal number of chains because the lines need not have the same length. The minimal number of chains is 14, which is equal to the number of lines. In such a case, all lines are lexically, semantically and gram-

¹ Univ. Ostrava (Czech Republic); mail: MMistecky@seznam.cz

matically different. But this is rather an exception. Studying 47 Czech sonnets by Machar, such a structure was not found. The maximal number of chains – 13 – was found in *Sonet večerní* (“An Evening Sonnet”), the topic of which is formed by a discontinuous set of genre images.

Table 1
Belza chains in Machar's *Letní sonety*

Sonnet	Vector	Number of Chains	Sum of Chain Lengths
E. Zolovi	8,1,1,1,2,1	6	14
Matce	2,2,2,5,1,4	6	16
Sonet cynický	1,1,2,2,1,1,3,1,1,1	10	14
Sonet de vanitate	3,1,1,1,2,3,2,1,2,2	10	18
Sonet elegický	2,1,1,1,1,2,2,3,2	9	15
Sonet ironický	2,1,1,3,2,1,2,1,3	9	16
Sonet k sociální otázce	1,1,1,1,1,1,2,1,2,2,2,1	12	16
Sonet k teorii; Boj o život	5,2,2,4,2,2	6	17
Sonet materialistický	2,2,4,1,4,1	6	14
Sonet mystický	2,2,1,1,1,1,5,1	8	14
Sonet na Chopinovu melodii	2,2,2,1,1,2,3,1,2,2,2	11	20
Sonet na sentenci z Goetha	2,4,2,1,1,1,2,2,3	9	16
Sonet na sklonku století	2,2,2,3,1,1,1,1,2	9	15
Sonet nad verši z mládí	1,2,2,3,3,3,2,2,2	9	20
Sonet noční	2,1,1,2,2,1,1,1,1,2	10	14
Sonet o antice a vlasech	2,3,2,2,2,1,1,1,2,2	10	18
Sonet o bídě	1,2,5,2,1,2,2,2	8	17
Sonet o hodinách	2,2,2,1,3,2,1,5	8	18
Sonet o lásce	1,1,1,1,1,1,1,1,1,3,1,1	12	14
Sonet o minulosti	1,1,2,2,5,1,1,1,1	9	15
Sonet o Panně Marii	1,1,2,2,2,3,2,2,3	9	18
Sonet o rokoku	1,3,2,2,2,2,3,1,2	9	18
Sonet o staré metafoře	2,2,2,3,1,3,1,2	8	16
Sonet o starém líci a rubu	2,8,3,3,1,1,3,3	8	24
Sonet o třech metaforách	1,1,1,1,2,2,2,1,1,1,1	11	14
Sonet o třetí hodině v červenci	2,2,1,2,1,1,4,2,1	9	16
Sonet o vídeňských kosech	3,5,1,4,2,2,1	7	18
Sonet o západu slunce	3,2,1,1,2,1,2,2	8	14
Sonet o zlatém věku naší poezie	1,1,1,1,2,1,1,2,1,3	10	14
Sonet o životě	2,8,2,2,2	5	16
Sonet patologický	1,1,1,1,2,4,1,1,1,1,1,1	12	16
Sonet polední	1,1,1,1,1,2,1,1,2,2,1	11	14
Sonet sarkastický	3,1,3,4,2,1,2,1	8	17
Sonet svatební	2,12,3	3	17
Sonet úvodní	2,1,1,2,2,1,2,2,2,2,1	11	18

Sonet večerní	2,1,1,1,1,1,1,1,1,1,1, 1	13	14
Sonet z dvacátého září	1,1,2,2,2,2,1,1,1,3	10	16
Sonet-apostrofa	1,3,2,1,1,1,3,2,1	9	15
Sonet-epilog čtenáři	1,3,1,1,1,1,1,2,1,2	10	14
Sonet-intermezzo ₂	2,6,2,2,2	5	14
Sonet-intermezzo	1,1,2,11	4	15
Sonety-Causerie I.	3,1,4,2,2,2,2,1	8	15
Sonety-Causerie II.	2,1,1,2,2,9,2,2	8	21
Sonety-Causerie III.	1,4,2,2,1,1,1,1,1,1	10	15
Sonety-Causerie IV.	2,4,2,1,1,1,1,3,1,2	10	18
Sonety-Causerie V.	1,1,1,3,1,1,2,2,1,1	10	14
Své ženě s předešlým sonetem	3,6,2,3,3,1	6	18

2. Distributions and Association

As is usual, one tries to capture the data by a function. However, 47 sonnets, of the chain lengths between 3 and 13, do not yield a satisfactory picture. Nevertheless, there was an endeavour to find a preliminary model. The data are presented in Table 1; here, only the length 7 is “anormal”. In order to smooth the data, the sum of the neighbouring length and the mean of the lengths are taken into calculations. In this way, one obtains the results in the second half of the table. This sequence can be captured by the Lorentzian function used frequently in linguistics, namely

$$(1) \quad y = \frac{a}{1 + \left(\frac{x-b}{c}\right)^2}.$$

Table 2
Distribution of Belza chains in *Letní sonety* by Machar

Number of Chains	Frequency	Averaged	Frequency	Lorentzian Ft.
3	1	3.5	2	1.79
4	1	5.5	7	3.86
5	2	7.5	10	11.06
6	5	9.5	20	19.73
7	1	11.5	7	7.14
8	9	13.5	1	2.79
9	11			
10	9			
11	4			
12	3			
13	1			
		a = 20.7112, b = 9.1150, c = 1.7296, R ² = 0.9390		

If the sums of chain lengths are considered, one is offered, again, with numbers between 14 and 24, and their frequency could be modelled. However, one sees a possibility of modelling in the length in relation to the number of chains, too. For the sake of simplicity and in order to get some relative number, the sum of the chain lengths is divided by their number and the measure of association is obtained (see Table 3). This index suggests a rate of inter-connectedness of topics in a given poem.

Table 3
Associativity in Machar's *Letní sonety*

Sonnet	Association
E. Zolovi	2.33
Matce	2.67
Sonet cynický	1.40
Sonet de vanitate	1.80
Sonet elegický	1.67
Sonet ironický	1.78
Sonet k sociální otázce	1.33
Sonet k teorii; Boj o život	2.83
Sonet materialistický	2.33
Sonet mystický	1.75
Sonet na Chopinovu melodii	1.82
Sonet na sentenci z Goetha	1.78
Sonet na sklonku století	1.67
Sonet nad verši z mládí	2.22
Sonet noční	1.40
Sonet o antice a vlasech	1.80
Sonet o bídě	2.13
Sonet o hodinách	2.25
Sonet o lásce	1.17
Sonet o minulosti	1.67
Sonet o Panně Marii	2.00
Sonet o rokoku	2.00
Sonet o staré metafoře	2.00
Sonet o starém líci a rubu	3.00
Sonet o třech metaforách	1.27
Sonet o třetí hodině v červenci	1.78
Sonet o vídeňských kosech	2.57
Sonet o západu slunce	1.75
Sonet o zlatém věku naší poezie	1.40
Sonet o životě	3.20
Sonet patologický	1.33
Sonet polední	1.27
Sonet sarkastický	2.13
Sonet svatební	5.67
Sonet úvodní	1.64
Sonet večerní	1.08
Sonet z dvacátého září	1.60

Sonet-apostrofa	1.67
Sonet-epilog čtenáři	1.40
Sonet-intermezzo ₂	2.80
Sonet-intermezzo	3.75
Sonety-Causerie I.	1.88
Sonety-Causerie II.	2.63
Sonety-Causerie III.	1.50
Sonety-Causerie IV.	1.80
Sonety-Causerie V.	1.40
Své ženě s předešlým sonetem	3.00

Now, one may endeavour to investigate the modelling of the association numbers. In order to get a better fit, means of intervals will be counted with; the value of “1.25” thus means the interval of <1; 1.5). Again, the Lorentzian function yields a satisfactory fit.

Table 4
Distribution of Associations

Association Interval Mean	Frequency	Lorentzian Ft.
1.25	11	11.12
1.75	19	18.87
2.25	7	7.79
2.75	5	3.16
3.25	1	1.62
3.75	1	0.97
5.25	1	0.34
a = 19.4774, b = 1.6643, c = 0.4780, R ² = 0.9820		

3. Motif Length of Belza Chains

The chains can be written in their original form, e.g. in the sonnet *E. Zolovi* (“To *É. Zola*”) by Machar, one finds the sequence [8,1,1,1,2,1]. Using some generalizations (cf. Liu, Liang 2017), one can construct more abstract units, namely motifs. Quantitative motifs are constructed in the form of non-decreasing numbers; in the above case, the division gives [8]; [1,1,1,2]; [1]. Now, quantitative motifs have a sum (here, 8, 5, 1), a length (here 1, 4, 1), and a range, which is computed as a difference between the greatest and the smallest number (here 0, 1, 0). The list of the motifs and their lengths is presented in Table 5.

Table 5
Motifs and their lengths in Machar’s *Letní sonety*

Sonnet	Chain Motifs	Chain Motifs Length
E. Zolovi	[8]; [1,1,1,2]; [1]	1,4,1
Matce	[2,2,2]; [5]; [1,4]	3,1,2
Sonet cynický	[1,1,2,2]; [1,1,3]; [1,1,1]	4,3,3
Sonet de vanitate	[3]; [1,1,1,2,3]; [2]; [1,2,2]	1,5,1,3
Sonet elegický	[2]; [1,1,1,1,2,2,3]; [2]	1,7,1

Belza Chains in Machar's Letní sonety

Sonet ironický	[2]; [1,1,3]; [2]; [1,2]; [1,3]	1,3,1,2,2
Sonet k sociální otázce	[1,1,1,1,1,2]; [1,2,2,2]; [1]	7,4,1
Sonet k teorii; Boj o život	[5]; [2,2,4]; [2,2]	1,3,2
Sonet materialistický	[2,2,4]; [1,4]; [1]	3,2,1
Sonet mystický	[2,2]; [1,1,1,1,5]; [1]	2,5,1
Sonet na Chopinovu melodii	[2,2,2]; [1,1,2,3]; [1,2,2,2]	3,4,4
Sonet na sentenci z Goetha	[2,4]; [2]; [1,1,1,2,2,3]	2,1,6
Sonet na sklonku století	[2,2,2,3]; [1,1,1,1,2]	4,5
Sonet nad verši z mládí	[1,2,2,3,3,3]; [2,2,2]	6,3
Sonet noční	[2]; [1,1,2,2]; [1,1,1,1,2]	1,4,5
Sonet o antice a vlasech	[2,3]; [2,2,2]; [1,1,1,2,2]	2,3,5
Sonet o bídě	[1,2,5]; [2,1]; [2,2,2]	3,2,3
Sonet o hodinách	[2,2,2]; [1,3]; [2]; [1,5]	3,2,1,2
Sonet o lásce	[1,1,1,1,1,1,1,1,3]; [1,1]	10,2
Sonet o minulosti	[1,1,2,2,5]; [1,1,1,1]	5,4
Sonet o Panně Marii	[1,1,2,2,2,3]; [2,2,3]	5,3
Sonet o rokoku	[1,3]; [2,2,2,2,3]; [1,2]	2,5,2
Sonet o staré metafoře	[2,2,2,3]; [1,3]; [1,2]	4,2,2
Sonet o starém líci a rubu	[2,8]; [3,3]; [1,1,3,3]	2,2,4
Sonet o třech metaforách	[1,1,1,1,2,2,2]; [1,1,1,1]	7,4
Sonet o třetí hodině v červenci	[2,2]; [1,2]; [1,1,4]; [2]; [1]	2,2,3,1,1
Sonet o vídeňských kosech	[3,5]; [1,4]; [2,2]; [1]	2,2,2,1
Sonet o západu slunce	[3]; [2]; [1,1,2]; [1,2,2]	1,1,3,3
Sonet o zlatém věku naší poezie	[1,1,1,1,2]; [1,1,2]; [1,3]	5,3,2
Sonet o životě	[2,8]; [2,2,2]	2,3
Sonet patologický	[1,1,1,1,2,4]; [1,1,1,1,1,1]	6,6
Sonet polední	[1,1,1,1,1,2]; [1,1,2,2]; [1]	6,4,1
Sonet sarkastický	[3]; [1,3,4]; [2]; [1]; [2]; [1]	1,3,1,1,1,1
Sonet svatební	[2,12]; [3]	2,1
Sonet úvodní	[2]; [1,1,2,2]; [1,2,2,2,2]; [1]	1,3,5,1
Sonet večerní	[2]; [1,1,1,1,1,1,1,1,1,1,1]	1,12
Sonet z dvacátého září	[1,1,2,2,2,2]; [1,1,1,3]	5,4
Sonet-apostrofa	[1,3]; [2]; [1,1,1,3]; [2]; [1]	2,1,4,1,1
Sonet-epilog čtenáři	[1,3]; [1,1,1,1,1,2]; [1,2]	2,6,2
Sonet-intermezzo(2)	[2,6]; [2,2,2]	2,3
Sonet-intermezzo	[1,1,2,11]	1
Sonety-Causerie I.	[3]; [1,4]; [2,2,2,2]; [1]	1,2,4,1
Sonety-Causerie II.	[2]; [1,1,2,2]; [9]; [2,2]	1,4,1,2
Sonety-Causerie III.	[1,4]; [2,2]; [1,1,1,1,1,1]	2,2,6
Sonety-Causerie IV.	[2,4]; [2]; [1,1,1,1,3]; [1,2]	2,1,5,2
Sonety-Causerie V.	[1,1,1,3]; [1,1,2,2]; [1,1]	4,4,2
Své ženě s předešlým sonetem	[3,6]; [2,3,3]; [1]	2,3,1

Here, merely the motif length of the complete collection will be studied. The obtained results are listed in Table 6. Here, the Lorentzian function would yield excellent results but an effort is made to present the models in the simplest possible way. Here, the simple exponential function will be applied. Since the zeroes are omitted, the formula reads

$$(2) \quad y = 1 + a^{-bx} .$$

Table 6
Motif lengths of Belza chains in *Letni sonety* by Machar

Lengths of Belza Chains	Frequency	Exponential Ft.
1	44	46.97
2	38	32.96
3	23	23.21
4	17	16.44
5	12	11.73
6	7	8.46
7	3	6.19
10	1	2.74
12	1	1.84
a = 66.1347, b = 0.3636, R ² = 0.9751		

From the interpretational point of view, it is still an open question what the motifs actually mean as to the semantics of a poem. In general, one can say that motifs are higher-level units created conventionally, just as morphemes or syllables; their placing into the Köhlerian control cycle is a task for future. In this case, the rising number of chains indicates a growing or stagnating tendency to complicate the structure of the poem by means of associations; the sutures between the motifs then point at places where this trend is breached, and the poem's association plummets. These reflections may be made use of in literary criticism.

4. Climax Sequences

In order to make the concept of the motif useful for a less formal approach to texts, the present research has tried to sketch a new unit on their basis, namely climax sequences – i.e., in difference to Köhler's quantitative motifs, which are non-decreasing, we try to establish purely increasing sequences and call them climax sequences. Climax sequences (CS) are founded upon the idea that the gradation of topics is carried out via a rise in Belza chains as the poem or any text increases / progresses; this means that a sequence contains only increasing lengths of Belza chains, and once the length remains the same or drops, a new climax-sequence starts. To exemplify it, let us have a look at the aforementioned sonnet *E. Zolovi*: here, the motif parsing yields the sequence [8]; [1,1,1,2]; [1], whereas the climax-sequences partitioning is [8]; [1]; [1]; [1,2]; [1]. It is logical that the lower the number of climax sequences is in a poem, the more climax-oriented it is to be considered. As to the general counts, all possibilities which are open for motifs are available for climax sequences as well.

Table 7 shows the results and what may be counted out of them. First, the CS averages are presented, the formula of which is

$$(3) \quad \phi_{cs} = \frac{\text{Number of CS}}{\text{Number of Belza chains}} ;$$

they may indicate the measure of a poem's inclination to gradation: the lower the value is, the more it tends to be end-focused. On the other hand, if the numbers are high, it demonstrates that there are many local climaxes throughout the poem, its structure thus being multi-faceted. This turns the CS into a usable stylistic indicator. Second, the general average of the CS numbers may be calculated, the value of which it is possible to compare with other sonnet collections. Here, the count yields 6.70.

Table 7
Climax sequences and their averages in Machar's *Letní sonety*

Sonnet	Climax Sequences	Number of Belza Chains	Number of Climax Sequences	Average CS Length
E. Zolovi	[8]; [1]; [1]; [1,2]; [1]	6	5	1.20
Matce	[2]; [2]; [2,5]; [1,4]	6	4	1.50
Sonet cynický	[1]; [1,2]; [2]; [1]; [1,3]; [1]; [1]; [1]	10	8	1.25
Sonet de vanitate	[3]; [1]; [1]; [1,2,3]; [2]; [1,2]; [2]	10	7	1.43
Sonet elegický	[2]; [1]; [1]; [1]; [1,2]; [2,3]; [2]	9	7	1.29
Sonet ironický	[2]; [1]; [1,3]; [2]; [1,2]; [1,3]	9	6	1.50
Sonet k sociální otázce	[1]; [1]; [1]; [1]; [1]; [1,2]; [1,2]; [2]; [2]; [1]	12	10	1.20
Sonet k teorii; Boj o život	[5]; [2]; [2,4]; [2]; [2]	6	5	1.20
Sonet materialistický	[2]; [2,4]; [1,4]; [1]	6	4	1.50
Sonet mystický	[2]; [2]; [1]; [1]; [1]; [1,5]; [1]	8	7	1.14
Sonet na Chopinovu melodii	[2]; [2]; [2]; [1]; [1,2,3]; [1,2]; [2]; [2]	11	8	1.38
Sonet na sentenci z Goetha	[2,4]; [2]; [1]; [1]; [1,2]; [2,3]	9	6	1.50
Sonet na sklonku století	[2]; [2]; [2,3]; [1]; [1]; [1]; [1,2]	9	7	1.29
Sonet nad verši z mládí	[1,2]; [2,3]; [3]; [3]; [2]; [2]; [2]	9	7	1.29
Sonet noční	[2]; [1]; [1,2]; [2]; [1]; [1]; [1]; [1,2]	10	8	1.25
Sonet o antice a vlasech	[2,3]; [2]; [2]; [2]; [1]; [1]; [1,2]; [2]	10	8	1.25
Sonet o bídě	[1,2,5]; [2]; [1,2]; [2]; [2]	8	5	1.60

Sonet o hodinách	[2]; [2]; [2]; [1,3]; [2]; [1,5]	8	6	1.33
Sonet o lásce	[1]; [1]; [1]; [1]; [1]; [1]; [1]; [1,3]; [1]; [1]	12	10	1.20
Sonet o minulosti	[1]; [1,2]; [2,5]; [1]; [1]; [1]; [1]	9	7	1.29
Sonet o Panně Marii	[1]; [1,2]; [2]; [2,3]; [2]; [2,3]	9	6	1.50
Sonet o rokoku	[1,3]; [2]; [2]; [2]; [2,3]; [1,2]	9	6	1.50
Sonet o staré metafoře	[2]; [2]; [2,3]; [1,3]; [1,2]	8	5	1.60
Sonet o starém líci a rubu	[2,8]; [3]; [3]; [1]; [1,3]; [3]	8	6	1.33
Sonet o třech metaforách	[1]; [1]; [1]; [1,2]; [2]; [2]; [1]; [1]; [1]; [1]	11	10	1.10
Sonet o třetí hodině v červenci	[2]; [2]; [1,2]; [1]; [1,4]; [2]; [1]	9	7	1.29
Sonet o vídeňských kosech	[3,5]; [1,4]; [2]; [2]; [1]	7	5	1.40
Sonet o západu slunce	[3]; [2]; [1]; [1,2]; [1,2]; [2]	8	6	1.33
Sonet o zlatém věku naší poezie	[1]; [1]; [1]; [1,2]; [1]; [1,2]; [1,3]	10	7	1.43
Sonet o životě	[2,8]; [2]; [2]; [2]	5	4	1.25
Sonet patologický	[1]; [1]; [1]; [1,2,4]; [1]; [1]; [1]; [1]; [1]; [1]	12	10	1.20
Sonet polední	[1]; [1]; [1]; [1]; [1,2]; [1]; [1,2]; [2]; [1]	11	9	1.22
Sonet sarkastický	[3]; [1,3,4]; [2]; [1,2]; [1]	8	5	1.60
Sonet svatební	[2,12]; [3]	3	2	1.50
Sonet úvodní	[2]; [1]; [1,2]; [2]; [1,2]; [2]; [2]; [2]; [1]	11	9	1.22
Sonet večerní	[2]; [1]; [1]; [1]; [1]; [1]; [1]; [1]; [1]; [1]; [1]; [1]; [1]	13	13	1.00
Sonet z dvacátého září	[1]; [1,2]; [2]; [2]; [2]; [1]; [1]; [1,3]	10	8	1.25
Sonet-apostrofa	[1,3]; [2]; [1]; [1]; [1,3]; [2]; [1]	9	7	1.29
Sonet-epilog čtenáři	[1,3]; [1]; [1]; [1]; [1]; [1,2]; [1,2]	10	7	1.43
Sonet-intermezzo(2)	[2,6]; [2]; [2]; [2]	5	4	1.25

Sonet-intermezzo	[1]; [1,2,11]	4	2	2.00
Sonety-Causerie I.	[3]; [1,4]; [2]; [2]; [2]; [2]; [1]	8	7	1.14
Sonety-Causerie II.	[2]; [1]; [1,2]; [2,9]; [2]; [2]	8	6	1.33
Sonety-Causerie III.	[1,4]; [2]; [2]; [1]; [1]; [1]; [1]; [1]; [1]	10	9	1.11
Sonety-Causerie IV.	[2,4]; [2]; [1]; [1]; [1]; [1,3]; [1,2]	10	7	1.43
Sonety-Causerie V.	[1]; [1]; [1,3]; [1]; [1,2]; [2]; [2]; [1]; [1]	10	9	1.11
Své ženě s předešlým sonetem	[3,6]; [2,3]; [3]; [1]	6	4	1.50

As to the rank-frequency distribution of climax sequences, it is possible to model it according to several functions, the best fit of which is provided by the Zipf-Alekseev one. Its formula reads

$$(4) \quad y = 1 + cx^{a+b \cdot \ln(x)} ;$$

the precision of the fit, which is shown in Table 8, is also attested by the high value of determination coefficient (0.9923).

Table 8
Rank-frequency distribution of CS as modelled by the Zipf-Alekseev function

Rank	Type of CS	Frequency	Zipf-Alekseev +1
1	[1]	126	126.07322
2	[2]	87	86.44854
3	[1,3]	15	18.67291
4	[3]	11	4.15266
5	[2,3]	10	1.58564
6	[1,4]	4	1.11825
7	[2,4]	4	1.02612
8	[1,2,11]	3	1.00628
9	[2,5]	2	1.00163
10	[1,2,3]	2	1.00046
11	[1,4]	2	1.00014
12	[1,5]	2	1.00004
13	[2,8]	2	1.00001
14	[8]	1	1.00000
15	[5]	1	1.00000
16	[1,2,5]	1	1.00000
17	[3,5]	1	1.00000
18	[1,2,4]	1	1.00000
19	[1,3,4]	1	1.00000
20	[2,12]	1	1.00000

21	[2,6]	1	1.00000
22	[2,9]	1	1.00000
23	[3,6]	1	1.00000
a = 1.5557, b = -3.0374, c = 125.0732			

Moreover, the length of CS may be modelled, too. Due to a low number of ranks – there are one-chain, two-chain, or three-chain CS only –, a good fit is provided by the casual Zipf power function. The formula

$$(5) \quad y = ax^{-b}$$

yields the results with the determination coefficient of 0.9972. The details are given in Table 9.

Table 9
Rank-frequency distribution of CS lengths as modelled by the Zipf function

Rank	Length of CS	Frequency	Zipf function
1	1	226	226.29772
2	2	46	41.61313
3	3	8	15.45325

The lengths of CS thus demonstrate a synergetic tendency (cf. Köhler 2005): the longer they are, the less frequent they get. This may probably be caused by difficulties linked to holding a steady pace in developing the gradation for a long time.

The definitions and computations presented above are one of the first steps into a quantitative analysis of sonnets. This kind of poetry has the advantage of being of, *mutatis mutandis*, the same structure in all languages, but the disadvantage of being too short. Nevertheless, the results won quantitatively are representative, and can be tested easily.

References

- De Beaugrande, R. – Dressler, W. U. (1981). *Introduction to Text Linguistics*. London / New York: Longman.
- Halliday, M. A. K. – Hasan, R. (1976). *Cohesion in English*. London: Routledge.
- Köhler, R. (2005). „Synergetic Linguistics.“ In: Köhler, R. – Altmann, G. – Piotrowski, R. G. (eds.), *Quantitative Linguistics. An International Handbook*. Berlin/New York: Walter de Gruyter, pp. 760–775.
- Liu, H., Liang, J. (eds.) (2017). *Motifs in Language and Text*. Berlin/Boston: de Gruyter Mouton.
- Van Dijk, T. (1977). *Text and Context*. London: Longman.

Quantifying Comprehensibility of Christmas and Easter Addresses from the Ukrainian Greek Catholic Church Hierarchs

Andrij Rovenchak¹, Olha Rovenchak²

Abstract. We propose a set of parameters suitable for quantification of text comprehensibility. From the analysis of 36 texts based on word frequencies we found that lower values of entropy correspond to better comprehensibility, while the same is observed for higher values of the fraction of dis legomena and the scaling factor between the text length and the h -point squared. Our observations based on relative values of parameters will be useful to define a numerical measure for text comprehensibility.

Keywords: Ukrainian, text comprehensibility, dis legomena, repeat rate, h -point, entropy, type-token, hapax legomena

1. Introduction

Texts are different in various aspects. Some of them are easily measurable, like length or vocabulary size, some other – like mean syllable length or sentence length – require a little more effort, mostly linked with definitions of linguistic units. But there are certainly much more hidden properties of texts, for which no simple approach is readily available. Being to a certain extent subjective, i. e., reader- or listener-related, such properties still can be attempted to get measured. For instance, the complexity of discourse was studied by Kockelman (2009) and thematic concentration of text was analyzed by Čech et al. (2015). The property we are going to discuss in the present paper is text comprehensibility. It is related to text difficulty or text readability (Mikk 1995; Mikk & Elts 1999; Pires et al. 2017).

Obviously, a belletristic short story is perceived easily comparing to a scientific paper or a diplomatic agreement. Similarly, manuals for first-graders are easier to read than those for fourth-graders. The things become more complicated, however, when one tries to classify in this respect texts of one type. It was shown in particular that sermons, being the closest genre to the subject of this study, are quite homogeneous with respect to several parameters (Buk et al 2010; Rovenchak & Buk 2011). We are going to analyze addresses from church hierarchs which are typically announced during liturgies on major Christian Holidays, namely, Easter and Christmas. Based both on authors' personal impression and on feedbacks from numerous people we can state that comprehensibility of such addresses differs depending on their authors. We will try to quantify these differences.

¹ Mail: andrij.rovenchak@gmail.com

Department for Theoretical Physics, Ivan Franko National University of Lviv

² Mail: rowentschak@gmail.com

Department for Sociology, Ivan Franko National University of Lviv

The paper is organized as follows. Section 2 contains some background information and description of texts we analyze. Results are presented in Section 3 and also summarized in Table 1 in the Appendix. Brief discussion is given in Section 4.

2. Material overview

To put the Readers into a wider context, brief historical overview and biographical notes are given below. Ukrainian Greek Catholic Church (UGCC), or more officially Ukrainian Catholic Church of the Byzantine Tradition, was founded in the late 16th century as a result of the conclusion of the Brest Church Union in 1596 and the transition to the jurisdiction of the Pope the vast majority of Bishops and Metropolitan of Kyiv Metropolitanate. The activity of UGCC was systematically prohibited on the territories of Ukraine under Russian and later Soviet rule in 17–20th centuries. During 1946–1989 it acted as an underground church in the Soviet Union. This denomination is the third highest in Ukraine by the number of parishes with the majority in its Western part (Rovenchak 2008: 140–147). The Church is ruled by Supreme Archbishop of Kyiv-Halyč[yna], Metropolitan of Kyiv. Present Head of UGCC is Sviatoslav Shevchuk. The UGCC headquarters were moved from Lviv to Kyiv in 2005 (UGCC 2004–2018). Due to wide geography of Ukrainian emigrants settling and great significance they attach to their church life, UGCC has its departments in Western Europe, Americas, and Australia and serves there as both religious and social institute (Rovenchak & Volodko 2015).

The addresses from the following three hierarchs were included in the analysis: Lubomyr *Cardinal* Husar, Sviatoslav Shevchuk, and Ihor Vozniak. The biographical data are compiled from several sources (Catholic Pages 1996–2007; Cheney 1996–2017; Dziuba et al. 2001–2017).

Lubomyr *Cardinal* Husar M.S.U (Liubomyr Huzar, Любомир Гузар; *26.02.1933–†31.05.2017). He was born in Lviv. In 1944 his family moved to Salzburg (Austria) and later in 1949 to Stamford, CT (USA). He was ordained a priest in 1958. In 1969–1972, Lubomyr Husar studied at the Pontifical Urbanian University in Rome obtaining his doctorate in theology. He was consecrated a bishop in 1977. Lubomyr Husar was named auxiliary bishop of Lviv of the Ukrainians by the Synod of Bishops of the Ukrainian Church in 1996. He was elected Major Archbishop of Lviv in 2001. Lubomyr Husar was elevated to Cardinal in 2001. He resigned his post of Major Archbishop of Kyiv in 2011.

Sviatoslav Shevchuk (Святослав Шевчук; *05.05.1970). He was born in Stryi, Lviv Oblast, Ukraine. In 1992–1994 he studied in Holy Spirit Ukrainian Catholic Seminary and received the priesthood in 1994. In 1999 he received the Doctorate with Summa cum laude in theological anthropology and moral theology from the Pontifical University of Thomas Aquinas. In 2002–2005, Sviatoslav Shevchuk was Head of the secretariat and personal secretary of Lubomyr Husar, also Head of the Patriarchal Curia in Lviv. Bishop's ordination took place in 2009. In 2011, the Electoral Synod of the Bishops of the UGCC elected Bishop Sviatoslav Shevchuk as the Supreme Archbishop of Kyiv-Halyč, Head of the Ukrainian Greek Catholic Church.

Ihor Vozniak, C.S.S.R. (Ihor Voznyak, Ігор Возняк; *03.08.1952). He was born in Lypytsi near Mykolayiv, Lviv Oblast. In 1973 he became a member of the Congregation of the Most Holy Redeemer. In the 1970s, Ihor Vozniak studied philosophy and theology at the underground seminary in Lviv. He professed perpetual vows in 1981. In 2005–2011 Ihor Vozniak was Archbishop of the Lviv Archeparchy. He was the Administrator of the UGCC from February 10

*Quantifying Comprehensibility of Christmas and Easter Addresses
from the Ukrainian Greek Catholic Church Hierarchs*

to March 27, 2011 and is Metropolitan of Lviv since 2011.



Sviatoslav Shevchuk



Ihor Vozniak



Lubomyr *Cardinal* Husar

Our set consists of 36 texts. These are all Christmas and Easter addresses we were able to collect online; they represent the time span from 2006 to 2018. Eleven texts were authored by Lubomyr *Cardinal* Husar, 11 by Ihor Vozniak, and 14 by Sviatoslav Shevchuk. For each text, a frequency list of wordforms was compiled yielding the rank–frequency distribution. Wordforms (or orthographic words) are understood as alphanumeric sequences between two spaces and/or punctuation marks. For simplicity, we refer to them as ‘words’.

Note that it is reasonable to remove the opening and final greetings from the addresses as they are standard formulas. For rather short texts like the analyzed ones keeping such items would generate unnecessary blurring in frequency data. So, we decided to remove the opening and final greetings from the analysis.

Now it is important to mention that addresses of Lubomyr *Cardinal* Husar are often recognized as being clearer hence better understood in comparison with those of Ihor Vozniak and Sviatoslav Shevchuk. In the next Section, we will try to find differences in some text parameters for these three authors and thus approach the problem of quantification of text comprehensibility.

3. Results

The idea of this study is in particular to find a parametrization of text comprehensibility based on simple indicators. That is why we do not consider, for instance, sentence length (cf. (Mikk 1995; Mikk & Elts 1999)). Eighteen parameters were calculated for each text. The complete set of data is summarized in Table 1 given in the Appendix. The parameters are as follows:

- number of words (tokens) N ;
- number of different words (types) V ;
- type-token ratio $TTR = V/N$;
- number of hapax legomena N_1 ;
- number of dis legomena (types occurring twice in a given sample) N_2 ;

- fraction of hapax legomena N_1/N ;
- fraction of dis legomena N_2/N ;
- relation of hapax and dis legomena N_1/N_2 ;
- entropy S

$$S = \sum_{i=1}^V p_i \ln p_i = \sum_{i=1}^V \frac{f_i}{N} \ln \frac{f_i}{N},$$

where f_i are absolute word frequencies and $p_i = f_i/N$ are probabilities (note the natural logarithm in the definition);

- mean word length in syllables m_1

$$m_1 = \frac{1}{N} \sum_{i=1}^N x_i,$$

where x_i is number of syllables per word, which is calculated in Ukrainian very simply, as the number of letters for vowels ⟨a, e, є, и, і, ї, o, y, ю, я⟩ (Mačutek & Rovenchak 2011);

- dispersion of word length in syllables (second central moment) m_2

$$m_2 = \frac{1}{N} \sum_{i=1}^N (x_i - m_1)^2;$$

- dispersion quotient d

$$d = \frac{m_2}{m_1 - 1};$$

- fraction of four-syllabic words p_4 ;
- fraction of five-syllabic words p_5 ;
- repeat rate R (Zörnig et al. 2016):

$$R = \sum_{i=1}^V p_i^2 = \frac{1}{N^2} \sum_{i=1}^V f_i^2;$$

- relative repeat rate R_{rel}

$$R_{\text{rel}} = \frac{1 - R}{1 - 1/V};$$

- h -point h (Popescu & Altmann 2008):

*Quantifying Comprehensibility of Christmas and Easter Addresses
from the Ukrainian Greek Catholic Church Hierarchs*

$$h = \begin{cases} r & \text{if there exists a solution for } r = f_r \\ \frac{f_1 r_2 - f_2 r_1}{r_2 - r_1 + f_1 - f_2} & \text{otherwise (here, } f_1 > r_1 \text{ and } f_2 < r_2 \text{)} \end{cases};$$

- index a being the scaling coefficient for the h -point (Popescu et al 2009):

$$a = \frac{N}{h^2}.$$

Some of these parameters are known to yield good genre and author attribution. In particular, the pair $(d; p_4)$ was used by Kelih et al. (2005) and parameters R_{rel} and a were among those applied for multivariate analysis by Zörnig et al. (2016).

In Fig. 1 we demonstrate some of the parameters depending on the authors. It appears that only a few of them are suitable for at least subtle author discrimination. Quite interestingly, it is Lubomyr Husar whose addresses stand out comparing to other two hierarchs, Ihor Vozniak and Sviatoslav Shevchuk. This result is what we expected to see in view of generally better comprehensibility of Lubomyr’s texts, as mentioned at the end of the previous Section.

The best separation is achieved for the pair of parameters N_2/N (fraction of dis legomena) and S (entropy), see Fig. 2. Another observation concerns the minimal value of the S parameter for Sviatoslav, corresponding to the Easter address in 2011. This is his first text in our set, just one month after he became the Head of the Church after Lubomyr. Such a situation could occur had both hierarchs used the same speechwriter. However, there are no other evidences for such claims according to our data. Moreover, even a superficial stylistic analysis suggests that all Sviatoslav’s texts are written by the same author. This is reflected in particular by the usage of quotations from the Bible.

It was expected that entropy S could be a parameter relevant to text comprehensibility. Such data were obtained for several genres of Ukrainian texts (Buk & Rovenchak 2004) with belles-lettres having the lowest value comparing to the highest ones in scientific and official styles. Entropy also proved to be a good discriminating parameter in frequency studies of nucleotide sequences (Rovenchak 2018).

Some observations regarding the parameter a are known. Being a scaling factor of the h -point it is considered independent (or weakly-dependent) on sample size. Across languages, smaller a s are considered as “a sign of analytism, i.e. the number of word forms is smaller” (Popescu et al. 2009: 23). On the other text, within one language this parameter, according to Zörnig et al. (2016), “can only be understood as an indicator for the different degree of the grammatical and morphosyntactical organisation of texts”. From our results we obtain that higher values of a correspond to better comprehensibility of texts.

The fraction of dis legomena N_2/N is a new parameter found in our work. As with the a parameter, higher values of a correspond to better comprehensibility of texts. Interestingly, hapax legomena – on the other hand – appear not to have any discriminating feature for the analyzed texts.

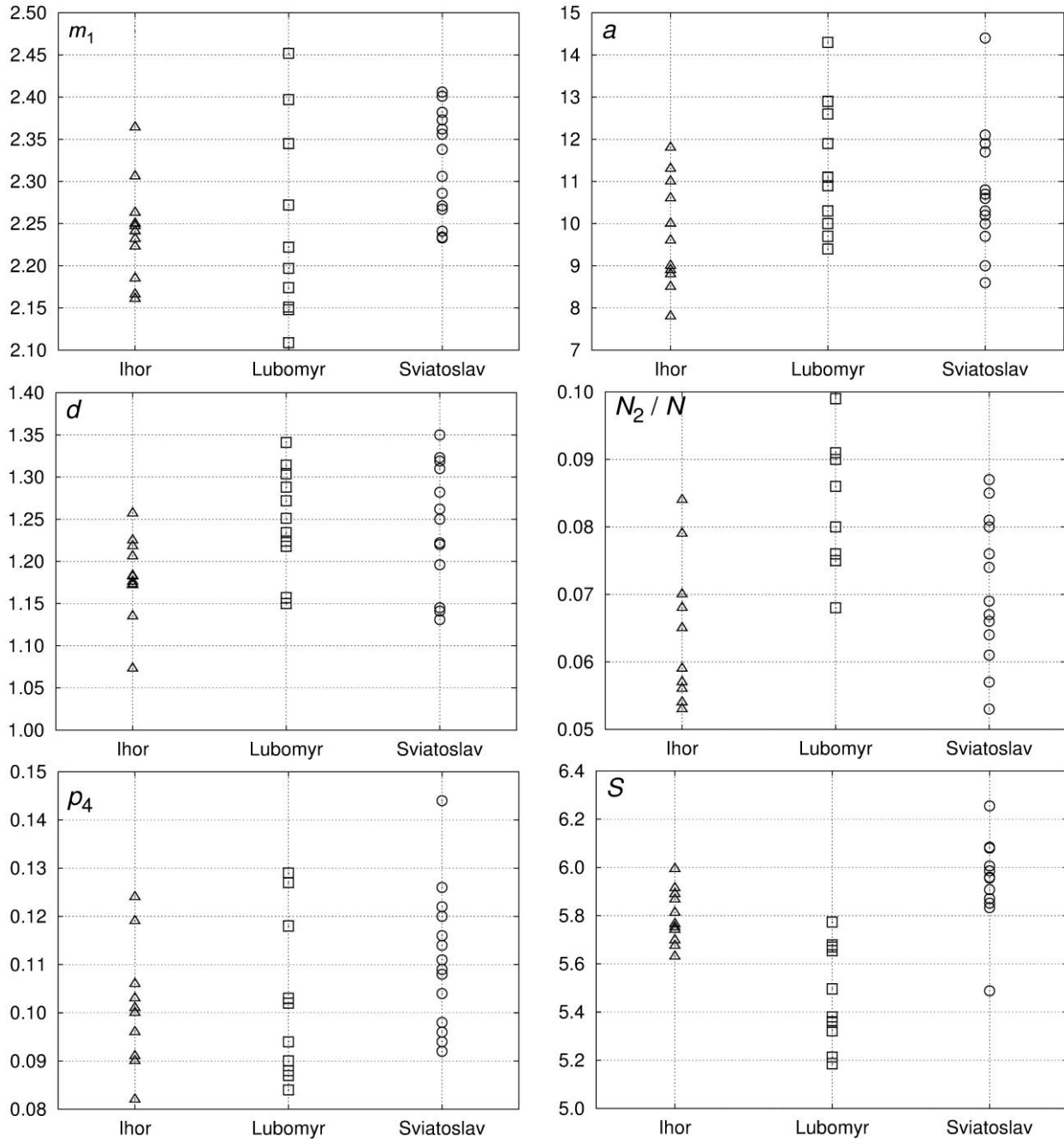


Fig. 1. Parameters exhibiting very little to none correlation (left column) and parameters with noticeable correlation with respect to authors (right column).

Texts are plotted in Fig. 2 on the $(S; N_2/N)$ plane. The left upper region thus corresponds to better text comprehensibility. In this domain, 7 of 11 Lubomyr’s addresses are concentrated. Both Sviatoslav’s and Ihor’s texts are located in right lower region. These texts appear hardly separable based on the values of the calculated S and N_2/N parameters – in fact, from the majority of other parameters as well, see Table 1 in the Appendix. Note that the mean entropy of Sviatoslav addresses is highest among the three authors: $\langle S \rangle = 5.46$ for Lubomyr, $\langle S \rangle = 5.79$ for Ihor, and $\langle S \rangle = 5.95$ for Sviatoslav.

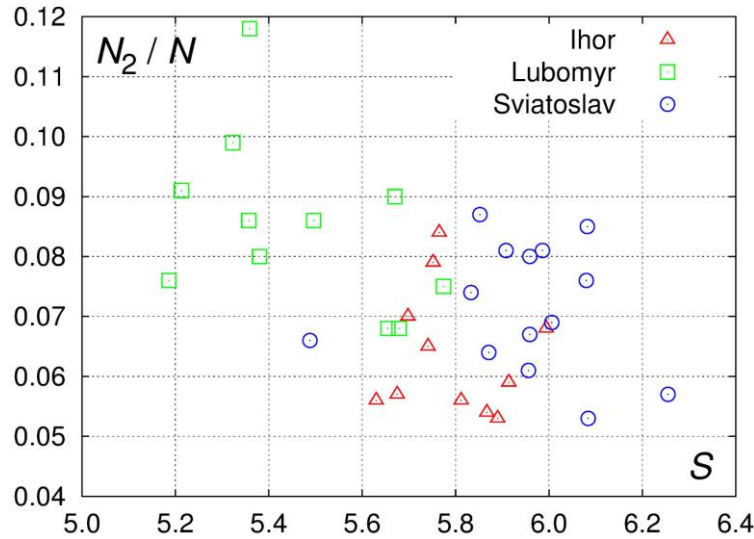


Fig. 2. Fraction of dis legomena (vertical axis) versus entropy (horizontal axis) for all 36 analyzed texts.

4. Discussion

From the conducted analysis we propose the following parameters as relative quantitative measures of text comprehensibility. The first one is entropy: the lowest entropy the better text comprehensibility. This result was expected on the basis of some previous studies and might be given a rather simple interpretation. The notion of entropy was coined in physics yet by Clausius (1865), later developed by Boltzmann and Gibbs (Müller & Müller 2009), and introduced into the information theory by Shannon (1948). Entropy is considered as a measure of disorder (Müller & Müller 2009: 139), so with respect to text comprehensibility smaller entropy might be treated as lower disorder hence less effort required to perceive the information from the text.

The second parameter is the fraction of dis legomena, which is a bit unexpected result lacking so far an interpretation: higher fraction corresponds to better comprehensibility. A similar statement concerns the third parameter, a , being the scaling factor in the h -point. Being related to the grammatical organization of text, these parameters might serve as measures of grammar simplicity.

Presently, we can only provide a relative measure for text comprehensibility. Further studies with a larger number of texts as well as an independent estimation of text comprehensibility are required to confirm our conclusions and to find an appropriate absolute definition of this property.

References

- Buk, S., Humenych, O., Mal'tseva, L. & Rovenchak, A.** (2010). "Word-length-related parameters of text genres in the Ukrainian language. A pilot study". In: *Text and Language: Structures - Functions - Interrelations. Quantitative perspectives*. Ed. by P. Grzybek, E. Kelih

- & J. Mačutek. Wien: Praesens, pp. 13–19.
- Buk, S. & Rovenchak, A.** (2004). “Rank-frequency analysis for functional style corpora of Ukrainian”. In: *Journal of Quantitative Linguistics* 11.3, pp. 161–171. DOI: 10.1080/0929617042000314912.
- Catholic Pages** (1996–2007). “Cardinal Husar”. Available online at: http://www.catholic-pages.com/hierarchy/cardinals_bio.asp?ref=192.
- Cheney, D. M.** (1996–2017). *The Hierarchy of the Catholic Church. Current and historical information about its bishops and dioceses*. In particular, see: <http://www.catholic-hierarchy.org/bishop/bhusar.html>; <http://www.catholic-hierarchy.org/bishop/bshevchuk.html>; <http://www.catholic-hierarchy.org/bishop/bvozniak.html>.
- Clausius, R.** (1865). “Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie”. In: *Annalen der Physik und Chemie* 125.7, pp. 353–400. DOI: 10.1002/andp.18652010702.
- Čech, R., Garabík, R. & Altmann, G.** (2015). “Testing the thematic concentration of text”. In: *Journal of Quantitative Linguistics* 22.3, pp. 215–232. DOI: 10.1080/09296174.2015.1037157.
- Dziuba, I. M., Žukovskyj, A. I., Romaniv, O. M., Železnjak, M. H. et al. (eds.)** (2001–2017). *Encyklopedija sučasnoji Ukrajinijy [Encyclopedia of Modern Ukraine]*. Kyiv: NASU Institute of Encyclopaedic Research. For the online version, see <http://esu.com.ua>.
- Kelih, E., Antić, G., Grzybek, P. & Stadlober, E.** (2005) “Classification of author and/or genre? The impact of word length”. In: *Classification – The Ubiquitous Challenge*. Ed. by C. Weihs & W. Gaul. Heidelberg: Springer, pp. 498–505.
- Kockelman, P.** (2009). “The complexity of discourse”. In: *Journal of Quantitative Linguistics* 16.1, pp. 1–39. DOI: 10.1080/09296170802514146.
- Mačutek, J. & Rovenchak, A.** (2011). “Canonical word forms: Menzerath–Altmann law, phonemic length and syllabic length”. In: *Studies in Quantitative Linguistics II: Issues in Quantitative Linguistics, Vol. 2*. Ed. by E. Kelih, V. Levickij, Yu. Matskulyak. Lüdenscheid: RAM-Verlag, pp. 136–147.
- Mikk, J.** (1995). “Methods for determining optimal readability of texts”. In: *Journal of Quantitative Linguistics* 2.2, pp. 125–132. DOI: 10.1080/09296179508590041.
- Mikk, J. & Elts, J.** (1999). “A reading comprehension formula of reader and text characteristics”. In: *Journal of Quantitative Linguistics* 6.3, pp. 214–221. DOI:10.1076/jqul.6.3.214.6158.
- Müller, I. & Müller W. H.** (2009). *Fundamentals of Thermodynamics and Applications: With Historical Annotations and Many Citations from Avogadro to Zermelo*. Berlin–Heidelberg: Springer.
- Pires, C., Cavaco, A. & Vigário, M.** (2017). “Towards the definition of linguistic metrics for evaluating text readability”. In: *Journal of Quantitative Linguistics* 24.4, pp. 319–349. DOI: 10.1080/09296174.2017.1311448.
- Popescu, I.-I. & Altmann, G.** (2008). “On the regularity of diversification in language”. In: *Glottometrics* 17, pp. 94–108.
- Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L. & Vidya, M. N.** (2009). *Word Frequency Studies*. Berlin–New York: Mouton de Gruyter.
- Rovenchak, A.** (2018). “Telling apart *Felidae* and *Ursidae* from the distribution of nucleotides in mitochondrial DNA”. In: *Modern Physics Letter B* 32.5, art. 1850057 [16 p.].
- Rovenchak, A. & Buk, S.** (2011). “Application of a quantum ensemble model to linguistic

*Quantifying Comprehensibility of Christmas and Easter Addresses
from the Ukrainian Greek Catholic Church Hierarchs*

analysis". In: *Physica A* 390.7, pp. 1326–1331. DOI: 10.1016/j.physa.2010.12.009.

Rovenchak, I. I. (2008). *Heohrafija kul'tury: problemy teoriji, metodolohiji ta metodyky doslidžennja [Geography of Culture: Problems of Theory, Methodology and Methods of Research]*. Lviv: Lviv University Press.

Rovenchak, O. & Volodko, V. (2015). *Mižnarodna migracija: teorija ta praktyka [International Migration: Theory and Practice]*. Lviv: Lviv University Press.

Shannon, C. E. (1948). "A mathematical theory of communication". In: *Bell System Technical Journal* 27.3, pp. 379–423. DOI:10.1002/j.1538-7305.1948.tb01338.x.

UGCC (2004–2018). *Official site of the Ukrainian Greek Catholic Church*. Available online at: <http://ugcc.ua>.

Zörnig, P., Kelih, E. & Fuks, L. (2016). "Classification of Serbian texts based on lexical characteristics and multivariate statistical analysis". *Glottology* 7.1, pp. 41–66. DOI: 10.1515/glot-2016-0004.

Appendix

Table 1. Parameters of texts. Easter or Christmas (Xmas) address are followed by a two-digit year mark corresponding to 2006–2018.

Type	Author	N	V	TTR	N_1	N_2	N_1/N	N_2/N	N_2/N_1	S	m_1	m_2	d	p_4	p_5	R	R_{rel}	h	a
Easter13	Ihor	539	367	0.681	304	30	0.564	0.056	10.13	5.630	2.364	1.463	1.073	0.124	0.035	0.0123	0.9904	7.0	11.0
Easter14	Ihor	725	452	0.623	362	47	0.499	0.065	7.70	5.741	2.185	1.400	1.182	0.090	0.022	0.0151	0.9871	9.0	9.0
Easter15	Ihor	678	432	0.637	334	57	0.493	0.084	5.86	5.765	2.161	1.365	1.176	0.091	0.027	0.0125	0.9898	8.0	10.6
Easter16	Ihor	720	464	0.644	378	40	0.525	0.056	9.45	5.812	2.263	1.538	1.218	0.106	0.038	0.0120	0.9902	9.0	8.9
Easter17	Ihor	869	540	0.621	436	51	0.502	0.059	8.55	5.914	2.247	1.528	1.225	0.119	0.020	0.0129	0.9890	9.5	9.6
Xmas13	Ihor	565	391	0.692	330	32	0.584	0.057	10.31	5.675	2.232	1.548	1.257	0.124	0.028	0.0118	0.9907	8.0	8.8
Xmas14	Ihor	764	495	0.648	406	41	0.531	0.054	9.90	5.867	2.241	1.468	1.183	0.103	0.039	0.0116	0.9904	9.5	8.5
Xmas15	Ihor	902	536	0.594	422	48	0.468	0.053	8.79	5.890	2.166	1.407	1.206	0.082	0.035	0.0143	0.9876	9.5	10.0
Xmas16	Ihor	957	600	0.627	484	65	0.506	0.068	7.45	5.994	2.306	1.531	1.172	0.101	0.032	0.0134	0.9883	9.0	11.8
Xmas17	Ihor	632	409	0.647	326	44	0.516	0.070	7.41	5.698	2.250	1.466	1.173	0.100	0.028	0.0132	0.9892	9.0	7.8
Xmas18	Ihor	721	433	0.601	323	57	0.448	0.079	5.67	5.752	2.223	1.388	1.135	0.096	0.029	0.0139	0.9884	8.0	11.3
Easter06	Lubomyr	501	324	0.647	249	43	0.497	0.086	5.79	5.496	2.222	1.554	1.272	0.084	0.042	0.0159	0.9871	6.5	11.9
Easter07	Lubomyr	340	223	0.656	169	26	0.497	0.076	6.50	5.186	2.197	1.605	1.341	0.103	0.032	0.0177	0.9867	6.0	9.4
Easter08	Lubomyr	589	385	0.654	304	40	0.516	0.068	7.60	5.679	2.452	1.816	1.251	0.129	0.048	0.0123	0.9903	6.8	12.9
Easter09	Lubomyr	622	386	0.621	294	42	0.473	0.068	7.00	5.655	2.174	1.542	1.314	0.088	0.043	0.0146	0.9880	7.5	11.1
Easter10	Lubomyr	360	263	0.731	217	31	0.603	0.086	7.00	5.357	2.397	1.617	1.157	0.156	0.036	0.0132	0.9906	5.5	11.9
Xmas06	Lubomyr	329	225	0.684	174	30	0.529	0.091	5.80	5.213	2.109	1.447	1.304	0.094	0.024	0.0157	0.9887	5.5	10.9
Xmas07	Lubomyr	357	255	0.714	197	42	0.552	0.118	4.69	5.359	2.345	1.638	1.218	0.118	0.050	0.0127	0.9912	5.0	14.3
Xmas08	Lubomyr	831	468	0.563	339	62	0.408	0.075	5.47	5.774	2.148	1.479	1.288	0.087	0.034	0.0169	0.9852	9.0	10.3
Xmas09	Lubomyr	477	293	0.614	218	38	0.457	0.080	5.74	5.380	2.151	1.323	1.150	0.090	0.027	0.0198	0.9836	7.0	9.7
Xmas10	Lubomyr	403	271	0.672	211	40	0.524	0.099	5.28	5.323	2.174	1.449	1.234	0.102	0.025	0.0185	0.9851	5.7	12.6
Xmas11	Lubomyr	589	379	0.643	289	53	0.491	0.090	5.45	5.670	2.272	1.556	1.224	0.127	0.041	0.0125	0.9902	7.7	10.0
Easter11	Sviatoslav	458	319	0.697	265	30	0.579	0.066	8.83	5.488	2.362	1.554	1.141	0.144	0.024	0.0142	0.9889	6.7	10.3
Easter12	Sviatoslav	772	481	0.623	365	67	0.473	0.087	5.45	5.852	2.356	1.553	1.145	0.122	0.028	0.0129	0.9891	8.0	12.1
Easter13	Sviatoslav	924	575	0.622	446	75	0.483	0.081	5.95	5.986	2.382	1.688	1.222	0.111	0.060	0.0133	0.9884	8.0	14.4
Easter14	Sviatoslav	1021	573	0.561	437	65	0.428	0.064	6.72	5.871	2.286	1.622	1.262	0.120	0.026	0.0214	0.9804	10.0	10.2
Easter15	Sviatoslav	951	577	0.607	449	64	0.472	0.067	7.02	5.959	2.338	1.769	1.323	0.126	0.040	0.0152	0.9866	9.0	11.7
Easter16	Sviatoslav	960	608	0.633	470	82	0.490	0.085	5.73	6.082	2.271	1.520	1.196	0.104	0.034	0.0102	0.9915	9.0	11.9
Easter17	Sviatoslav	1040	580	0.558	420	83	0.404	0.080	5.06	5.959	2.241	1.514	1.220	0.096	0.034	0.0160	0.9857	10.3	9.7
Xmas12	Sviatoslav	780	482	0.618	372	58	0.477	0.074	6.41	5.833	2.406	1.803	1.282	0.108	0.046	0.0138	0.9882	8.5	10.8
Xmas13	Sviatoslav	1309	787	0.601	617	74	0.471	0.057	8.34	6.255	2.373	1.811	1.319	0.092	0.036	0.0110	0.9903	11.0	10.8
Xmas14	Sviatoslav	900	526	0.584	385	73	0.428	0.081	5.27	5.908	2.401	1.585	1.131	0.109	0.044	0.0140	0.9879	10.0	9.0
Xmas15	Sviatoslav	1070	668	0.624	543	57	0.507	0.053	9.53	6.084	2.233	1.541	1.250	0.098	0.026	0.0130	0.9885	10.0	10.7
Xmas16	Sviatoslav	1170	640	0.547	489	71	0.418	0.061	6.89	5.956	2.234	1.667	1.350	0.094	0.039	0.0189	0.9827	11.7	8.6
Xmas17	Sviatoslav	1282	723	0.564	551	98	0.430	0.076	5.62	6.080	2.267	1.598	1.262	0.116	0.032	0.0175	0.9839	11.0	10.6
Xmas18	Sviatoslav	898	585	0.651	473	62	0.527	0.069	7.63	6.006	2.306	1.711	1.310	0.114	0.046	0.0119	0.9898	9.5	10.0

Some Properties of Adjectives in Texts

Gabriel Altmann

Abstract. In the present article only some of the infinite number of properties of adjectives will be discussed. One can propose and examine the properties of adjective without end. Especially the quantification is somewhat underdeveloped. We shall consider the semantic classes and their ranking in English, the length of adjectives in basic form, their compositional structure, position, and the POS basis. More attention is dedicated to their occurrence in poetry: the adjectival richness of strophes, the Busemann's coefficient and its development in texts.

Keywords: Adjectives, semantic classes, length, composition, position, Busemann's coefficient, English, Slovak

1. The problems

Adjectives occur in every text and may have different grammatical functions, e.g. *The new book* or *The book is new*, a direct or metaphoric meaning, there can be other categories derived from them (*news, newly, renewal, anew*), they can be simple, derived or compound, they can have synonyms, hypernyms, hyponyms, in some languages they can be synthetically graded, etc. In any case one can classify them in various ways. Classification is no theory, it is merely an elementary concept formation, just as it was made at the beginning in biology, sociology, etc., even if ordering in classes means a step upstairs in the hierarchy. Classifications may concern different aspects but there is none that would comprise all possibilities because the properties are established by us and our views develop. As a matter of fact, classifications do not represent reality but our views of reality.

But whatever the kind of classification, one can apply it in order to scrutinize the occurrence of adjectives in texts. It can be more or less regular, e.g. they can occur in (ir)regular intervals, the distances between them follow a special distribution, there may be a fixed rank-frequency distribution of classes, etc. This all can be observed only in texts, though various classifications are possible also on the basis of a dictionary.

Adjectives express some property, and each property can be quantified and measured, hypotheses may be set up, tested and inserted in a control cycle from which other hypotheses can be derived. Arriving at this point one stays at the threshold of a theory.

The adjectival structure of texts may be different even for a single writer, e.g. in his evolution, or for text types or languages. Translated texts need not have the same adjectival structure as the original text because one can replace adjectives by other expressions, e.g. metaphors or other parts of speech.

Adjectives may be contrasted with other word classes. They represent first order predicates of nouns and both in the language-learning by children and in style, one can differentiate degrees of ornamentality and activity. There are well-known indicators, e.g. Busemann's verb-adjective ratio, its relations to other indicators, and a number of concrete results from texts in different languages.

Adjectives themselves can be ordered in different classes. There are a great number of enterprises of this kind, unfortunately, they stop at achieving a classification because in qualitative linguistics the first ordering at the surface is sufficient. Here, we shall show at least one of them:

Yesypenko (2009) used 18 semantic adjectival classes: 1. *Traits of characterization*. 2. *Physical/natural condition*. 3. *Intellectual capacity*. 4. *Appearance*. 5. *Senses*. 6. *Age/time*. 7. *Temperature/sound*. 8. *Shape/size*. 9. *Flavour*. 10. *Weight*. 11. *Degree/intensity*. 12. *Color*. 13. *Actions done to the object*. 14. *Positive evaluation*. 15. *Evaluation of length/distance/position of the object*. 16. *Evaluation of value/function of the object*. 17. *Material*. 18. *Negative evaluation*. Other classifications and much more classes can be found in Warren (1984), Trost (2006), Schmale (2011), Wilson (2009), etc. Needless to say, in each class both a further sub-classification and several scaling are possible, hence one can obtain a quantified sub-classification. On the other hand, every linguist choosing a special aspect could perform the attribution of an adjective to a class differently, new aspects can be developed, an adjective can be ascribed to various classes (cf. e.g. *hard*, *good*) according to its environment, etc. If one considers any of the above mentioned Yesypenko classes, one sees that in each of them various kinds of Osgood's semantic differential could be applied. If this could be performed for a set of texts in a language, one could see which of the scalings is fruitful for setting up hypotheses and which of them can be linked with another property. If a construction of a link is possible, then the given scaling is fruitful theoretically, otherwise it is merely a descriptive means.

We conjecture that the ascription of adjectives to classes abides by a ranking distribution characterizing a text; further, if in a class scaling has been performed, the occurrence of individual values follows a function which can be derived theoretically. This is simply the consequence of the theme choice, of the text type, of the attitude and momentary mood of the writer, etc. All of these possibilities can be represented by the parameters of the model. The parameters may differ according to the text type or language. Finally, each of the possible aspects is linked with other properties of the text but the finding of the links is a task for the future. If one analyzed many texts in one language and applied a specific function to capture the representation of adjectives, then the differences between parameters are signs of the existence of boundary conditions. A rejection of the fitting of a well documented function (law) to a text does not mean definitive rejection: one must try to find the boundary conditions present in the given text, one must check the data, one must modify the hypothesis, etc. At this point literary scientists must begin to contribute by interpreting the facts. Last but not least, the membership in a class may be considered either as fuzzy or weighted in some way. In text analysis one uses mostly dichotomic decisions. The process of classification and attribution will never end but it must begin somewhere.

We recommend to solve at least the following problems:

- (1) To find the rank-frequency order of classes of adjectives in individual texts (for any type of classification).
- (2) To scale the adjectives in individual classes (i.e. to quantify the properties) and compare the classes.
- (3) To study the length of adjectives in the basic or given form.
- (4) To study their derivational and compositional structure.
- (5) To study their left-right position in relation to the respective noun in texts. The left-right position may lead to quite special classification types.
- (6) To study the polysemy of adjectives.
- (7) To study the basic part-of-speech from which the given adjective has been derived. Then evaluate the situation not only in the given text but in the whole language. This is a very complex task, especially in strongly synthetic languages.
- (8) If the given language is strongly synthetic, one may study the frequency of individual case of grammatical categories associated with each adjective (e.g. gender, case, number).
- (9) In neurolinguistics and psycholinguistics one frequently examines the sequence of two or more adjectives in front of a noun. Which class stays at the first place, at the second,

etc.? This is rather an expression of our relation to the reality but it may lead to a scaling and may be important for semantics.

(10) Not to speak about theory in cognitive linguistics until one did not set up a background system of basic relations enabling us to derive from it some hypotheses, test them and consider the well tested ones as laws. For theory building classification is not enough. It is merely the first step of ordering the facts.

Further problems will automatically be created and the domain will be extended step by step.

2. Ranking of semantic classes

The rank-frequency relation of all words of any text follows some variant of the Zipf distribution, as is known from the rich literature. Now, if we consider only special words, e.g. adjectives, we may ask whether the law holds true. This can be done in various ways: (a) each adjective (type) is taken and observed separately, e.g. in connection with its context. (b) The adjectives are ordered in classes and the frequencies of classes are scrutinized. Since variant (a) necessitates very long texts, it is much simpler to set up semantic classes of adjectives a priori. It must be remarked that different authors devised many different classifications. To decide which one is “correct” or “better” is not possible before one finds a theoretical background from which the given relation may be deduced. The criteria for decision can be restricted to a small number of recommendations: (1) Choose the simplest function derivable from a differential equation leaning against the unified theory (cf. Wimmer, Altmann 2005) and test it; (2) If not necessary, do not use polynomials. (3) If the test was positive, test the hypothesis on many texts in many languages. (4) Ignore the difference between distributions and functions; distributions are merely normalized functions; but if you have the respective software use one of the possibilities. Formulas do not express truth but give us a possibility to process the problem formally.

In order to illustrate the procedure (b) we use ready-made data and study the forming of the distribution of classes as presented by N. Yesypenko (2009) for three English texts. Yesypenko ordered the adjectives in 18 classes as given above. It is to be noted that many adjectives are polysemic and their assigning to a special class depends on the context, it cannot be performed mechanically. The same holds true for any other parts of speech. On the other hand, one will always find adjectives which do not belong to any of Yesypenko’s classes. Nevertheless, her classification can be used as a good starting point because she brought the respective numbers.

The ranking in Table 1 does not represent the numeration according to the above classification, it is a simple ranking according to frequency. Needless to say, the adjectives in these classes can be further classified or the properties expressed may be scaled according to various criteria. We remain at the first level and try to find a model common to three texts, namely “A Handful of Dust” by E. Waugh, “Gulliver’s Travels” by J. Swift and “The Adventures of Tom Sawyer” by M. Twain as presented by N. Yesypenko (2009). One automatically expects that ranking obeys by some Zipfian approach but here we may have to do with stylistic differences representing boundary conditions. That means, if it is necessary, one may add a parameter, otherwise one should strive for the simplicity of the respective function. Here we begin with a very simple approach, namely using the conjecture that the relative rate of change of frequencies in individual classes changes according to the attained frequency, i.e. we consider dy/y . Since we begin with rank 1, we may subtract it and obtain $dy/(y-1)$. We consider this relative repeat rate as constant and obtain

$$(1) \quad \frac{y'}{y-1} = -b$$

where $y' = dy/dx$, whose solution yields the simple exponential function

$$(2) \quad y = a \cdot \exp(-bx) + I.$$

Here, parameter a depends on the first frequency, parameter b shows the decline of frequencies from rank to rank. Of course, one can find excellent discrete probability distributions and other, “better”, functions with more parameters but we adhere to the above mentioned principles. The fitting of this function to the data presented by N. Yesypenko are displayed in Table 1.

Table 1
Fitting the exponential function to English adjective classes occurring in three texts
(Yesypenko 2009)

Rank	E.Waugh, <i>A Handfull of Dust</i>		J.Swift, <i>Gulliver`s Travels</i>		M.Twain, <i>The Adventures of Tom Sawyer</i>	
	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	122	129.77	120	121.90	93	101.63
2	118	110.06	99	104.98	87	89.94
3	100	93.36	93	90.42	81	79.60
4	68	79.22	75	77.91	75	70.46
5	67	67.25	66	67.14	69	62.39
6	59	57.11	63	57.88	63	55.25
7	48	48.52	60	49.92	54	48.95
8	45	41.24	57	43.07	48	43.37
9	44	35.08	54	37.18	42	38.45
10	42	29.86	33	32.12	42	34.09
11	22	25.45	18	27.76	21	30.25
12	22	21.70	12	24.02	18	26.85
13	19	18.53	6	20.80	12	23.84
14	7	15.85	3	18.03	12	21.19
15	6	13.58				
16	2	11.65				
17	2	10.02				
	a = 152.0448 b = 0.1662 R ² = 0.9646		a = 140.5740 b = 0.1508 R ² = 0.9221		a = 113.8698 b = 0.1236 R ² = 0.9236	

Of course, one can find other functions or distributions with “better” fit but as long as the determination coefficient R^2 is not smaller than 0.80 one may consider the fit as sufficient. The fit at high ranks is not quite satisfactory but preliminarily we accept it.

Since there are many classifications of adjectives – but tests are made only sporadically – we must adhere to the principle that that classification is fruitful which obeys some regularity expressed by the given hypothesis. Since we have few data, we performed the fit on the empirical basis, i.e., inductively, and cannot declare the law-likeness of the result. Many data of many languages must be collected in order to obtain a theoretically based result.

If one wants to make a step into the depth, one must show the link joining the given model with the model of another text property.

The comparison of the three texts presented above may be performed using the chi-square test comparing the frequencies in two texts.

The ranking of semantic classes has nothing to do with the so-called premodifying, i.e. the place of individual classes if there are several adjectives in front of a noun. Nevertheless, the results could be at least compared and the correlation could be studied.

3. Length of adjectives

The use of adjectives of different length is not only a problem of style but also of language and could be studied using translations. The adjectives created in strongly synthetic languages can be expressed only using several words in other languages, for example, the Hungarian *legmegszentségtelenítettebbeknek* can be translated into Slovak *najneznesvätiťnejším* but in German one needs several words: *den am meisten unentweihbaren* or *denen, die am wenigsten entweihbar sind*, and there are languages needing still more words. Automatically the question arises whether one should measure the length of the complete adjectival expression or merely that of the true adjective. Further, length of adjectives will surely be different in poetry where the writer must care for rhythm. The question is whether the situation is the same in all languages. Further, in monosyllabic languages the question is irrelevant; but in all other ones one can conjecture that there is a special distribution of lengths depending both on the given language (synthetic, analytic) and on the text type.

Since Yesypenko does not present the English adjectives, we consider them on the basis of the short Slovak poem *Kykymora* by A. Sládkovič and his long poem *Marína*. Here we do not distinguish individual classes but consider the adjectives as one set. Further, we ignore the case, number, gender and other category endings which may prolong the adjective because this is given only in strongly synthetic languages. Again, we apply formula (2) and present the data concerning *Kykymora* in Table 2.

Table 2
Length of adjectives in *Kykymora*

Length	Frequency	Exp+1
2	42	43.70
3	27	20.30
4	5	9.73
5	1	4.95
a = 208.8591, b = 0.7938, R ² = 0.9230		

As can be seen, the exponential function is a sufficient model for measuring the length of adjectives (in the given text). Usually, there is a small number of length classes, hence one must strive for finding a function with a small number of parameters. For *Kykymora* where the frequencies decrease monotonically, the exponential function with added 1, i.e. (2), is sufficient.

For the long poem *Marína*, formula (2) would be adequate, too, but there are too few monosyllabic adjectives. Compared with disyllables, one can see that there are only 2 monosyllables in the whole poem. The above formula yields a sufficient determination coefficient but the longest adjectives are not adequately captured. We make the following operation: we pool the class $x = 1$ with $x = 2$ and apply the Menzerathian function which is a generalization of the exponential, namely

$$y = ax^b \exp(-cx).$$

The results of fitting are presented in Table 3.

Table 3
Length of adjectives in *Marína*

Length	Frequency	Menzerath
2	1038	1038.03
3	494	493.77
4	72	73.88
5	15	5.87
6	2	0.31
a = 14584.6305, b = 9.8188, c = 4.7246, R ² = 0.9999		

It is to be remarked that in Slovak, there is a very small number of monosyllabic adjectives (e.g. *zlý, mdlý, ľvív, psí*). In English or French, this length is well represented and must be taken into account.

4. The compositional structure of adjectives

As every word, adjectives may be simple, affixal, compounded or representing whole expressions. Differentiating these aspects one can obtain four different results in some languages, however, the individual weights can be made more detailed, for example, a compositional adjective in which the basic adjective is the main member may have a greater weight than a compound in which it is secondary. Even affixes can be weighted. Other classifications are possible, too, and will appear stepwise in science. This aspect is mostly very simple and leads to quite restricted result in texts. In poetic texts, compounding of adjectives is quite seldom, because the writer must care for rhythm. In scientific texts the adjectives may be longer but it depends on the extent of text whether a regular distribution will be found. In other texts they consist mostly of one component and it is not very convincing to apply a function to the data. Perhaps the analysis of a dictionary would show the differences between languages.

In the poem *Kykymora* by Sládkovič we obtain the results presented in Table 3. As can be seen the fit is very good in spite of the fact that the sequence is quite short. Since all adjectives in Slovak are built by some categorical affix, we consider an affix only if it is not used for expressing adjectivity, e.g. *bezprostredny* (immediate) has a prepositional prefix *bez*, and even *pro-* can be considered a prefix. The adjective forming affix *-ný* is not taken into account. There is a possibility that an adjective is at the same time compound and has one or more affixes. Such construction will be called compound affixal and obtain preliminarily the weight 4.

Table 4
Compositional structure of adjectives in *Kykymora*

Rank	Frequency	Exp + 1
1 (Simple)	61	61.07
2 (Affixal)	13	13.36
3 (Compound)	2	3.54
a = 291.8949, b = 1.5807, R ² = 0.9986		

The exponential function with added 1 is sufficient here.

5. The position of adjectives

In some languages only one sole position of the adjective in relation to its respective noun is possible, e.g. in Indonesian the adjective is always behind the noun (postmodifying), in Hungarian always in front of the noun (premodifying). In other ones, both positions are possible but one of them may have a secondary meaning, e.g. poeticity. But there are languages in which both positions are equally possible. Before performing a count, one must decide about the variant used. There are also possibilities to present the adjective in two parts (left and right) but this is a seldom case in poetry.

In Slovak, the place of adjectives is in front of the noun, its placing behind the noun evokes a poetic color. In *Kykymora* we find the following results: In front of = 34, behind = 43. Since more than a half of adjectives are placed behind the noun, the text is strongly poetic. One could compute the probability of “behind” using the binomial distribution if one knew the basic probability p . For prosaic texts it is 1, hence one could not obtain a reasonable result. Neither the use of the chi-square yields reasonable results, because one does not expect equality. Hence, preliminarily, one must comply with the proportion of $B(\text{ehind})$ which is $43/77 = 0.558$. The study and ordering of poetic texts would, perhaps, lead to the weighting of lyricism, epicism, etc.

6. The basis of adjectives

Adjectives may be basic words, differing from other POS or not (e.g. in strongly isolating languages where one can identify them only on the basis of their position), or they may be derived from other POS, e.g. the German *mächtig* is derived from the noun *Macht*, active or passive participles may be derived from verbs (E. *-ing*, *-ed*) and the verb itself is already derived from a noun, e.g. in Slovak: *strach* (noun), *strašit`* (verb), *postrašený* (passive participle = adjective). The last one can be considered an adjective of the second level because there is also the adjective *strašný*. But there are also bases which produce nouns, adjectives, verbs, adverbs, etc. at the same time and one cannot decide what is the synchronic base, e.g. in Slovak: *celok*, *celý*, *celkom*, *ucelit`* (the whole, whole, wholly, make complete). In other languages it may be quite complex or very simple. In such cases, we shall consider the adjective as basic. There are even adjectives derived from other adjectives, e.g. Slovak *hluchy* (deaf) may be used to express the process of becoming deaf, *ohluchlý* (got deaf). An analysis of this phenomenon could, perhaps, help to language typology, text characterization, text type comparison, etc. but if the basis is not quite evident, the classification becomes problematic.

One can easily find many other properties which could be quantified. To each, at least the ranking and its model should be found, otherwise one stays at the level of classification.

7. The adjectival richness of strophes

Whatever the length of strophes (in terms of lines), one can count the number of adjectives in individual ones. The more adjectives there are, the more poetic is the poem. The number of adjectives in a strophe yields a distribution of poeticity. Our question is: Is there a regularity? Do the frequencies abide by a distribution i.e. can they be expressed by a formula?

The poem *Kykymora* is too short – it contains only 11 strophes but *Marína* contains 291 strophes and must be adequate to be modeled. The number of adjectives in individual strophes is shown in Table 5. In order to be as simple as possible, we fit the non-normalized function representing otherwise the Poisson distribution, i.e.

$$y = a^x/x!$$

which yields the values from $x = 0$ to $x = 14$ as shown in Table 5.

Table 5
Number of adjectives in individual strophes of *Marína*

Number of adjectives	Number of strophes	Theoretical frequency
0	1	1.00
1	11	5.65
2	18	15.99
3	32	30.14
4	36	42.61
5	49	48.19
6	48	45.42
7	39	36.69
8	25	25.94
9	15	16.30
10	6	9.22
11	6	4.74
12	1	2.23
13	2	0.97
14	3	0.39
a = 5.6550, R ² = 0.9738		

The determination coefficient is very high, hence we can accept this first approximation. One automatically asks whether other properties – which must first be defined – e.g. verb valency - abide by the same regularity. If so, a new domain of investigation may be opened and the individual properties may be compared with those in other text types.

Various characterizations of the empirical frequencies may be used as degrees of poeticity. Again, this depends also on the number of lines in the strophe. For comparative purposes one must relativize the resulting indicator. If the strophes are not equal, e.g. in sonnets, each “strophe-weight” must be divided by the number of lines. In sonnets, the first two numbers must be divided by 4; in the last two strophes, the frequencies must be divided by 3. In spite of this “normalizing”, there are merely 4 numbers (expressing the poeticity of the 4 strophes). Here no distribution can be sought but the average and other indicators can be computed.

Each part of speech contributes in some way to the poeticity but no evaluations are known. Evidently, it is better to process long texts because they yield more reliable results.

8. The Busemann coefficient

Since adjectives represent a rather static phenomenon, one compares their occurrence frequently with that of verbs which express a rather dynamic phenomenon. Of course, not all adjectives are static and not all verbs are dynamic, e.g. “to be”. Besides, some forms of verbs may be classified as adjectives (e.g. participles), hence statics and dynamics are properties that can be scaled. This can be done either by responses of test persons or authoritatively, by definition.

Busemann’s coefficient (1925; cf also Boder 1949; Schliessmann 1948; Antosch 1953; Fischer 1969; Altmann 1978; Altmann, Köhler 2015) can be defined here as the relation of adjectives to the sum of adjectives and verbs, i.e.

$$Q = \frac{A}{A+V}$$

where A is the number of adjectives in text, and V the number of verbs, that is, it is a simple proportion. It expresses a state of the text. There is also another possibility to study the dynamic behavior of the text, namely by considering the number of adjectives before the x^{th} verb. One obtains an increasing function which can be expressed formally, even for short texts. Consider, for example the Slovak text *Kykymora* for which one obtains:

[a,v,a,a,a,a,v,a,a,a,v,a,v,v,a,a,v,a,v,v,a,v,a,v,v,v,a,v,a,v,a,a,v,a,a,a,a,a, a,v,a,v,v,v,a,v,v,a,v,a,v,a,v,a,a,v,a,a,v,a,a,v,a,v,a,v,v,a,v,a,v,v,a,v,v,a,a,a,a, v,a,a,a,v,a,v,v,a,v,v,v,v,v,v,v,a,v,a,v,a,a,v,v,a,v,a,v,v,a,v,a,v,v,v,v,a,a,a,v,a,v,v,v, v,a,v,v,a,a,v,v,a]

Writing the rank of v as x and the number of a in front of it we obtain the result presented in Table 6.

Table 6
Number of adjectives in front of the x^{th} verb in the Slovak poem *Kykymora*

Rank of v	Number of a	Rank of v	Number of a	Rank of v	Number of a	Rank of v	Number of a
1	1	21	22	41	50	61	67
2	6	22	24	42	50	62	67
3	10	23	30	43	51	63	69
4	11	24	31	44	51	64	70
5	11	25	31	45	52	65	70
6	13	26	31	46	52	66	71
7	14	27	31	47	57	67	73
8	14	28	32	48	60	68	74
9	15	29	32	49	61	69	74
10	16	30	34	50	61	70	74
11	16	31	34	51	62	71	74
12	16	32	36	52	62	72	78
13	17	33	37	53	62	73	79
14	18	34	40	54	62	74	80
15	18	35	43	55	62	75	80
16	18	36	45	56	62	76	80
17	18	37	46	57	62	77	81

18	19	38	48	58	62	78	81
19	20	39	48	59	63	79	83
20	22	40	49	60	64	80	83

It can easily be tested that the increase of adjectives (*a*) takes place on a straight line $y = 4.9699 + 1.1071x$. Since the parameter *b* is approximately 1 and $R^2 = 0.9880$ one may speak about a dynamic ornamentality/activity equilibrium. However, it can be expected that not all texts display this structure. If the number of verbs is too great, the straight line need not be adequate. Therefore we fit the power function defined as

$$y = a * x^b$$

yielding the results presented in Table 7

Table 7
Fitting the power function to the number of adjectives in front of the x^{th} verb in the Slovak poem *Kykymora*

Rank	Number of a	Power ft.	Rank	Number of a	Power ft.
1	1	1.96	41	50	47.69
2	6	3.55	42	50	48.68
3	10	5.04	43	51	49.67
4	11	6.45	44	51	50.66
5	11	7.81	45	52	51.65
6	13	9.14	46	52	52.64
7	14	10.43	47	57	53.62
8	14	11.70	48	60	54.60
9	15	12.95	49	61	55.57
10	16	14.18	50	61	56.55
11	16	15.39	51	62	57.52
12	16	16.58	52	62	58.48
13	17	17.76	53	62	59.45
14	18	18.93	54	62	60.41
15	18	20.09	55	62	61.37
16	18	21.24	56	62	62.33
17	18	22.37	57	62	63.29
18	19	23.50	58	62	64.24
19	20	24.62	59	63	65.19
20	22	25.73	60	64	66.14
21	22	26.83	61	67	67.09
22	24	27.92	62	67	68.03
23	30	29.01	63	69	68.97
24	31	30.09	64	70	69.91
25	31	31.16	65	70	70.85
26	31	32.23	66	71	71.79
27	31	33.30	67	73	72.72
28	32	34.35	68	74	73.65
29	32	35.40	69	74	74.58
30	34	36.45	70	74	75.51
31	34	37.49	71	74	76.44
32	36	38.53	72	78	77.36

Some Properties of Adjectives in Texts

33	37	39.56	73	79	78.28
34	40	40.59	74	80	79.20
35	43	41.62	75	80	80.12
36	45	42.64	76	80	81.04
37	46	43.65	77	81	81.96
38	48	44.66	78	81	82.87
39	48	45.67	79	83	83.78
40	49	46.68	80	83	84.69
a = 1.9592, b = 0.8595, R ² = 0.9887					

For the German poem *Der Erlkönig* by Goethe we obtain a quite different picture:

[v,v,v,v,v,v,v,a,v,v,a,v,v,a,v,v,a,v,v,a,a,v,v,a,v,v,v,a,v,v,v,a,v,v,a,a,v,v,a,v,v,v,
v,v,v,v,a,v,v,a]

The number of adjectives is small (16), that of verbs much greater (39).

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
0 0 0 0 0 0 0 1 1 2 3 3 4 4 4 5 7 7 8 8 8 9 9 9 9 10 10 12

30 31 32 33 34 35 36 37 38 39
12 13 14 14 14 14 14 14 15 15

The last adjective is omitted because there is no verb behind it. As can be seen, the increase is not a straight line and cannot be captured by the above formula. However, it can be captured by another simple formula, namely the well known power function $y = a \cdot x^b$ yielding the results presented in Table 8

Table 8
Number of adjectives in front of the xth verb in *Der Erlkönig* by Goethe

Rank of v	Number of a	Power f.	Rank of v	Number of a	Power f.
1	0	0.10	21	8	7.01
2	0	0.27	22	8	7.48
3	0	0.48	23	9	7.95
4	0	0.71	24	9	8.44
5	0	0.96	25	9	8.93
6	0	1.24	26	9	9.42
7	0	1.53	27	10	9.93
8	0	1.85	28	10	10.44
9	1	2.17	29	12	10.96
10	1	2.51	30	12	11.49
11	2	2.87	31	13	12.02
12	3	3.23	32	14	12.56
13	3	3.61	33	14	13.10
14	4	4.00	34	14	13.66
15	4	4.40	35	14	14.22
16	4	4.81	36	14	14.80
17	5	5.24	37	14	15.35
18	7	5.67	38	15	15.93

19	7	6.11	39	15	16.51
20	8	6.56			
a = 0.1040, b = 1.3831, R ² = 0.9695					

If the number of verbs strongly increases, one may fit the sequence using the straight line but the parameter $b < 1.00$. Consider, for example the German text taken from the German newspaper “Der Bote”, November 12, 2017 (*Hohe Lebenszufriedenheit*). One obtains the results resented in Table 9.

Table 9
Number of adjectives in front of the xth verb in a German newspaper text
(Der Bote, November 12.2017, Hohe Lebenszufriedenheit)

Rank of v	Number of a	Comp	Rank of v	Number of a	Comp
1	0	1.12	21	15	14.13
2	0	1.77	22	16	14.78
3	0	2.42	23	17	15.43
4	3	3.07	24	17	16.08
5	3	3.72	25	17	16.73
6	6	4.37	26	18	17.38
7	6	5.02	27	18	18.03
8	6	5.67	28	18	18.68
9	6	6.32	29	19	19.33
10	6	6.97	30	19	19.98
11	7	7.62	31	20	20.63
12	7	8.27	32	22	21.29
13	9	8.92	33	23	21.93
14	10	9.57	34	23	22.58
15	11	10.22	35	24	23.23
16	12	10.87	36	24	23.88
17	12	11.52	37	24	24.53
18	12	12.17	38	24	25.18
19	15	12.82	39	24	25.83
20	15	13.47	40	24	26.48
a = 0.4692, b = 0.6503, R ² = 0.9789					

We may state that an equilibrium exists if the increasing sequence can be captured by a straight line whose parameter $b = 1$. It is, of course possible to test the difference of two b-s but for the time being we do not have a sufficient number of texts from various languages.

If we fit the straight line to *Der Erlkönig*, we obtain $a = 0$, $b = 0.3768$ and $R^2 = 0.9295$. One sees that the text is rather active than ornamental.

9. Conclusion

Adjectives, just as any other part of speech, have, so to say, an infinite number of properties. They can be phonic, semantic, grammatical, positional, simple, compounded, they can behave in special ways, etc. Many studies in various languages are necessary in order to present at

least a preliminary picture. The descriptions in school grammars are mostly grammatical and semantic, capturing some qualities of adjectives. Here we tried to show a view of their behavior in texts and the first steps in quantification.

References

- Bache, C.** (1978). *The Order of Premodifying Adjectives in Present-Day English*. Odense: Odense University Press.
- Hundsnurscher, F., Splett, J.** (1982). *Semantik der Adjektive des Deutschen. Analyse der semantischen Relationen*. Opladen: Westdeutscher Verlag.
- Krause, M.** (2011). Adjektive multifunktional vs. monofunktional. graduierbar vs. nicht graduierbar: Fragen. In: G. Schmale (ed.). *Das Adjektiv im heutigen Deutsch. Syntax. Semantik. Pragmatik* (Eurogermanistik 29): 15- 27. Tübingen: Stauffenburg.
- Leitzke, E.** (1989). *(De)nominale Adjektive im heutigen English*. Tübingen: Niemeyer.
- Oguy, A., Mgeladze, M.** (1993). *Sistema prilagatelnych v „Pesne o Nibelungach“: rekonstrukcija srednevekovoj ozenocnoj sistemy*. Czernivczi: Chernivtsy UP.
- Schmale, G.** (ed.) (2011). *Das Adjektiv im heutigen Deutsch – Syntax. Semantik. Pragmatik* (= Eurogermanistik; 29). Tübingen: Stauffenburg.
- Trost, I.** (2006) *Das deutsche Adjektiv. Untersuchungen zur Semantik. Komparation. Wortbildung und Syntax*. Hamburg: Buske.
- Warren, B.** (1984). *Classifying Adjectives*. Gothenburg: Acta Universitatis Gothoburgensis.
- Wilson, A.** (2009). The well-formedness of two psychoanalytic word categories in Portuguese texts. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of Text Analysis: 285-307*. Chernivcy: CNU.
- 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: de Gruyter.
- Yesypenko, N.** (2009). An integral qualitative-quantitative approach to the study of concept realization in the text. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of Text Analysis: 308-327*. Chernivcy: CNU.

José María de Oleza Arredondo, S.J. (1887–1975)

He was born in Palma de Mallorca, Spain (14 July 1887), and died in Santa Vera Cruz, Bolivia (6 Sept 1975). “S.J.” are the initials of “Society of Jesus” (“*Societas Iesu*” in Latin and “Compañía de Jesús” in Spanish), a Catholic religious congregation whose members are called Jesuits¹. Reviewing the origins of quantitative linguistics in Spain – and Catalonia in particular – requires attention to be paid to some singular figures, such as the Jesuit Father José María de Oleza Arredondo. He is a singular case because, strictly speaking, his only known contribution to Quantitative Linguistics is *Spanische Lautdauer*² (Menzerath & De Oleza 1928); this work was co-authored with the German phonetician Paul Menzerath (1883–1954). The actual role of de Oleza in this publication is unclear.

However, as we will see, the open-minded mentality and philosophy of José María de Oleza, S.J., was actually that of a quantitative linguist who tried to study languages scientifically, as revealed by his active participation in the foundation of the *Oficina Romànica de Lingüística i Literatura*³ (OR), an institution that promotes the study of the Catalan language in all its manifestations (Iglésias 2005). Interestingly, De Oleza did not preside over it, because of his modern scientific convictions. For example, he wanted to make use of an international phonetic notation (that he considered better than others promoted by other local Catalan scholars), to change some Catalan spelling rules, or to pursue an ideology-detached scientific study of languages (Iglésias 2005).

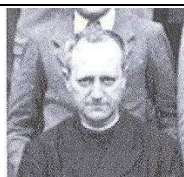


Figure 1: José María de Oleza Arredondo, S.J., (1887–1975), in addition to following his religious vocation, was a remarkable Romanesque philologist and grammarian. His timely collaboration with Paul Menzerath in the “*Spanische Lautdauer*” (1928), which may have initiated quantitative linguistics in Spain, was cut short by both his personal and historical circumstances. Photograph courtesy of *l'Arxiu dels Jesuïtes de Catalunya*.

José María de Oleza was born in Palma de Mallorca on May 14th, 1887; he was a son of the infantry captain Don (Mr.) Manuel de Oleza and Doña (Ms.) Concepción Arredondo⁴. He joined the *Societas Iesu* on August 8th, 1908, in Gandia, and received a presbytery in the village of Sarrià (Barcelona, Spain) in 1920, after his university training. First, he studied humanities in Veruela, Spain (1910–1914), where most Jesuits were trained. Then, he pursued postgraduate studies in philosophy at Gemer (in the

¹ It is usual to place these initials at the end of the full name of the members of this community, as it appears for example on the cover of the book *Spanische Lautdauer*.

² In German. English translation: “Duration of Spanish sounds”.

³ In Catalan. English translation: “Romance Bureau of Linguistics and Literature”.

⁴ The main biographical information for this article has been extracted from the personal file that was kindly provided by Father Francesc Casanovas, S.J., from the Archive of Jesuits of Catalonia (*Arxiu dels Jesuïtes de Catalunya*), as well as from the article published by Jesuitas Bolivia in *Diáspora*, on October 18th, 1975, supplied by Father Antonio Menacho, S.J., and from the article “A philologist, from the Rhine to the Andes” by Tambor Vargas, which was published in *La Patria*, Cultural Magazine *El Duende*, on September 30th, 2012 (<http://lapatriaenlinea.com/index.php?t=un-philologist-del-rin-a-los-andes¬e=121151>).

Netherlands) between 1914 and 1917, in the difficult context of the First World War. Finally, De Oleza graduated in theology at Barcelona and Valkenburg (Netherlands) in 1921. One has the impression that, without detriment to his religious vocation, both social circumstances and his scholarly interests (possible linguistic concerns) led José María de Oleza to become a priest at the age of 33 – a symbolic age in Catholicism, the practice being common for Jesuits in the early 20th century.

In fact, after he completed his training as a Jesuit and priest, his early love for philology led him to be part of the group of Catalan philologists, many of whom were also priests – such as Antoni Griera i Gaja (1887–1973), Antoni M. Alcover (1862–1932), and Josep Calveras, S.J. (1890–1964) - who studied in Germany and founded various institutions in Barcelona aimed at promoting the scientific study of Romance languages. These included, for example, the OR and the *Societat Catalana de Filologia*⁵ (SCF). Later, the objectives of these parallel institutions were taken over by the *Institut d'Estudis Catalans*⁶ (IEC), an institution that has been promoting Catalan studies up to the present day. These were difficult years for various reasons, including the dictatorships of Primo de Rivera (1923–1930) and of Francisco Franco (1939–1975); in these long periods, Catalan was banned and persecuted (Iglésias 2005). In between, in the periods of the Spanish Republic (1931–1936) and Civil War (1936–1939), the Jesuits were also persecuted and forced into exile.

Father De Oleza enrolled at the Rheinischen Friederich-Wilhelms-Universität zu Bonn, where he studied Romance Philology, Classical Languages, and General Phonetics. On December 14th, 1927, he defended his PhD thesis supervised by the eminent Romanist Wilhelm Meyer-Lübke (1861–1936): *Zur Bestimmung der Mundart der Katalanischen Version der Graalsage*⁷ (*Codex I.79, Ambrosiana, Milano*). The professors of the University wanted the superiors of the *Compañía de Jesús* to allow him to stay at the faculty, but De Oleza was in the end appointed teacher of Latin, Greek, and probably German, to the junior Jesuits in Veruela.

However, De Oleza took full advantage of his stay in Bonn, as evidenced by the epistolary relationships he had, especially with another Jesuit father, Josep Calveras (Iglésias 2007). During the period of his doctoral thesis, De Oleza greatly influenced his supervisor Meyer-Lübke in regard to considering Catalan as a fully-fledged Romance language and not merely an Occitan dialect, as Meyer-Lübke's publication "*Das Katalanische*" (1925) shows. In this respect, de Oleza was a free thinker and pioneer. However, his academic passion led him to go much further in his stay in Bonn – he collaborated closely with Paul Menzerath in his experimental phonetic study of Spanish. De Oleza probably sensed that his time in academia would end when he finished his thesis, since his superiors decided that he should return to Veruela to train young Jesuits.

The collaboration with Menzerath was probably carried out in 1927, although the *Spanische Lautdauer* was published the following year. As proved by several letters he sent to Josep Calveras, S.J., de Oleza desperately sought the funding necessary to publish the copies of his thesis, which the University of Bonn required for a positive evaluation to grant a PhD degree (Iglésias 2007: 134–140). In fact, only part of his thesis was published in the end, probably due to economic constraints (de Oleza 1928b).

Spanische Lautdauer is a seminal work in both phonetics and quantitative linguistics. Menzerath chose Spanish for this study as it is a relatively transparent

⁵ Catalan Association of Philology.

⁶ Institute of Catalan Studies.

⁷ In German. Translation into English: "On the determination of the dialect of the Catalan version of the Graalsage (Codex I.79, Ambrosiana, Milano)".

language, that is a language with an almost direct correspondence between the graphemes (or letters in the written language) and the phonemes, so that there is little difference between the acoustic segmentation of the syllables and the corresponding one in the written language. Menzerath's need for a native speaker and a philologist expert in Spanish is perhaps the reason for the involvement of de Oleza. Curiously, *Spanische Lautdauer* neither appears in the list of works at the Archive of the Jesuits of Catalonia, nor is it mentioned in the correspondence in a period in which de Oleza sought funds for his thesis (but was not successful in securing these). It could be that de Oleza feared that his superiors would consider such work an unnecessary distraction, at a time when the OR was also starting, or even that he humbly thought that his contribution to the book did not actually deserve an authorship, according to high ethical standards. We may never know the real reason.

In *Spanische Lautdauer* (1928), after classifying the words according to the number of syllables they contain and their accentuation, the authors measured the duration of hundreds of words and each of their syllables, using an electromechanical device – the kymograph – that Menzerath had refined. The segmentation of the signals seemed crucial to Menzerath. In his laboratory, he could carry research with various mechanical devices that he had developed in previous years; among these was the redesigned kymograph, which allowed him to record on paper the sounds emitted by the human voice. The kymograph had been invented by Pierre-Jean Rousselot (1898), and subsequently evolved into various models by Scripture (1906), Meyer (1911), and Panconcelli-Calzia (1924), according to Hess (1983: 93–103).

De Oleza knew how Juan de Pablo Bonet (1573–1633) had been reflecting on the teaching of deaf-mutes much earlier, in 1620, in his *Reducción de las letras y arte para enseñar a ablar los mudos*⁸, one of the first treaties about modern phonetics and speech therapy. De Pablo Bonet had pointed out that all the articulating organs collaborate jointly in the production of speech sounds, establishing the popular metaphor of the Spanish guitar, which would be revisited by the famous Spanish phonetician Tomás Navarro Tomás (1884–1979) in his *Manual de Pronunciación Española*⁹ (Navarro Tomás 1918: 156). According to this metaphor, the position of the sound producing organs when a voice is produced is reminiscent of how fingers are placed on the strings of a Spanish Guitar when it is being played.

Menzerath, on the other hand, was familiar with the 1923 German translation of the *Manual de pronunciación española*, a work cited as a reference manual at the beginning of *Spanische Lautdauer* (Menzerath & de Oleza 1928: 1). In many respects, Menzerath was an heir to the line of Germanic research inaugurated by the physiologist Ernst Wilhelm von Brücke (1819–1892), who highlighted the need to separate phonetics from philology, connecting – in his *Grundzüge der Physiologie und Systematik der Sprachlaute für Linguisten und Taubstummlehrer* (1856) – the former with human anatomy. In his works, Brücke described for the first time the articulatory characteristics of the phonetic production. However, he treated the language elements as isolated, and with a unique phonetic correspondence. By contrast, Menzerath, beginning in *Spanische Lautdauer*, tried to overcome the static conception proposed by Brücke, and to complete it – jointly with the Portuguese linguist Armando Soeiro Moreira de Lacerda (1902–1984) – with a definition and early understanding of the phenomenon of coarticulation. This was the intuition that Juan de Pablo Bonet had three hundred years before with the metaphor of the Spanish guitar.

⁸ In Spanish. In English: “Reduction of letters, and the art of teaching the dumb to speak”.

⁹ In Spanish. In English: “Handbook of Spanish pronunciation”.

Menzerath and de Oleza tabulated the mean durations of the words according to their numbers of syllables and the position of the stress, and presented the results graphically (Menzerath & de Oleza 1928). It is important to note that Menzerath and de Oleza prefer the concept of 'sound' (*laut*) rather than that of the phoneme; thereby, they refer to a modern materialist (physical) perspective that began in the 19th century and which many phoneticians defend, as it is a way to avoid the controversial debate about the actual existence of phonemes beyond their abstract or theoretical conception. In their joint book, Menzerath and de Oleza focus on formulating and confirming linguistic hypotheses that range from certain phonetic particularities of Spanish to some more general ones, such as the laws that are highlighted by Best and Rottmann (2017: 100) when reviewing the history of quantitative linguistics:

The Menzerath-Altmann law processes a different perspective of language: the “vertical” one that connects linguistic levels with each other. A decisive step towards that was the measurement of the duration of sounds, syllables and words in Spanish; they resulted in the formulation of general and specific quantitative laws by Menzerath & de Oleza (1928: 68ff.). The most important ones were as follows (marking as in the original):

“1. The average duration of the sound in the word becomes *smaller*, when the number of sounds in the word *increases*.” (68)

“4. A sound becomes *shorter* when the number of syllables of the word increases.” (70)

“6. The average duration of syllables in general *decreases* when the number of syllables of the word... *increases*.” (71)

“9. The average duration of a word *increases* when the number of sounds in a word *increases*.” (73)

This is how Menzerath and de Oleza deduced that Spanish words with longer durations tend to contain shorter syllables and sounds, with few statistical variations depending on the position of the accent (Menzerath & de Oleza 1928).

However, it was not until 1954 that the so-called *Menzerath's law* would be established and generalized in linguistics through the work *Die Architektonik des deutschen Wortschatzes* (Menzerath 1954). Menzerath proposed this as a principle of general linguistic economy according to which “the larger a construct, the smaller its components”. This 1954 work is often cited as the origin of Menzerath's law, though this actually dates back to the *Spanische Lautdauer* of 1928. Only in a few cases is the origin of Menzerath's Law connected to his presentation at the First International Congress of Linguistics of Leiden (1928), at which time he already had the results of the experimental study of Spanish (Menzerath 1928). As is well-known, since the *Spanische Lautdauer* more than fifty years passed until Gabriel Altmann (1931-) formalized Menzerath's law mathematically in his *Prolegomena to Menzerath's Law*. Altmann's formalization is commonly known as the Menzerath–Altmann Law (Altmann 1980).

Finally, we will never know if the contribution of de Oleza in the *Spanische Lautdauer* is a mere anecdote in the history of quantitative linguistics, or whether his contributions were substantial. The subsequent trajectory of Paul Menzerath suggests that the contribution of de Oleza could have been secondary; however, it may be that de Oleza, in addition to serving as a native informant, made some relevant contributions at least in the reflections about the phonetics of Spanish, of which he was an expert, as some of his letters confirm (Iglésias 2007). De Oleza's academic career was cut short at that point, so we may never know the magnitude of his influence.

With regard to de Oleza's biography, after he finished his thesis there were very difficult years for him, due to various reasons. First, he was ostracised for being a Catalanist, then persecuted for being a Jesuit, and finally only tolerated during the Franco regime. So in 1928, he had to return to Veruela – in the midst of the dictatorship of Primo de Rivera – in order to continue his evangelizing task and to train junior Jesuits. Although *a priori* he was not suspected of radical Catalanism – as he was a son of a soldier and a Spanish-speaker – his thesis on Catalan linguistics and his participation in the OR with Father Calveras, S.J., could have been the reason why the order sent him discretely to Veruela. Then, he had to move – first to Bollengo (Italy) and then to the Netherlands, both during the Spanish Second Republic, and during the Spanish Civil War (1936–1939) – taking with him the junior Jesuits he had been educating in Veruela. One needs to bear in mind that the Second Republic devised a new Spanish Constitution that abolished the *Compañía de Jesús*, and Jesuits were persecuted in Spain as a result of that.

The Spanish Civil War and the arrival of Francisco Franco's dictatorship ended the OR and any possibility of studying Catalan linguistics in Spain, but restored the *Compañía de Jesús*, which had been banned during the Republic. De Oleza returned to Valladolid initially and then moved to Barcelona, where he taught at a secondary school run by Jesuits (Sagrat Cor de Jesús¹⁰). In the following years, he channelled his passion for linguistics into practically the only thing that was probably within his reach – he published several Greek and Latin grammars of pedagogical and scholastic scopes (De Oleza 1942a, 1942b, 1945a, 1945b).

With humility, he taught students at secondary school and new Jesuits until 1952, when he was already 65 years old, at which time he volunteered to go to work for the Bolivian Jesuit missions. After many years humbly devoted to teaching, this was possibly the beginning of a new stage of his life. He settled first in the parish of the *Compañía de Jesús* in Cochabamba (Bolivia), and then in the *Colegio de Sucre*, a Jesuit school in the city of Sucre. His insatiable curiosity, intellectual interest in languages, and mission led him to learn Quechua, so that he could preach, take confession, and teach the native population. Not surprisingly, he followed with interest the Inter-American Indian Congress held in La Paz (1954) and, in 1955, he prefaced and revised a Quechua grammar and Quechua-Spanish Dictionary developed by the Jesuit missionaries Joaquín Herrero and Jorge Urioste (Herrero, de Oleza and Urioste 1955).

From 1960 until his death in Cochabamba on September 6th, 1975, Father José María de Oleza continued to teach languages, Gregorian, and liturgy in the novitiate and house of studies of the parish of Santa Vera Cruz. At the end of the obituary that the Jesuits dedicated to de Oleza (Jesuitas Bolivia 1975), one reads: “Those who knew him bear witness to the simplicity, humility, optimism – to all proofs and a good humour of this wise man, excellent religious, and fellow traveller.”

Antoni Hernández-Fernández (antonio.hernandez@upc.edu, Univ. Politècnica de Catalunya) & **Ramon Ferrer-i-Cancho**, Barcelona (ramon.ferrericancho@gmail.com, Univ. Politècnica de Catalunya)

¹⁰ The school is still active: <http://www.casp.fje.edu/>.

Acknowledgements

Firstly, we would like express our sincere gratitude and indebtedness to fathers Francesc Casanovas, S.J. (*Arxiu dels Jesuïtes de Catalunya*), and Antonio Menacho, S.J., for providing us with crucial advice and materials for this investigation. We are also very grateful to Gabriel Altmann for his encouragement and support. Stuart Semple, Naomi Escabias Potgieter, Michal Místeký and Peter Grzybek's feedback has been crucial for improving the quality of earlier versions. All remaining errors are ours. This research was supported by the grant TIN2017-89244-R from MINECO (*Ministerio de Economía, Industria y Competitividad*; Spanish Government), and the recognition 2017SGR-856 (MACDA) from AGAUR (*Generalitat de Catalunya*).

References

- Altmann, Gabriel** (1980). Prolegomena to Menzerath's law. *Glottometrika* 2, 1–10.
- Arxiu dels Jesuïtes de Catalunya** (2017). *Personal file about José María De Oleza Arredondo*. Supplied by the father Francesc Casanovas, S.J.
- Best, Karl-Heinz & Rottmann, Otto** (2017). *Quantitative Linguistics, an invitation*. Lüdenscheid: RAM-Verlag.
- Brücke, Ernst** (1876). *Grundzüge der Physiologie und Systematik der Sprachlaute für Linguisten und Taubstummenlehrer*. Wien: Druck und Verlag Von Carl Gerold's Sohn.
- Hess, Wolfgang** (1983). *Pitch Determination of Speech Signals. Algorithms and Devices*. Berlin: Springer-Verlag.
- Iglésias, Narcís** (2005). L'Oficina Romànica de Lingüística i Literatura (1928–1936). *Llengua & Literatura: 16*, 289–362.
- Iglésias, Narcís** (2007). *Epistolari de l'Oficina Romànica*. Montserrat: Publicacions de l'Abadia de Montserrat.
- Jesuitas Bolivia** (1975). José María Oleza Arredondo. *Diáspora*, October 18th, 1975. Supplied by father Antonio Menacho, S.J.
- Menzerath, Paul** (1928). Über einige phonetische Probleme. *Actes du premier Congrès international de linguistes*. Sijthhof: Leiden, 104–105.
- Menzerath, Paul & Lacerda, Armando** (1933). *Koartikulation, Steuerung und Lautabgrenzung: eine experimentelle Untersuchung, Volumen 1 de Phonetische Studien*. Berlin / Bonn: F. Dümmlers Verlag.
- Meyer-Lübke, Wilhelm** (1925). *Das Katalanische. Seine Stellung zum Spanischen und Provenzalischen, sprachwissenschaftlich und historisch dargestellt*. Heidelberg: Winter.
- Navarro Tomás, Tomás** (1918). *Manual de Pronunciación Española*. Madrid: Centro de Estudios Históricos.
- Vargas, Tambor** (2012). Un filólogo, del Rhin a Los Andes. *La Patria, Cultural Magazine "El Duende"*, September 30th, 2012.

De Oleza's Works

- De Oleza, S.J., José María** (1927). *Zur Bestimmung der Mundart der Katalanischen Version der Graalsage (Codex I.79, Ambrosiana, Milano)*. Thesis advised by Wilhelm Meyer-Lübke. Bonn: Universität zu Bonn.

- De Oleza, S.J., José María** (1928a). Diccionari català-valencià-balear. *Anuari de la Oficina Romànica de Lingüística i Literatura I*: 382–390.
- De Oleza, S.J., José María** (1928b). *Zur Bestimmung der Mundart der katalanischer Version der Graalsage. Inaugural-Dissertation zur Erlangung der Doktorwürde genehmigt von der philosophischen Fakultät der Rheinischen Friederich-Wilhelms-Universität zu Bonn*. Barcelona: Edicions Biblioteca Balmes [This is the only published part of de Oleza's PhD thesis].
- Menzerath, Paul & De Oleza, S.J., José María** (1928). *Spanische Lautdauer. Eine experimentelle Untersuchung*. Berlin / Leipzig: de Gruyter.
- De Oleza, S.J., José María** (1930). Recensió al Diccionari Català-Valencià-Balear (I), *Anuari de la Oficina Romànica de Lingüística i Literatura III*: 340–342.
- De Oleza, S.J., José María** (1942a). *Primer curso de lengua Griega. Método completo conforme al cuestionario oficial de 14 de abril de 1939, año de la Victoria. Preliminares. Morfología. Sintaxis. Geografía. Historia. Prácticas*. Barcelona: Ibérica.
- De Oleza, S.J., José María** (1942b). *Segundo curso de lengua Griega. Método completo conforme al cuestionario oficial de 14 de abril de 1939, año de la Victoria. Morfología. Sintaxis. Prácticas*. Barcelona: Ibérica.
- De Oleza, S.J., José María** (1945a). *Primer libro de traducción griega* de Antonio Guasch, S.J., (1915). Barcelona: Casals. [Second Edition of the original of 1915 by Antonio Guasch, S.J., improved by José María De Oleza, S.J.].
- De Oleza, S.J., José María** (1945b). *Gramática de la Lengua latina, según el método del P. Manuel Álvarez, S.J.* Barcelona: Ibérica. [2nd edition in 1944. Barcelona: Eugenio Subirana].
- Urioste, S.J., Jorge, De Oleza, S.J., José María & Herrero, Joaquín** (1955). *Gramática de la lengua quechua y vocabulario quechua-castellano, castellano-quechua de las voces más usuales*. La Paz: Editorial Canata.

Book Reviews

Haitao Liu, Junying Liang (eds.) (2017), *Motifs in Language and Text*. Berlin/ Boston: De Gruyter Mouton, pp. 271. (Quantitative Linguistics Vol. 71).

Reviewed by **Hanna Gnatchuk** (agnatchuk@gmail.com, Universität Klagenfurt)

Motifs of whatever linguistic entity are, so to say, the first step towards sequential generalization of any linguistic units. They were developed by R. Köhler (2006, 2008a,b, 2015) on the basis of musical motifs (cf. Boroda 1982) and developed further by many linguists. Motifs are sequential entities and just as any other entities in language they are the results of our definitional activity concerning language itself – not the surrounding reality. From the time of their creation, omnibus volumes were published (cf. Mikros, Mačutek 2015, cf. also Grzybek, Kelih, Mačutek 2010) and the definition of various motifs increases every year. The reviewed book is quite international but the majority of articles were written by Chinese authors. This fact is very pleasing because it testifies to the expansion of quantitative linguistics in the world. At the same time, it joins China and Europe in their common endeavor to develop quantitative linguistics. For any linguistic entity sequences may be constructed. The entities may be quantified and measured but even for qualitative entities Köhler developed method for their construction. Hence they represent a new dimension in linguistic investigations. There are two tasks: (1) to define as many different motifs as possible and (2) to find their place in the Köhlerian self-regulation circuit. To this end many hypotheses will be necessary and still more tests in many languages.

In the first article, *Persistency of Higher Order Motifs* (1-12), André Pascal Beyer asks a basic question: can we construct motifs of motifs? That means to make a next step on the ladder of abstraction? Though his example (evaluating the length of English words in terms of letter numbers) is fortunately only an example (later on, he takes syllables into account) but his question is justified and the answer is simple: if we can set up some hypothesis concerning motifs of motifs than we can and must test them; if there are no hypothesis, the step is not necessary. The author computes the entropy, Hurst-exponent and length for different orders of abstraction. The basic data are not given, so that interested readers can neither check the results nor use the data for setting up hypotheses. Unfortunately, the author presents everything only in form of figures but there is no formula expressing the relations studied. The study is based on Russian and Italian texts and DNA sequences.

In the second article, *On Motifs and Verb Valency* (13-36) by Čech, Vincze and Altmann only the first level of motifs, namely that of verb valencies, is examined in Czech and a Hungarian translation of Orwell's text. The data are completely presented in the Appendix. The authors investigated (1) the types of valency motifs and for their ranking proposed the Zipf-Mandelbrot distribution though they mention also some other ones. The spectrum of the frequencies is obtained by the transformation of the Zipf-Mandelbrot distribution and yields also excellent results. The motifs have a certain length, and the relation of frequency to average length is captured by the Lorentzian function used today in many linguistic investigations. For the Hungarian data one obtained the simple power function. It has not been said what are the boundary conditions for Hungarian but the linguists have a free access to all data.

In the next article, *Chinese Word Length Motif and its Evolution* (37-64) by H. Chen and J. Liang, the authors compare word length motifs in spoken and written Chinese. In the spoken Chinese they obtained for all texts excellent results for the rank-order/frequency

relation in form of the power function and the same for written Chinese. All “numbers” are presented and the functions are shown graphically. For length and frequency they obtain good results using the Hyper-Pascal distribution. The function shown as an example (Text 1) is concave, hence three parameters are adequate. Both in spoken and in written text they found only two exceptions (S9 and W13) in which the probability is slightly low. Further, the authors perform a very useful investigation. They divide the history in 6 periods and compare the parameters of the rank frequency relationship (a and b). They obtain once an increasing trend in a

and a decreasing one in b . In the second case a decreases and b increases (Tables 11 and 12) but it is not clear which of them in written and spoken Chinese). The same is done for length and frequency relation. The results are excellent and show that motifs are “good” abstractions and one may consider them also historically.

In the fourth article, *Quantitative text Classification Based on POS-motifs*.(65-85), Ruina Chen studies qualitative sequences and their properties in Chinese and English, namely parts of speech. She used many texts and many indicators (richness, entropy, repeat rate, Gini’s coefficient, hapax legomena, type-token relation) and performed a discrimination of texts. She introduced also the method of random forests, discriminant analysis and presented a classification of text types.

In the fifth article, *L-motif TTR for Authorship Identification on Hougloumeng and its Translation* (87-108), Yu Fang uses the length motifs in order to solve a problem of authorship of the *Hougloumeng* text. The author uses formulas for comparison of two translations, shows the sequences of TTR, compares parameters, studies their evolution, presents everything in figures and numbers, there is a number of examples. As can be seen, the motifs are useful even in analyses of higher levels of literary studies. It is not enough to make descriptions, comparisons and evaluations using merely words: some background theory is formed on the basis of motifs.

In the sixth article, *Length Motifs of Words in Traditional and Simplified Chinese Scripts* (109-132), Wei Huang measures the length of Chinese signs in terms of stroke numbers. Some signs become simplified and this must be very honored for any language – unfortunately, it happens quite seldom. The author compares traditional and simplified texts in terms of motif lengths. All numbers are shown, the power law is fitted and the results are presented so that a reader can check them and use for his purposes. For motif lengths the author uses the Hyper-Pascal distribution yielding good results. Everything is well documented with numbers, formulas, many tables and figures. Perhaps the only error is the abbreviation of European names in references where the Chinese and Hungarian way of writing the names is used. Thus G. Jozef, is, as a matter of fact: Genzor, J. But the Europeans make the same errors with Chinese names.

In the seventh article, *Dependency Distance Motifs in 21 Indo-European Languages* (133-150), Y. Jing and H. Liu analyze distances between dependent parts of sentence. They distinguish between decreasing, increasing and equal motifs, show some regularities and make (without tests) some conclusions concerning preferences. They set up some questions, conjecture that a special order could reduce the complexity of languages, formulate some hypotheses but the only results is the ordering of Indo-European languages according to some results. There are no formulas and no tests but the ideas can be exploited for studying the dependency distance in its relation to other grammatical properties. The distance is well defined, hence for each sentence in the 21 languages one could perform a comparison with another measured property of the sentence.

In the eighth article, *Word Length Distribution and Text Length: Two Important Factors Influencing Properties of Word Length Motifs* (151-163), G.K. Mikros and J. Mačutek study the situation in Greek and analyze the above mentioned relation. They state

that motif properties carry some information not present in word length distribution. They corroborate the power function for the type-token relation of motifs, find the relation of the exponential parameter to text length. All formulas and figures are attached, they refer also to 70 Ukrainian texts analyzed by Mačutek (2015).

In the ninth article, *Quantitative Genre Analysis Using Linguistic Motifs* (165-180), Yaqin Wang applies Köhler's L-, F- and T- motifs in order to distinguish genres, i.e. for text sort classification. This is a further relation of motifs to other subsystems. The rank-frequency relation of L-motifs follows evidently Zipf-Mandelbrot (test and figures are shown). A special table shows that parameters a and b are different for different genres, and they are correlated. The next logical step would be finding further properties and their different presence in different genres. Needless to say, it is a beginning of an infinite work.

The eleventh article, *The Rank-frequency Distribution of Parts-of-speech Motifs and Dependency Motif in the Deaf Learners' Compositions* ((181-200) is quite special. Here, the authors asks whether motifs can be used also in slightly different environment. The answer must be positive because they come from the music and were applied also for the study of DNA. Here the dependency relations in sentences of deaf persons is studied. It is shown that parts-of-speech motifs follow the usual Zipf-Mandelbrot dependency. The age groups (primary, junior, senior) are distinguished. We miss some ranking tests for all comparisons but the field of applications of statistical methods is here very extensive.

The twelfth article, *Quantitative Properties of Polysemy Motifs in Chinese and English* (201-216) by Jiang Yang is a semantic comparative study. The authors uses two dictionaries and two corpora, constructs the polysemy motifs. Their rank-frequency distribution follows Zipf-Mandelbrot but as is well known, when the frequencies are very large the P-value is not reliable and one uses rather the C-value. Fitting the function itself one obtains the determination coefficient which is here in all cases very long and reliable. All tests are shown, one finds all figures. The next problem was to look at the length of polysemy motifs. The use of the mixed negative binomial distribution was possible because of the enormous number of data. Again, the chi-square does not say anything but the determination coefficient is greater than 0.9. Though the author strives for substantiating the mixed negative binomial distribution, one can expect that a simple (not probabilistic) function had captured the results slightly better. The study shows also the usual problems in the modern quantitative linguistics: one uses corpora representing a mixture of texts, hoping that the size of the corpus eliminates deviations. One hopes that the corpora are uniform. Still more problematic is the use of translations, especially of religious texts, etc. Thus, there are not only mathematical problems but also linguistic ones. Up to now, nobody said what can be compared.

The thirteenth article, *The Words and F-motifs in the Modern Chinese Version of the Gospel of Mark* (217-229), Cong Zhang analyzes six versions of the Chinese translations of Gospel of Mark. It is questionable whether different translations of a religious text can say anything about the differences between the versions of the target language. The author uses correct methods, fits successfully the power function to the F-motifs, shows all results (parameters and determination coefficients). Then types and tokens of words and F-motifs are examined and the correlation is stated. It should be remarked that computing the correlation is only the first step, the inductive one. In Table 5 and 6 only numbers can be found. The complete procedure is somewhat problematic: one studies the development and versions of a language using a translation of a religious text from another language(s). Here an enormous field of research in history and dialectology opens – if the texts are genuine (not translated).

The last article, *Motifs of Generalized Valencies* (231-260) by Hongxin Zhang and Haitao Liu considers the valency in Tesnière's sense. Beginning with verbs, all words of the sentence obtain a numerical characteristic expressing the number of words they govern. An excellent example enables the reader to do the same. Then motifs are constructed. Their types

are ranked and the Zipf-Alekseev distribution is fitted. They analyze Chinese and English with excellent results. For motif lengths the same holds. The study of time series is not very elucidating. The relation of length to frequency is captured by the Hyperpoisson distribution. Since all numbers can be found in the Appendix, the reader can perform his own experiments and search for other function.

The book as a whole surpasses the qualitative, inductive limits of the official linguistics and in the first step, it performs a visit in an abstract direction. If we consider the number of languages, the number of texts in them, their development etc. we may see that even this step leads into the infinity. It is to be hoped that similar volumes studying various properties of motifs will follow.

References

- Boroda, M.** (1982). Häufigkeitsstrukturen musikalischer Texte. In: Orlov, J., Boroda, M., Nadarejshvili, I (Eds.), *Sprache, Text, Kunst. Quantitative Analysen: 231-262*. Bochum: Brockmeyer.
- Köhler, R.** (2006). The frequency distribution of the length of length sequences. In: Genzor, J, Bucková, M. (eds.), *Favete linguis. Studies in honour of Viktor Krupa: 145-152*. Bratislava: Slovak Academy Press.
- Köhler, R.** (2008a). Word length in text. A study in the syntagmatic dimension. In: S. Mislovičová (ed.), *Jazyk a jazykoveda v pohybe: 421-426*. Bratislava: Veda.
- Köhler, R.** (2008b). Sequences of linguistic quantities. Report on a new unit of investigation. *Glottology 1(1), 115-118*.
- Köhler, R.** (2015). Linguistic Motifs. In: Mikros, G., Mačutek, J. (eds.), *Sequences in Language and Text: 89-108*. Berlin/Boston: De Gruyter.
- 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.
- Mačutek, J.** (2015). Type-token relation for word length motifs in Ukrainian texts. In: Tuzzi, A., Benešová, M., Mačutek J. (eds.), *Recent Contributions to Quantitative Linguistics: 63-73*. Berlin/Boston: de Gruyter-
- Mikros, G., Mačutek, J.** (Eds.) (2015). *Sequences in Language and Text*. Berlin/Boston: De Gruyter Mouton.
- 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: 183-193*. Wien: Praesens.

Mikhail Kopotev, Olga Lyashevskaya, & Arto Mustajoki. (Eds.) (2017). *Quantitative Approaches to the Russian Language*. New York: Routledge. ISBN:978-1-138-09715-5, 220 pp.

*Reviewed by Heng Chen*¹

Russian linguistics has a tradition of quantitative/mathematical linguistic studies, which can be dated back to the 19th century. Quantitative Linguistics (QL) in Russian had been developed synchronously with international QL studies. There was actually a boom for Russian quantitative studies in the 1960s-1980s, the most famous of which, including Piotrovsky's "Statistika Reči" ("Parole Statistics") group, Tuldava's series of quantitative studies regarding lexical systems, and Arapov's work of *Quantitative Linguistics*, etc. The experts in the "Parole statistics" group are not only from linguistics, but also other disciplines such as computer science, mathematics, psychology, and statistics, etc. However, the QL studies merely vanished after the end of the Soviet Union, although there are still several excellent QL researchers such as B.B. Kromer and A.A. Polikarpov. Kelih (2008) conducted a more systematic historical investigation of the application of quantitative methods in Russian linguistics and literature science, for a review, see Liu (2010).

Nevertheless, many large and deeply annotated corpora are available for extensive quantitative studies nowadays, such as the Russian National Corpus, ruWac, and ruTenTen, just to name a few. Most of these articles in this volume are achievements of a workshop entitled *Quantitative Approaches to the Russian Language*, which took place in August of 2015 in Helsinki, Finland, co-organized with a symposium called New Developments in the Quantitative Study of Languages. This volume is a new attempt in this field by applying the latest new techniques such as NLP tools, mathematical models, and machine learning algorithms to quantitative analyses of Russian big language data, meanwhile, the methods are also evaluated. This volume is focused on quantitative methodology and data processing of Russian language, representing state-of-art research in Russian QL. There are ten articles in this volume including the first introduction chapter, which are organized into four parts around the following topics:

Part I, Introductory chapters, including 2 contributions, opens with an introductory article titled "**Russian challenges for quantitative research**" by **Mikhail Kopotev, Olga Lyashevskaya, and Arto Mustajoki**, who are also editors of this volume. The authors begin by stating that the goal of the present volume is to present current trends in examining Russian QL, to evaluate the new research methods and models vis-a-vis Russian data, and to show the advantages and disadvantages of the methods and models. Then they describe the

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

main features of the Russian language and look back upon the quantitative (corpus) studies in Russian (2000-2010s), concluding that many topics in grammar and lexicon need to be covered, and more examples of quantitative approaches need to be provided. Next, the contributions in this volume are introduced. The inventory of internet sources and quantitative methods used in this volume are summarized at the end, which makes it advantageous for the inquiry.

The other contribution in this part, “**Big data and word frequency: measuring the consistency of Russian corpora**” by **Maria Khokhlova**, aims to compare linguistic phenomena across the main Russian corpora of different sizes. Specifically, 3 linguistic phenomena are examined, i.e., syntactic relations involving nouns, high-frequency nouns, and low-frequency nouns; the corpora include the ruWac and the ruTenTen; the main quantitative methods are log-likelihood score and Spearman’s correlation coefficient. The results obtained for the syntactic relations involving nouns in ruWac and ruTenTen are compared with each other, and the analyses show that the two corpora are largely similar in featuring syntactic relations. The results of the high- and low-frequency Russian nouns were compared with data published in *A Frequency Dictionary of Modern Russian*, which indicate that there are different situations for high- and low- frequency distributions comparisons. Further researches are needed for more parts of speech and other better metrics.

Part II, Topics in semantics, to be more specific, lexical semantics, contains 3 contributions. It begins with an article titled “**Looking for contextual cues to differentiating modal meanings: a corpus-based study**” by Olga Lyashevskaya, Maria Ovsjannikova, Nina Szymor, and Dagmar Divjak. An important property of modal words is that they are largely ambiguous. Thus the modals can be assumed to be “word-like elements which are poly-functional in the sense that they express no less than two types of modality”. The authors propose that the availability of large corpus data paves the way for a study of the empirical reliability of existing classifications originally proposed by philosophers. Thereupon, in order to test if contextual cues, i.e., 12 formal and semantic features (of the modals) can predict the type and function of modal words, the most frequent 6 Russian verbs were chosen, and for each word, 250 sentences were extracted from the RNC. The annotation of contextual cues for each word in the sentences was done by two experts manually. To achieve the aim, two visualization techniques, i.e., multiple correspondence analysis and shaded mosaic plots, and two inferential statistical methods, i.e., polytomous logistic regression, and classification and random forest were used. The results show that, generally, type or function can be predicted from context cues, also with some exceptions, which need further investigations in the future.

The study titled “**Automated word sense frequency estimation for Russian nouns**” by Anastasiya Lopukhina, Konstantin Lopukhin, and Grigory Nosyrev, is the first study on sense frequency distributions in the Russian language. The article begins with a well-known observation by G. k. Zipf (1945), stating that words used more frequently usually have more senses than words that are used less frequently. Although information about word frequency is widely available nowadays, sense frequencies and their distributions remain a neglected area in linguistics. In this paper, the authors present a method for determining noun sense frequency distributions automatically from raw text, an evaluation of the methods, its comparison to state-of-art system, and a discussion of its applications. The method is actually

based on word sense disambiguation techniques usually used in computational linguistic or NLP, using distributed vector representations with weighting. Distributed vector representations is a way of representing words as low-dimensional dense real-valued vectors, and is known as the famous word2vec family of methods. The linguistic hypothesis here is that words occur in similar contexts tend to have similar meaning. The evaluation results show that the frequency estimation error of the model is 11-15 percent. The results of the 440 nouns sense frequency information as well as source code are online for further consultation.

The third contribution in this part is **“Two centuries in two thousand words: neural embedding models in detecting diachronic lexical changes”** by Andrey Kutuzov and Elizaveta Kuzmenko. Similar to the above research by Lopukhina et al, this study traces Russian word semantic changes with state-of-art technique in lexical semantic similarity modeling: artificial neural networks (neural embedding models) in NLP. The central assumption here is that online training of such models with new textual data results in a “drift” of word vectors in the “semantic space”. The case presented in this study uses three sub-corpora from the RNC: texts produced before Soviet times (before 1917), during Soviet times (1918-1990), and after the fall of the USSR (since 1991). After training 3 neural embedding models on these 3 subcorpora, several algorithms to extract words with changing meanings are evaluated. Eventually, they came to conclusion that comparing nearest neighbor sets using Kendall’s τ distance works best, both on artificially created data and on short, manually compiled, gold standard data sets. The results of 2000 nouns and adjectives that have undergone the most significant changes are online for further consolation.

PART III, Topics in the Lexicon-Grammar Interface, including 3 contributions, begins with **“The grammatical profiles of Russian biaspectual verbs”** by Alexander Piperski. Biaspectual words can be used to convey both perfective and imperfective meaning. In this study, three quantitative methods for determining the status (more imperfective or perfective-like) of biaspectual verbs (over time) were evaluated: estimating the relative frequency of their perfective and imperfective gerunds, classifying their grammatical profile using the k Nearest Neighbors algorithm, and conducting an experiment on the perception of the inherent aspect of biaspectual verb forms. The results show that their applications are in agreement with each other.

A study conducted by Lidia Pivovarova, Daria Kormacheva, and Mikhail Kopotev, titled **“Evaluation of collocation extraction methods for the Russian language”**, begins with a distinction between lexical and empirical collocations, of which the later is the focus of this study. Then the authors review the main existing measures for collocation extraction, including t-score, log-likelihood, mutual information, Dice, and wFR. Next, the evaluation of automatically obtained collocations is conducted by comparing both with dictionary data and native speakers’ responses. Two comparisons both show that t-score performs slightly better than the other measures. However, they all provide similar results, which means that it may be more plausible to suppose that different measures are intended to identify different kinds of collocates.

The third contribution in this part is **“From quantitative to semantic analysis: Russian constructions with dative subjects in diachrony”** by Anastasia Bonch-Osmolovskaya. The author conducts a quantitative research into predicative and corresponding adjective constructions with dative arguments from a diachronic perspective. The core issue here is to

reveal behavior classes of lemmas defined in terms of dative argument frequency within the three forms (i.e., predicative, short adjective form, and long adjective form) and to study diachronic changes of the determined behavior classes. The data are from the RNC and the search is confined to two samples, one from the 18th century, the other from the 21st century. Eight lemmas are selected for the study. The investigation shows that the frequency rates of dative subjects are different from predicates, and diachronic trends are observed using hierarchical clustering methods.

PART IV, also the final part, turns our attention to **Topics in language acquisition**, including 2 contributions. **“Measuring bilingual literacy: challenges of writing in two languages”** by Aleksei Korneev and Ekaterina Protassova. This study focuses on a computer-based, contrastive assessment of bilingual Finnish-Russian primary students with different linguistic backgrounds, and examines their written language proficiency. To achieve this, experiments of four groups - the Russian Dominant Bilinguals with 15 children, the Finnish Dominant Bilinguals with 13 children, the Russian-speaking control group with 15 children, and the Finnish-speaking control group with 10 children - are conducted based on a computer handwriting assessment system. The handwriting parameters include mean time of writing a letter, the exact time to write separate letters, the stability of the edge of the line. To analyze the parameter differences among different groups of subjects and in different writing tasks (copying and dictation), the authors use the repeated measures ANOVA. The results show that the dominance of the language plays an important role in writing proficiency in bilinguals; the writing system is another important factor; the language of the environment might support the language skills, but training in a different language and in a different script supports the quality of writing.

The final contribution, **“When performance masquerades as comprehension: grammaticality judgments in experiments with non-native speakers”** by Robyn Orfitelli and Maria Polinsky. In language acquisition studies, many observations are based on experiments. However, inappropriate experimental design can be problematic, because it can be hardly replicated and re-examined. In this study, the authors criticize the grammaticality judgment tasks (GJTs) which are originally introduced in linguistics to measure the acceptability of particular language structures for native speakers, and are now misused for non-native speakers. Based on numerous instances of within- and across-task inconsistency, the authors argue that the metalinguistic demands imposed by the task - and the difficulty involved in identifying the root cause of any incorrect answers - render the task unsuitable for testing language comprehension with non-native speakers. Then the authors illustrate this problem by discussing two recent experiments conducted with Russian non-native speakers using GJTs and other tasks. The analysis suggests that poor performance on GJTs by non-native speakers may be related not to grammatical errors, but to extra-grammatical factors involving metalinguistic awareness and processing demands.

In conclusion, this edited collection presents a range of resources and new quantitative methods in Russian QL studies, which will promote the combination of classical QL with the latest techniques from the age of language big data. The authors show that those state-of-art techniques such as neural embedding models, word2vec, word sense disambiguation (WSD) algorithms and distributional semantic models, actually can and should be applied to quantitative studies of Russian language regarding modern linguistic questions. Moreover, a

series of evaluations of quantitative methods are conducted, and some theoretical problems are also scrutinized.

This volume was published with high typographical quality, and the index listed at the end of the book makes it very convenient for readers to read and refer. However, there are still a few critical points I am obliged to make. The “R2” in pp. 60, pp. 66, and pp. 72, should be “R²” or with the superscript “2”; the “Diagram 1” in pp. 168 should be “Figure 8.3”, and the “Figure 8.3” in pp. 169 should be “Figure 8.4”, otherwise it will be confusing. As for this collection, we believe that it would be greatly improved if it focuses more on quantitative linguistic laws (Köhler, 2012), and gives more in-depth linguistic interpretations and predictions. What is more, we are looking forward to more contributions in topics such as quantitative syntax analysis, linguistic complex systems/networks analysis, which are research focuses in QL today.

The Russians have contributed a lot to the development of QL as well as computational linguistics, for example, the most famous Markov Chain extensively used and developed in NLP practices, the Piotrowski-Altmann law in QL (one of the three main laws in QL), as well as the precise literary studies by B. I. Yarho that can be dated back to the early 20th century. Now, by applying the state-of-art techniques to Russian QL studies, this volume will promote developments in all these fields. I think this will be of great interest to graduate students and researchers in the area of quantitative and Slavic linguistics, both outside and inside Russia.

References

- Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin and Boston: De Gruyter Mouton.
- Liu, H. (2010). Review of Kelih, Emmerich (2008) *Geschichte der Anwendung quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft (History of the application of quantitative methods in Russian linguistics and literature)*. Hamburg: Kovač. *Journal of Quantitative Linguistics*, 17(4): 365-370.
- Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2), 251–256.

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

1. U. Strauss, F. Fan, G. Altmann, *Problems in Quantitative Linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen*. 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics 2*. 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics*. 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G. Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl*. 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro*. 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp.
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in Quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3*. 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4*. 2014, VI + 148 pp.
15. K.-H. Best, E. Kelih (Hrsg.), *Entlehnungen und Fremdwörter: Quantitative Aspekte*. 2014, IV + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, *Unified modeling of length in language*. 2014. III + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical approaches to text and language analysis*. 2014, IV + 230 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA. Quantitative Index Text Analyzer*. 2014, IV + 106 pp.
19. K.-H. Best (Hrsg.), *Studies zur Geschichte der Quantitativen Linguistik. Band 1*. 2015, III + 159 pp.
20. P. Zörnig et al., *Descriptiveness, activity and nominality in formalized text sequences*. 2015, IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015, III+146 pp.
22. P. Zörnig et al. *Positional occurrences in texts: Weighted Consensus Strings*. 2016. II+179 pp.

23. E. Kelih, E. Knight, J. Mačutek, A. Wilson (eds.), *Issues in Quantitative Linguistics Vol 4*. 2016, 287 pp.
24. J. Léon, S. Loiseau (eds). *History of Quantitative Linguistics in France*. 2016, 232 pp.
25. K.-H. Best, O. Rottmann, *Quantitative Linguistics, an Invitation*. 2017, V+171 pp.
26. M. Lupea, M. Rukk, I.-I. Popescu, G. Altmann, *Some Properties of Rhyme*. 2017, VI+125 pp.
27. G. Altmann, *Unified Modeling of Diversification in Language*. 2018, VIII+119 pp.
28. E. Kelih, G. Altmann, *Problems in Quantitative Linguistics, Vol. 6*. 2018, IX+118 pp.