Glottometrics 49
2020

RAM-Verlag

ISSN 1617-8351
e-ISSN 2625-8226

Glottometrics

(Open Access)

Indexed in ESCI by Clarivate Analytics 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

Editorial and Peer Review Process

Glottometrics is a peer-reviewed scientific journal with a rigorous editorial screening and assessment process made up of several stages**:** 1. First Check, 2. Editorial Review, 3. Peer Review, 4. Final Decision. For detailed information please see: (https://www.ram-verlag.eu/journals-e-journals/glottometrics/editorial-and-peer-review-process/)

Orders for CD-ROM or printed copies to: RAM-Verlag@t-online.de
Free PDF-Download: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 49 (2020), Lüdenscheid: RAM-Verlag, 2020. 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 49 (2020)
online/ e-version ISSN 2625-8226 (print version ISSN 1617-8351)

Contents

Glottometrics 49, 2020, 1–12

Nominal vs. Adjectival Adnominals in Russian Fiction: Relationship and Distribution

Sergey Andreev[1]

Abstract. The article studies the relationship between adjectival and nominal adnominals (nouns in attributive function) in Russian prose fiction. The corpora include the works of six Russian female writers whose novels represent two different genres – literary fiction (belles-lettres fiction) and genre fiction (entertaining fiction). The results obtained demonstrate that all the authors follow similar implicit rules of setting the same relationship between the two classes of adnominals, irrespective of the genre and the period of the writers' creative activity. The Zipf-Alekseev function proved to fit well the distribution of distances between adjectival adnominals in the texts. The counts of distances between them corroborated Skinner's hypothesis.

Keywords: distribution, adjectival and nominal adnominals, Zipf-Alekseev function.

Attributes (adnominals), the main means of description, playing a highly important role in elaborating topics, are characterized by an important feature – in verbal syntactic structures, their syntactic positions are not obligatory in most cases, and thus are highly optional, depending on the author's inclinations, literary taste, and may serve as an explicit criterion of the peculiarities of authors' styles.

The latter stimulated research aimed at finding out the level of freedom of the authors in using adnominals, their frequency, proportions of different types and patterns. A number of aspects related to the regularities of the use of adnominals and their distributions have been investigated on the material of different languages (Köhler, Altmann, 2014; Altmann, 2015; Andreev, Popescu, Altmann, 2017a; Místecký, 2019).

Depending on the part of speech, one can single out in Russian (as well as in many other languages) two main classes of adnominals – adjectival and nominal ones. Adjectival adnominals (A-ADs) in Russian include the following types:

A – adjective ("krasivaya rosa" – *a beautiful rose*);
AY – adjectival phrase ("ves'ma trudnoye zadaniye" – *a highly difficult task*);
PTA – adjectivized participle ("igral'nye karty" – *playing cards*);
DETF – demonstrative pronoun ("eta kniga" – *this book*);
DETN – negative pronoun ("nikakaya rabota" – *no work*);
DETH – indefinite pronoun ("kakaya-to kniga" – *some book*);
DETQ – qualifying pronoun ("vse knigi" – *all books*);

[1] Smolensk State University, Przhevalsky str. 4, Smolensk 214000, RF, e-mail: smol.an@mail.ru.

DETS – possessive pronoun ("yego drug" – *his friend*);
DETV – relative pronoun ("ya sprosil, kakoy knigi net" – *I asked which book was missing*);
DETW – interrogative pronoun ("kakoy knigi net?" – *which book is missing?*).

Inclusion of the above-mentioned types of pronouns into the category of adjectival attributes is based on their morphological, semantic, and syntactic features similar to those of accrual adjectives in Russian, which is why they are often called pronouns-adjectives (Shvedova 1980). It should be underlined that at a deeper stage of classification, they fall into an independent class of determiners. Comparison of determiners and adjectives in their attributive function on the grounds of the data-base of Czech sonnets of the 19th and 20th centuries was carried out by M. Místecký and brought about interesting and important results of the relations between these two classes (Místecký, 2019).

The other class of adnominals includes those which are expressed by a noun. In Russian, their structures are as follows:

G – genitive case ("kniga rasskazov" – *a book of short stories*);
PR – prepositional pattern ("kniga dlya detey" – *a book for children*);
AP – apposition ("Neznakomets, chelovek srednego vozrasta, podoshel ko mne" – *The stranger, a middle-aged man, came up to me*; "kapitan Smollett" – *Captain Smollett*);
N-Case – instrumental and dative cases ("Ocharovaniye knigoy" – *fascination with the book*; "pis'mo drugu" – *letter to a friend*).

The study of the relationship between adjectival and nominal adnominals is usually limited to their two, most frequent types: type A and type G (genitive construction) (Andreev, Místecký, Altmann, 2018: 45–50). In the present study, we set the task to analyze complete sets of types in adjectival and nominal classes.

Though both A-ADs and N-ADs participate in the description of the fiction world, they nevertheless display distinct differences, which consist in the manner how description is realized. If A-ADs give a direct, immediate, and to some extent straightforward description of a theme, nominal attributes combine at least two different basic functions – first of all, they denote so so-called fiction (poetic) motifs[2] (objects, notions), and only secondly exercise description. This dual nature is especially noticeable in some cases which the following examples can demonstrate.

(1) The book [1] of my (DET-Adj) friend (G) [1].
(2) Light (A) unevenness [2] in his (DETS) gait (PR) [2] of a brave (A) soldier (G) [1] (Galbraith).

In (1), the noun "friend" is modified by the possessive pronoun "my" and at the same time, it is a modifier (G) of the other noun, "book". In (2), the descriptive pattern is more complicated – a number of nouns are modified by adnominals (adjectival and nominal) and, at the same time, they realize the adnominal function of a modifier. Numbers in square brackets show the adnominal valence of the modified nouns, i.e. how many adnominals modify the given

[2] The term "motif" here is used as a literary one, meaning the smallest (minimal) plot-forming unit" (Gasparov, 1997). In quantitative linguistics, it is now used in a different meaning denoting a sequence of elements, organized according to the principles of non-descending quantity of a given feature, and was introduced by R. Köhler (Köhler, 2008; Köhler, Naumann, 2008; Köhler, Naumann, 2016).

noun. The adnominal patterns in (2) are: “light unevenness”; “unevenness in gait”; “his gait”; “gait of a soldier”; “brave soldier”.

In both examples, noun adnominals retain their nominal features, which are emphasized by the fact that they themselves are modified by attributes, but at the same time realize a descriptive function as adnominals.

During the investigation of the relationship of the two types of description, the following questions arise. What is the relationship of A-Ads and N-ADs in different works of the same author and in the works of different authors? Are the proportions constant for the same individual, or do they change over time? Is there any stable proportion of A-Ads and N-Ads in speech in general? Is there any order in their distribution?

To address these questions, the data-base which included 6 feminine Russian authors (V. Tokareva, T. Tolstaya, L. Ulitskaya, A. Marinina, T. Ustinova, and T. Polyakova) was organized. The choice was motivated by the following reasons. (1) All of the authors are of the same gender, which excludes or minimizes possible differences of style due to the gender factor. (2) All authors are very popular among the readers of different literary tastes, which presupposes that their style and manner of description are accepted by public at large. (3) The works represent different stages of the creative activity of the authors. They include one of the first, one of the latest novels, and the works written by each author during the intermediate period. The list of the authors and their works is given in the appendix. (4) The genres of their novels are rather different: three first authors are writers of the so-called belles-lettres style, the last three belong to the sphere of entertaining fiction (more exactly – detective literature). In the first case, the works are usually attributed to “literary fiction”, in the second case – to “genre fiction”.

Each author is represented by five samples of 1,000 words from 5 books. All of them were taken from the beginning of the novels.

To assess the relationship of A-type (adjectival attributes) and genitive constructions, the coefficient of attributiveness was introduced in Andreev, Místecký, Altmann (2018: 45–46), which is similar to Busemann’s coefficient (Altmann, 2015) and the formula of which is:

$$(1) \qquad T = \frac{A}{A+N},$$

where T is the coefficient of attributiveness, A – all the attributes (adjectival adnominals), N – all the nominal attributes.

The coefficient values can vary between 0 and 1. High values of this coefficient ($T > 0.5$) show that A-ADs play a more important role in description, low values of the coefficient ($T < 0.5$) indicate the predominance of N-ADs in the style of the author.

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

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

The coefficient is statistically significant with 1 degree of freedom and $p < 0.05$ if $\chi^2 > 3.84$.

In this study, we shall also use this coefficient. The results of the analysis are shown in Table 1.

Table 1
T-coefficient and Chi-square

Text	A-ADs	N-ADs	T-coef.	Chi-square
T1	82	40	0.67	14.46
T2	70	38	0.65	9.48
T3	75	32	0.70	17.28
T4	76	45	0.63	7.94
T5	107	58	0.65	14.55
T6	161	50	0.76	58.39
T7	134	67	0.67	22.33
T8	135	55	0.71	33.68
T9	104	41	0.72	27.37
T10	121	48	0.72	31.53
T11	97	53	0.65	12.91
T12	211	76	0.74	63.50
T13	205	89	0.70	45.77
T14	174	84	0.67	31.40
T15	137	51	0.73	39.34
T16	91	50	0.65	11.92
T17	114	48	0.70	26.89
T18	125	44	0.74	38.82
T19	60	40	0.60	4.00
T20	126	53	0.70	29.77
T21	128	52	0.71	32.09
T22	96	53	0.64	12.41
T23	86	33	0.72	23.61
T24	101	43	0.70	23.36
T25	113	60	0.65	16.24
T26	87	41	0.68	16.53
T27	93	22	0.81	43.83
T28	116	54	0.68	22.61
T29	105	33	0.76	37.57
T30	81	36	0.69	17.31

As seen from the table, all the values of T-coefficient are statistically significant. The attributive style is observed in all cases, but the range over which this coefficient varies in these texts is 0.6 – 0.81. This adjectival priority was to be expected as a straightforward strategy of description, but the difference between low and high values of T-coefficient in various novels should be recognized as rather substantial. This fact points out to certain differences of the visualization of the fiction world and raises the question of stability of the manner of depicting such a fiction world.

As has been mentioned above, each author in the present data-base is represented by five works. To analyze the variability of the A-ADs vs. N-ADs relations in different works of the same author, the coefficient of variation was used

$$V = \frac{\sigma}{M} * 100\% , \tag{3}$$

where V is the coefficient of variation, σ is the standard deviation, and M is the mean. The lower the coefficient is, the lower the level of dispersion is. The results are given in Table 2.

Table 2
Coefficients of variation of the values of T-coefficient

Author	Genre	Coefficient of variation (%)
Tokareva	Literary fiction	8.19
Tolstaya	Literary fiction	11.97
Ulitskaya	Literary fiction	12.22
Marinina	Genre fiction	17.23
Ustinova	Genre fiction	11.31
Polyakova	Genre fiction	20.90

The coefficient shows a rather small variability for all the authors. The lowest variability is observed among the works of Tokareva (the group of “high style” literary fiction), and the highest variability is demonstrated in Polyakova’s novels (“entertaining” genre fiction). On the whole, belles-lettres fiction authors have lower variability than the authors of entertaining fiction, but this difference is very small and in one case (Usinova) is not found altogether.

Since the variability is rather small, it makes sense to establish the overall index of attributiveness for each author.

Table 3 contains mean values of nominal and adjectival adnominals in the works of each author and the corresponding figures of T-coefficient.

Table 3
Mean values of T-coefficient for each author

Author	N-ADs mean	A-ADs mean	T-coefficient	Chi-square
Literary fiction				
Tokareva	42.6	82	0.66	12.74
Tolstaya	52.2	131	0.72	34.66
Ulitskaya	70.6	164.8	0.70	38.58
Total	*55.1*	*125.9*	*0.70*	*27.68*

Genre fiction (detective stories)				
Marinina	47	103.2	0.68	22.28
Ustinova	48.2	104.8	0.69	21.54
Polyakova	37.2	96.4	0.72	27.57
Total	*44.1*	*101.5*	*0.70*	*22.58*

In all these cases, the results are statistically significant and demonstrate nearly the same mean scores, irrespective of the genre or date. This is to some extent unexpected, as one might suppose that literary fiction possesses a more elaborate and less direct style of depicting plot motifs.

Speaking of stability of the type of description over time, it is possible to carry out research examining the relationship between the date of writing a novel and its description type. Table 4 contains the dates when the novels were written and gives the corresponding values of T-coefficient.

Table 4

Relations of the date of writing and the type of description

Tokareva			Tolstaya			Ulitskaya		
Text	Date	T-coef.	Text	Date	T-coef.	Text	Date	T-coef.
T1	1991	0.67	T6	1987	0.76	T11	1975	0.65
T2	1994	0.65	T7	1998	0.67	T12	1992	0.74
T3	2004	0.70	T8	2000	0.71	T13	1996	0.70
T4	2015	0.63	T9	2007	0.72	T14	2003	0.67
T5	2018	0.65	T10	2015	0.72	T15	2010	0.73

Marinina			Ustinova			Polyakova		
Text	Date	T-coef.	Text	Date	T-coef.	Text	Date	T-coef.
T16	1993	0.65	T21	1997	0.71	T26	2002	0.68
T17	1996	0.70	T22	2000	0.64	T27	2005	0.81
T18	2001	0.74	T23	2004	0.72	T28	2009	0.68
T19	2010	0.60	T24	2010	0.70	T29	2012	0.76
T20	2017	0.70	T25	2017	0.65	T30	2016	0.69

There does not seem to be any correlation between the date of writing and the extent to which the nominal style intensifies or decreases. These three tests have demonstrated that the relationship between the two strategies of description is rather stable.

One more aspect of exploring the relationship between adjectival and nominal adnominals is to analyze if there is any order in which these attributes, namely A-ADs, are arranged on the syntagmatic axis in the text in relation to N-ADs. In other words, one will be able to find out whether there is any order in the changeability of these two types of description over the text. Technically, this question may be solved by different methods, such as runs, measuring the number of homogeneous sequences of attributes of the same type (Andreev, Místecký, Altmann, 2018: 50), repeat-rate, which shows the concentration of elements (Altmann, Köhler, 2015), and some others.

In the present article, we chose a different method – the one of establishing distances between A-ADs in relation to N-ADs. This will also help us to check whether Skinner's hypothesis is observed here (Andreev, Popescu, Altmann, 2017b). According to this hypothesis, similar elements occur closer to one another, i.e. have smaller distances from one another in speech (Skinner, 1941).

As an example of establishing such distances in our study, let us take the first sentence from T4. After all adnominals were marked in this sentence, they formed the following sequence:

A, PR, A, A, APR, DETS, PR,

and after transforming them into adjectival (A-AD) and nominal (N-AD) classes, we get A-AD, N-AD, A-AD, A-AD, N-AD, A-AD, N-AD.

Between the first and the second adjectival adnominals, there is one nominal N-AD. This is why we count the distance as one ($D = 1$). The second adjectival adnominal is followed immediately by the third one (distance $D = 0$), the forth occurs after one nominal adnominal ($D = 1$), etc.

After counting all the distances between adjectival adnominals in the texts, these distances were ranked in descending order, and the Zipf-Alekseev function was used to fit their distribution (Hřebíček, 2002):

$$f_x = f_1 x^{a+b*\ln x}, \tag{4}$$

where f_1 is the maximum frequency of the most numerous distance, a and b – parameters, x – the frequency of the given distance.

The results are given in Table 5.

Table 5

Fitting the Zipf-Alekseev function to the distribution of lengths of distances in 30 novels (D – distances, Em – empirical data, Th – theoretical values counted on the basis of the function)

Tokareva														
T1			T2			T3			T4			T5		
D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th
0	12	12.00	0	15	15.00	0	10	10.00	0	19	19.00	0	17	17.00
1	8	8.57	1	7	7.59	1	5	6.11	1	10	10.09	1	12	12.82
3	6	5.94	3	5	4.85	2	5	4.18	2	6	5.59	2	10	8.14
2	5	4.24	4	5	3.45	4	4	3.07	3	3	3.34	5	4	5.18
4	4	3.12	2	2	2.61	3	3	2.36	9	2	2.11	3	3	3.38
5	2	2.37	5	2	2.07	5	1	1.88	4	1	1.40	4	3	2.28
8	1	1.83	11	1	1.69	6	1	1.53	5	1	0.97	7	1	1.58
10	1	1.45				7	1	1.27	6	1	0.69			

						8	1	1.07	7	1	0.50			
R^2 = 0.9731 a = -0.219 b = -0.384			R^2 = 0.9736 a = -0.904 b = -0.113			R^2 = 0.9405 a = -0.570 b = -0.203			R^2 = 0.9973 a = -0.572 b = -0.493			R^2 = 0.9690 a = 0.044 b = -0.650		

Tolstaya														
T6			T7			T8			T9			T10		
D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th
0	13	13.00	0	21	21.00	0	16	16.00	0	17	17.00	1	18	18.00
1	11	10.10	1	19	16.96	1	15	14.14	1	8	8.08	0	10	10.59
2	6	7.19	4	7	10.40	2	8	9.85	3	5	4.69	2	8	6.77
4	5	5.18	2	6	6.25	3	8	6.70	2	3	3.03	5	4	4.63
3	4	3.82	3	5	3.83	4	4	4.62	6	2	2.11	3	3	3.33
6	3	2.89	5	5	2.42	5	4	3.25	4	1	1.53	4	3	2.49
11	3	2.23	8	1	1.57	7	3	2.33	7	1	1.16	10	2	1.91
7	2	1.75	9	1	1.05	6	1	1.71	12	1	0.90	7	1	1.50
5	1	1.40	11	1	0.72	8	1	1.27	14	1	0.71			
14	1	1.13							18	1	0.58			
R^2 = 0.9793 a = -0.064 b = -0.433			R^2 = 0.9469 a = 0.258 b = -0.817			R^2 = 0.9689 a = 0.270 b = -0.648			R^2 = 0.9971 a = -0.902 b = -0.246			R^2 = 0.9872 a = -0.551 b = -0.309		

Ulitskaya														
T11			T12			T13			T14			T15		
D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th
1	17	17.00	0	21	21.00	2	23	23.00	1	28	28.00	0	13	13.00
0	14	13.29	1	18	16.17	0	21	22.09	0	17	18.68	2	12	12.79
2	7	8.15	4	8	10.92	1	15	15.29	2	13	12.05	1	11	8.51
4	5	4.94	2	7	7.42	3	12	10.09	3	11	8.06	3	3	5.33
3	3	3.07	3	7	5.17	4	10	6.70	4	5	5.61	4	3	3.35
5	3	1.96	6	5	3.70	5	2	4.52	5	3	4.03	5	2	2.14
6	1	1.29	8	2	2.71	6	1	3.12	7	2	2.98	8	2	1.40
7	1	0.87	12	2	2.03	7	1	2.19	8	2	2.25	6	1	0.94
8	1	0.60	5	1	1.54	8	1	1.57	6	1	1.73	7	1	0.64
			7	1	1.20	9	1	1.15	10	1	1.36	9	1	0.45
			10	1	0.94	16	1	0.85	7	1	0.50	14	1	0.32
			11	1	0.75									
			13	1	0.60									

$R^2 = 0.9886$
a = 0.181
b = -0.773
	$R^2 = 0.9659$
	a = -0.003
	b = -0.539
		$R^2 = 0.9617$
		a = 0.478
		b = -0.773
			$R^2 = 0.9784$
			a = -0.269
			b = -0.454
				$R^2 = 0.9423$
				a = 0.598
				b = -0.895

Marinina														
T16			T17			T18			T19			T20		
D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th
0	36	36.00	0	21	21.00	0	18	18.00	0	15	15.00	1	14	14.00
1	21	22.34	1	8	8.98	1	6	6.72	1	11	10.91	0	10	11.37
2	15	13.53	2	6	5.28	3	5	4.02	2	6	5.77	3	9	8.18
4	10	8.58	4	5	3.57	8	3	2.87	3	2	3.01	2	7	5.91
3	5	5.69	3	3	2.62	2	2	2.25	7	2	1.61	4	5	4.35
5	3	3.92	6	1	2.02	4	2	1.86	4	1	0.90	6	3	3.28
6	3	2.79	8	1	1.61	5	2	1.59	5	1	0.52	7	2	2.52
8	1	2.04	11	1	1.32	6	1	1.40	6	1	0.31	5	1	1.97
9	1	1.52	23	1	1.11	7	1	1.26				9	1	1.56
11	1	1.16				10	1	1.14						
23	1	0.90				12	1	1.05						
						14	1	0.98						
$R^2 = 0.9930$ a = -0.342 b = -0.499			$R^2 = 0.9845$ a = -1.174 b = -0.075			$R^2 = 0.9922$ a = -0.517 b = 0.139			$R^2 = 0.9905$ a = 0.242 b = -0.101			$R^2 = 0.9654$ a = 0.021 b = -0.464		

Ustinova														
T21			T22			T23			T24			T25		
D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th
0	18	18.00	0	19	19.00	0	11	11.00	0	13	13.00	0	19	19.00
1	18	17.70	1	12	12.53	1	7	6.75	1	9	9.37	1	12	12.90
2	10	9.91	2	9	7.42	4	4	4.29	2	7	6.41	2	10	8.02
4	4	5.09	3	3	4.51	3	3	2.88	4	4	4.50	3	4	5.11
3	2	2.62	5	3	2.85	5	2	2.03	5	4	3.25	4	3	3.39
5	2	1.39	4	2	1.87	2	1	1.48	3	2	2.42	5	3	2.32
6	2	0.76	6	1	1.27	6	1	1.11	8	2	1.85	7	1	1.64
8	2	0.43	7	1	0.89	7	1	0.86	12	1	1.44	9	1	1.18
7	1	0.25	10	1	0.63	11	1	0.67				12	1	0.87
9	1	0.15				14	1	0.54						
13	1	0.09												

$R^2 = 0.9816$
a = 0.863
b = -1.280
	$R^2 = 0.9835$
	a = -0.163
	b = -0.630
		$R^2 = 0.9926$
		a = -0.443
		b = -0.378
			$R^2 = 0.9859$
			a = -0.179
			b = -0.424
				$R^2 = 0.9778$
				a = -0.170
				b = -0.560

Polyakova														
T26			T27			T28			T29			T30		
D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th	D	Em	Th
0	14	14.00	1	8	8.00	0	16	16.00	0	7	7.00	0	11	11.00
1	14	13.76	4	3	3.22	1	13	12.04	3	7	6.54	2	10	9.88
3	3	4.05	0	2	2.12	2	6	8.10	2	5	5.04	1	6	5.71
2	2	0.99	2	2	1.66	3	6	5.51	4	3	3.79	3	2	3.11
5	2	0.24	5	2	1.42	4	5	3.86	6	3	2.87	4	2	1.71
7	2	0.06	8	1	1.27	5	4	2.77	1	2	2.21	5	1	0.97
4	1	0.02	11	1	1.17	6	1	2.04	8	2	1.72	6	1	0.57
9	1	0.01	14	1	1.10	9	1	1.54	9	2	1.36	10	1	0.34
17	1	0.00	20	1	1.05	10	1	1.18	5	1	1.09	15	1	0.21
$R^2 = 0.9497$ a = 1.861 b = -2.722			$R^2 = 0.9842$ a = -1.493 b = 0.259			$R^2 = 0.9573$ a = -0.052 b = -0.517			$R^2 = 0.9655$ a = -0.246 b = -0.497			$R^2 = 0.9801$ a = 0.602 b = -1.092		

Skinner's hypothesis on the whole holds, because short distances dominate. Still in some cases, this rule is less obvious. Thus in three novels by Ulitskaya (T11, T13, and T14), the biggest frequency is observed not for the shortest distance (0), but for distances 1 or 2. Such neutralization of Skinner's law is also observed in one text by Tolstaya (T10) and Marinina (T20). In several texts, the differences between the first three distance ranks are very small (T8, T15, T21, T26, and T30).

Judging by these results, the main opposition in the manner of description is observed between Tokareva and Ulitskaya, both authors belonging to the class of literary fiction. Marinina and Ustinova, highly popular among the readers of detective-stories and using mostly colloquial language, demonstrate Skinner's tendency much better than two literary fiction authors (Tolstaya and Ulitskaya).

The results also demonstrate that the Zipf-Alekseev function fits very well the distribution of distances between the adjectival adnominals, which corroborates an order in choosing different types of description.

Overall, the study of the relationship between two main strategies of description of the world of fiction by modern female Russian authors has demonstrated that the authors preferred direct adjectival description of the fiction world over the nominal strategy. The authors of belles-lettres style (literary fiction) in some works resort to this adjectival description a little stronger than the authors of mass entertaining literature (genre fiction), but on the whole, they do not differ in this respect very much.

Judging by the values of the coefficient of attributiveness, the ratio between these two strategies is approximately 3:1 and remains more or less constant over time for each author, regardless of the period of creative activity of the author or genre in which the she is writing. This means that each author's style – as regards the relations between nominal and adjectival strategies – is implicitly controlled by some common trend or similar pattern of combining the two types of description.

The Zipf-Alekseev function fits well the distribution of the distances of adjectival adnominals, showing that it is governed by some general rules.

The results of this study should be tested on a broader range of material, including the authors of different genres, genders, literary schools, and writings in various languages.

References

Altmann G. (2015). *Problems in Quantitative Linguistics 5*. Lüdenscheid: RAM-Verlag.

Altmann, G., Köhler, R. (2015). *Forms and Degrees of Repetitions in Texts. Detection and Analysis*. Berlin / Munich / Boston: de Gruyter Mouton.

Andreev, S., Místecký M., Altmann G. (2018). *Sonnets: Quantitative Inquiries. Studies in Quantative Linguistics 29*. Lüdenscheid: RAM-Verlag, 2018.

Andreev, S., Popescu I.-I., Altmann G. (2017a). Some properties of adnominals in Russian texts. *Glottometrics* 38, 77–106.

Andreev, S., Popescu I.-I., Altmann G. (2017b). Skinner's hypothesis applied to Russian adnominals. *Glottometrics* 36, 32–69.

Gasparov M. L. (1997). "Snova tuchy nado mnoyu..." Metodika analyza. *Izbranniye trudy. T. II. O styhah.* Moscow: Yaziki Russkoy kul'tury, 9–20.

Hřebíček, L. (2002). Zipf's Law and Text. *Glottometrics* 3, 27–38.

Köhler, R., Altmann G. (2014). *Problems in Quantitative Linguistics 4*. Lüdenscheid: RAM-Verlag.

Köhler, R. (2008). Sequences of linguistic quantities. Report on a new unit of investigation. *Glottotheory* 1(1),115–119.

Köhler, R. and Naumann S. (2008). Quantitative text analysis using L-, F- and T-segments. *Data Analysis. Machine Learning and Applications*. Berlin / Heidelberg: Springer, 2008, 637–646.

Köhler, R., Naumann S. (2016). Syntactic Text Characterisation Using Linguistic S-Motifs. *Glottometrics* 34, 1–8.

Místecký, M. Five Ways of Investigating Adnominals in Czech Sonnets of the 19th and 20th Centuries. *Glottotheory* 9(2), 173–200.

Skinner, B. F. (1941). A quantitative estimate of certain types of sound-patterning in poetry. *The American Journal of Psychology* 54, 64–79.

Shvedova, N. Yu. (1980; chief ed.). *Russkaya grammatika* [Russian Grammar]. Moscow: Akademiya Nauk, Nauka.

Zörnig, P. (2010). Statistical simulation and the distribution of distances between identical elements in a random sequence. *Computational Statistics & Data Analysis* 54, 2317–2327.

Zörnig, P. (2013a). A continuous model for the distances between coextensive words in a text. *Glottometrics* 25, 54–68.

Zörnig, P. (2013b). Distances between words of equal length in a text. In: Köhler, R. Altmann, G. (eds.), *Issues in Quantitative Linguistics 3. To honour Karl-Heinz Best on the occasion of his 70th birthday*. Lüdenscheid: RAM-Verlag, 117–129.

Appendix

Author	Title	Number	Year
V. Tokareva	Skazat’ – ne skazat’	T1	1991
	Den’ bez vranya	T2	1994
	Ptitsa schastya	T3	2004
	«Samyj schastlivyj den’ (Rasskaz akseleratki)»	T4	2015
	Nu i chto?	T5	2018
T. Tolstaya	Milaya Shura	T6	1987
	90-60-90	T7	1998
	Nozhki	T8	2000
	Reka	T9	2007
	Shodit v magazin	T10	2015
L. Ulitskaya	Lestnica yakova	T11	1975
	Sonechka	T12	1992
	Medeya i ee deti	T13	1996
	Iskrenne vash Shurik	T14	2003
	Zelenyj shater	T15	2010
A. Marinina	Igra na chuzhom pole	T16	1993
	Stilist	T17	1996
	Zakon trex otricanij	T18	2001
	Lichnye motivy	T19	2010
	Angely na ldu ne vyzhivayut	T20	2017
T. Ustinova	Moj general	T21	2002
	Dom-fantom v pridanoe	T22	2005
	Na odnom dyxanii	T23	2009
	Srazu posle sotvoreniya mira	T24	2012
	Zhdite neozhidannogo	T25	2016
T. Polyakova	Dengi dlya killera	T26	1997
	Baryshnya i xuligan	T27	2000
	Bochka no-shpy i lozhka yada	T28	2004
	Moe vtoroe ya	T29	2010
	Zmej-soblaznitel’	T30	2017

Glottometrics 49, 2020, 13–31

Polysemy of Rhyme Words: A Case Study of Two Slovak Poems

Natália Kolenčíková[1]
Michal Místecký[2]
Gabriel Altmann

Abstract. The goal of the article is to count and measure polysemy of rhyme words. The corpus of the research will include two poems by Andrej Sládkovič, a Slovak Romantic poet – *Marína* (primarily) and *Morava* (secondarily). The research is carried out on the basis of rank-frequency distributions, which are very well capturable by the expontential function. This proves the tendency of using one-meaning words in rhymes much more than those with multiple meanings.

Keywords: Andrej Sládkovič, exponential function, Marína, Morava, polysemy, rank-frequency distribution, rhyme words, Slovak.

1. Introduction and Corpus

The rhyme words in rhymed poetry have a special task – they not only serve the phonic rhyming, but also contain information the complexity of which may be measured. Rhyme words have a special position, and tend to carry the climax of a line.

Polysemy is a complex phenomenon, the investigation of which meets many an obstacle. One must decide whether, e.g., *biely, belosť, belieť sa, obelieť, zbelieť* in Slovak or *white*, *whiteness*, *whiten*, *make something white*, or *turn white* in English are to be taken separately, or together. In some languages, one uses affixes, separate words, compounds, etc., in order to express the state, the quality or the activity, etc. The decision depends on the aim of research, on the "school" of the researcher, on the dictionaries, on the given language. All this taken into account, the precise rules for polysemy of rhyme words will be defined in the section to come.

As to the corpus, we try to capture and measure the polysemy of rhyme words in Slovak poems *Marína* and *Morava*, written by Andrej Sládkovič. *Marína* is the most famous and the most frequently published Slovak composition. The poem, which was created like a reflection of the author's real love experience, has 291 stanzas, and was published in 1846 for the first time. Its primarily intimate-meditative nature is mixed with motifs of beauty, youth, nation, Slovak country, sense of poetry, etc. In *Morava* (1848), which consists of 40 lines only, the author leaves his personal space and focuses mainly on the period issues. What dominates in it is a reflexive and meditative style, and via natural symbols, the position of a nation and the historical origin of its independence are emphasised. We use the versions from *Dielo 1* (1961), edited by C. Kraus. It is to be noted that the polysemy of many words was quite different two

[1] Ľudovít Štúr Institute of Linguistics, Slovak Academy of Science; e-mail: natalia.kolencikova@juls.savba.sk.
[2] University of Ostrava, Ostrava, the Czech Republic; e-mail: mmistecky@seznam.cz.

hundred years ago, but taking into account the quality of Slovak lexicography in the last decades, we adhere to the present-day state of Slovak mainly.[3]

2. Methods

In order to capture the meanings to as much detail as possible, we make use of several dictionaries. Primarily, we stem from the codificative *Krátky slovník slovenského jazyka* (A Short Dictionary of the Slovak Language, 2003; henceforth KSSJ), which is a one-volume lexicographical handbook listing the most used vocabulary of standard Slovak (60,000 words). If an expression is not found in this dictionary, we search for it in the most updated, yet still unifinished *Slovník súčasného slovenského jazyka A–G*, *H–L*, and *M–N* (A Dictionary of the Contemporary Slovak Language A–G, H–L, and M–N – 2006, 2011, 2015; henceforth SSSJ); the three volumes, which have been published so far, contain more than 45,000 entries. This indicates, in comparison with the previous dictionary, the more detailed elaboration of the entries. As to the words that were found in KSSJ, but not in SSSJ, there are, for instance, *bedovať* ("cry over"), *bieloružový* ("white-pink"), *čerstvota* ("freshness"), *horovať sa* ("to mate"), *mámenie* ("delusion"), *nezakrytý* ("uncovered"). If this word is missing even from this dictionary, we use the six-volume *Slovník slovenského jazyka* (A Dictionary of the Slovak Language, 1959–68; henceforth SSJ), which contains more than 100,000 entries, this considerably surpassing the scope of KSSJ. Given this, the incompletion of SSSJ, and the fact that the last volume of SSJ was published more than half a century ago, it is understandable that it is in this dictionary that we find almost 90 expressions unincluded in any other lexicographical book. These are, for instance, *búra* ("storm"), *pokonný* ("final"), *sprostiť sa* ("free somebody from something"), *vojvodiť* ("rule over"), *všesvet* ("universe"), *zalkať* ("start to mourn"), *žiadúci* ("needed"), etc. Despite the effort to found the number of meanings of the rhyme words on the grounds of the dictionaries, we do meet expressions that are unattested in any of the books (e.g., *divodivý* ["most wondrous"], *hromplesk* ["clap of thunder"], *obleva* ["thaw"], *poľúbok* ["kiss"], *praprapraotec* ["great-great-grandfather"], *výstrojiť* ["dress up"]). In such case, we assign to the word one meaning only, namely the one in which it is used in the analysed poem. It is to be noted, though, that this concerns mostly poetic lexical devices, or the ones belonging to the author's idiolect.

Investigating the number of meanings of rhyme words in *Marína*, we observe the following rules:

– If a rhyme word is poetically or otherwise modified (phonetically or morphologically), and it is not attested in this form in the dictionaries, but its lexicographical variant is contextually obvious, we assign to the word the number of meanings of the variant (e.g., variants of "pain" – *boľast* / *bolesť*, variants of "lummox" – *mamlas* / *mamľas*, variants of "blossoming" – *zakvetlý* / *zakvitnutý*, variants of "disappear" – *zmizeť* / *zmiznúť*).

– If a rhyme word is non-standard, which means the definition of it may not be found found in a dictionary, we give the number of meanings of the standard variant that the dictionary usually refers to (e.g., variants of "breastfeed" – *kojiť* / *dojčiť*).

– If a rhyme word is a homonym, we respect it, assigning to each of the homonyms its individual number of meanings; our cases only concern words *drahý*, *milý* ("dear" vs. "a dear one", "nice" vs. "a nice one"), *mať* ("to have" vs. "mother"), and *vystaviť* ("to exhibit" vs. "to build up").

[3] As an example of the changes in this sphere – and in direct connection to the polysemy of rhyme words in *Marína* –, word "družica" can be mentioned, which, given the extralinguistic reality, cannot be, as soon as 1846, interpreted as "a body circulating around a bigger cosmic object, a satellite" (KSSJ: 141; translated by the authors), as it is stated in the present lexicographical handbooks. It is used in its first meaning as a "girl in a festive clothing at the nuptial, or funeral, etc., ceremony" (KSSJ: 141; translated by the authors).

– If a rhyme word is a proper name, we assign to it one meaning only, respecting the idea that they fulfil not only the nomination, but also the identification and differentiation functions, or other functions, too (see Blanár, 1996).

– The verbal nouns are regarded as nouns only if their nounal meaning is lexicalized, and can be captured as such; if this situation has not occurred, we treat the word as an infinitive of the given verb, and assign to it the number of meanings of the verb (e.g., “bubbling” = “to bubble”, “swaying” = “to sway”, “sighing” = “to sigh”). The same approach is adopted in case of past participles with the adjectivization potential (e.g., “loved” = “to love”, “sparkled” = “to sparkle”, “wrapped up” = “to wrap up”).

– The comparatives are given in the base forms, and we presuppose the comparison may appear in all the capturable meanings, even though this may not be true in practice (“more miserable”, “more famous”).

– We strictly distinguish the non-reflexive verb forms from the reflexive ones, which express different meanings, and do thus possess different numbers of them (e.g., *boriť* / *boriť sa* = “to ruin” / “to fall into ruin”, *kryť* / *kryť sa* = “to cover” / “to take cover”, *rozkladať* / *rozkladať sa* = “to dismantle” / “to rot”, *spojiť* / *spojiť sa* = “to put together” / “to join”, *zastrieť* / *zastrieť sa* = “to cover” / “to stop”). The same applies in cases of the verbs differing in the verbal aspect (e.g., *dostať* / *dostávať* = “to get” / “to keep getting”, *dať* / *dávať* = “to give” / “to keep giving”, *dať sa* / *dávať sa* = “to give oneself” / “to keep giving oneself”, *padať* / *padnúť* = “to keep falling” / “to fall”, *prijať* / *prijímať* = “to accept” / “to keep accepting”).

– As to the negative forms of verbs, we treat them as separate words, despite the fact that they are not captured as such by the dictionary entries. We lean on the immanent property of the modification morpheme (*ne*-, in our case) to change the basic lexical meaning of the word; and even though we assign to these forms the same number of meanings as is given in their non-negative counterparts, semantically, negation does not stretch over all of their meanings (e.g. *[ne]lúčiť sa* – “[not] to say goodbye”, *[ne]očariť* – “[not] to enchant”, *[ne]prestrašiť* – “[not] to lure”, *[ne]rozdvojiť* – “[not] to halve”, *[ne]zasmútiť* – “[not] to sadden”). Not even in this case, the language practice does not have to correspond to the aforementioned principles absolutely.

For the sake of the example, consider the first strophe of *Marína*, in which we have the following rhyme words:

krása, nadšený, ohlas, zavrieť, svietiť, letieť, pohýnať, život, jednota, Marína.

Respecting all of the aforementioned principles, we have obtained the results presented in the following table.

Table 1
The numbers of the meanings of the rhyme-words in strophe 1 of *Marína*

Rhyme-word	*krása*	*nadšený*	*ohlas*	*zavrieť*	*svietiť*	*letieť*	*pohýnať*	*život*	*jednota*	*Marína*
Number of meanings	2	1	3	7	4	9	5	12	3	1

We mentioned our aim is to state the frequencies of the words in the rhyme positions and the polysemies of these words. We conjecture that there are laws behind these phenomena.

3. Results

If one counts the polysemies of the individual rhyme words (see Appendix 1), one gets the results presented in Table 2. Here, we use the Zipf-Alekseev function of the formula

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

and obtain satisfactory results. There is thus a visible tendency in the distribution of meanings in *Marína*'s rhyme words.

Table 2
Polysemy of rhyme words in *Marína*

No. of Meanings	Frequency	Zipf-Alekseev
1	955	951.56
2	643	660.04
3	433	418.32
4	274	271.60
5	226	182.61
6	86	126.83
7	89	90.60
8	49	66.30
9	42	49.54
10	46	37.68
11	31	29.12
12	19	22.81
14	3	14.51
20	1	4.61

	a = -0.1511, b = -0.5434,
	c = 951.5638, R^2 = 0.9956	

As can be seen, the empirical data are not very "smooth", but this is to be anticipated. The poet does not like to use the same word in the same position because the number of rhyming words is restricted. In analytic languages, one could expect a much stronger steepness. However, even the length of the poem and the way of analysis are decisive. Further, each text can be a posteriori changed, not only by the author, but also by the editor. Preliminarily, we can accept the hypothesis that in long Slovak poems, the meanings of words in the rhyme position follow the Zipf-Alekseev function. This is notable, as generally, we assume that the words with a higher number of meanings are more frequent in texts; this is visibly not the tendency of rhyme words, as here, mostly autosemantics (i.e., one- or two-meaning units) are employed.

It would be interesting to study shorter texts in order to discover the complete mechanism into which boundary conditions (text type, author, language, historical epoch, etc.) could be inserted and interpreted. A comparison of *Marína* with other long poems would be relevant theoretically, too.

Let us consider the 40-line poem by Sládkovič, *Morava*. The meanings and the fitted Zipf-Alekseev function show that this is the relevant model for this case, too. The result is presented in Table 3 (for the data, see Appendix 2).

Table 3
Polysemy of rhyme words in *Morava*

No. of Meanings	Frequency	Zipf-Alekseev
1	19	19.03
2	9	8.81
3	5	5.30
4	4	3.60
6	1	2.01
9	2	1.07
	a = -1.0195, b = -0.1315, c = 19.0271, R^2 = 0.9902	

As can be seen, rhyme words follow a distribution that is unusual that can be found in other instances of language (see above). The numbers obtained in the present study, besides pointing at this fact, can be used in many types of comparisons (poets, literary school and periods, languages, poetry / fiction differences, etc.). Moreover, as the rhyme words with fewer meanings are considered to be semantically fuller, the distribution can also serve to prove the quality of a poet's rhymes (in the sense of their ungrammatical character). Here, the main task was to show the possible modelling and open space for research to come.

Acknowledgements

This research was supported by the public sources provided by *Slovak Arts Council* (*Fond na podporu umenia*).

References

Blanár, V. (1996). *Teória vlastného mena (Status, organizácia a fungovanie v spoločenskej komunikácii).* Bratislava: Veda.

Krátky slovník slovenského jazyka. (2003). J. Kačala – M. Pisárčiková – M. Považaj (eds.). Bratislava: Veda.

Sládkovič, A. (1961). *Dielo I.* Edited by C. Kraus. Bratislava: SVKL.

Slovník slovenského jazyka. (1959 – 1968). Š. Peciar (ed.). Bratislava: Vydavateľstvo SAV.

Slovník súčasného slovenského jazyka A – G. (2006). K. Buzássyová – A. Jarošová (eds.). Bratislava: Veda.

Slovník súčasného slovesnkého jazyka H – L. (2011). A. Jarošová – K. Buzássyová (eds.). Bratislava: Veda.

Slovník súčasného slovenského jazyka M – N. (2015). A. Jarošová (ed.). Bratislava: Veda.

Appendix 1

Frequencies and number of meanings of rhyme words in *Marína*

Word	Freq.	Meanings	Word	Freq.	Meanings	Word	Freq.	Meanings
akýsi	1	2	nestávať	1	10	spasenie	1	1
anjel	5	3	netrebný	1	1	spať	1	3
Arab	1	1	netušiť	1	1	špata	1	1
Baby-Hoľa	1	1	netúžiť	2	1	speniť sa	2	1
Barkochab	1	1	neublížiť	1	1	sperliť	1	1
báť sa	3	3	neuchytiť	2	1	spev	3	3
Batu	1	1	neusadiť sa	1	3	spievať	7	6
baviť sa	1	4	nevedieť	2	6	spiežový	1	1
bedovať	1	1	nevedomý	1	2	spínať	3	2
belieť sa	2	1	neveriť	2	6	spojiť	5	5
belosť	1	1	neverný	1	1	spojiť sa	2	4
bežať	1	5	neviazať	1	8	spomínať	7	1
bezmiestno	1	1	nevina	2	1	spomínať si	1	1
bezočivý	1	1	nevinný	3	3	spomnúť si	2	1
bezpečný	1	3	nevládať	1	1	spraviť	2	1
bieda	3	2	nevlažiť	1	1	spriateliťsa	1	2
biednejší	1	3	nevodiť sa	1	2	sprostiť sa	1	1
biedno	1	1	nevolávať	1	7	sprostý	1	3
bielo-mramor	1	1	nevravieť	1	1	spustiť	1	5
bieloružový	2	1	nevyhnutný	1	2	srdečný	3	1
biely	5	2	nezabronieť sa	1	2	srna	1	1
biť	7	4	nezahrávať	1	2	starý	1	10
blaho	5	2	nezakrytý	1	1	šťastlivý	1	1
blahý	1	1	nezasmútiť	1	1	šťastne	1	1
blankyt	1	1	nezaznieť	1	1	stať	1	2
blato	1	1	nezhniť	1	1	stáť	9	10
blažený	3	1	nezhojiť	1	1	statný	1	1
bláznovský	1	1	nezhubiť	1	1	stávať	1	10
blbotať	1	1	nezhynúť	1	1	stena	1	4
blčať	1	1	nežiadať	1	3	stesniť sa	1	1
bledý	2	3	nezjednotený	1	1	stierať	1	4
blesk	2	2	nezmilovať sa	1	1	stín	6	1
blížiť sa	1	3	neznámy	1	1	stíšiť sa	1	2
blízkosť	1	1	neznaný	1	2	stískať	1	3
blízky	1	2	neznieť	1	5	stôl	1	2
blud	5	1	nezrušiť	2	2	storaký	1	2

blúdiť	1	4	nezvratný	1	2	strach	1	2
bludný	1	3	nie	1	5	strana	1	9
blýskať	4	2	nitriansky	1	1	strap	1	1
blýskavý	1	1	nížina	1	1	strata	1	3
blyskoce sa	1	1	nízky	4	6	stratiť	3	4
Boh	5	1	noc	7	1	stráž	1	2
bohatý	2	2	nočný	1	1	strela	4	2
boj	5	2	nocoblúdivý	1	1	streliť	1	4
bôľ	1	1	nosidlo	1	1	strestať	1	1
boľasť	3	3	nosiť	3	2	stretnúť sa	3	2
bolesť	1	3	nosiť sa	2	2	strieborný	1	3
bolestno	1	1	núkať sa	1	2	stroj	1	5
boriť	1	2	nútiť	1	2	strojiť	1	2
boriť sa	1	5	oberať sa	1	1	strom	1	1
boží	2	2	obeť	2	3	struna	1	2
bozkať	1	1	obetovať sa	1	3	stud	1	2
bozkávať	3	1	objať	2	1	studnička	1	2
božskosť	2	2	objatie	1	1	stvorenie	1	3
bralo	1	1	objem	3	3	stvoriť	4	2
brána	1	3	objímať	12	1	stvrdiť	1	1
breh	1	2	oblak	4	2	sud	1	1
brod	2	1	oblek	1	1	súd	1	5
brodiť	1	1	obleva	2	1	súdiť	1	5
brodiť sa	4	1	obliekať	3	2	súdny	1	2
brojiť	1	1	obloha	1	1	šuhaj	5	1
buch	1	2	oblok	1	1	šumieť	2	2
búchať	1	4	obluda	2	2	šumný	1	1
budúci	2	1	obrana	1	3	sused	1	2
búra	2	1	obrátiť	2	3	sušiť	1	3
búriť	1	4	obraz	8	6	šušťať	1	1
bydlo	1	1	obrovský	1	1	súžiť	2	1
byť	6	7	obsadnúť	1	2	svadba	1	3
bytosť	1	2	obstarný	1	1	švárny	2	1
bývať	6	7	obstať	4	1	sväte	1	4
čakať	1	5	obstáť	2	3	svätý	9	7
čalún	1	2	obťažený	1	1	svätyňa	1	2
Canóva	1	1	obutý	1	2	svet	22	8
cárica	1	1	obzrieť sa	1	4	svetlica	1	1
čarovný	2	2	odcloniť sa	1	1	svetlo	1	4

čas	10	9	oddať sa	1	3	svetlonos	1	1
čelo	5	3	odievať sa	3	1	svetovíťaz	1	1
celý	3	3	odkliaty	1	1	svietiť	6	4
ceniť	1	3	odkročiť	1	1	svietiť sa	1	1
čerstvota	1	1	odkryť	2	3	svitanie	2	1
červenieť sa	1	2	odložiť	1	3	svitať	6	2
červený	1	3	odlúčený	1	1	svitnúť	4	2
chcieť	3	4	odmeniť	1	1	svoj	2	6
cherubín	1	1	odobrať	1	4	svojhlavý	1	1
chiméra	1	2	odpadnúť	1	5	syn	6	2
chlad	2	2	odpínať	1	1	sýtny	1	1
chladiť sa	1	3	odpor	1	3	tajný	2	2
chládok	1	2	odrieknuť sa	1	1	táto	2	5
chlieb	4	2	odroniť	1	1	ťažoba	1	1
chodievať	4	9	odstrániť	1	3	telo	4	2
chodiť	3	9	odstrieť	2	1	ten	1	7
chór	2	2	odstupovať	1	5	tento	1	5
choroba	1	4	odšumieť	1	2	tešiť	1	2
chrám	2	1	odvrátiť	1	2	tešiť sa	2	1
chudoba	1	3	odznieť	2	2	tichamilovný	1	1
chvála	2	1	ohlas	4	3	tichý	4	8
chváliť	1	2	ohlušiť	2	2	tieto	1	5
chvieť sa	1	2	ohnivý	1	5	tisíc	1	2
chýr	1	2	okamženie	3	1	tisícoraký	1	2
chytiť	1	7	oko	11	2	tíšina	3	2
čiastka	1	3	okolo	1	2	tknúť sa	1	2
čierňava	1	1	okrádať	1	1	tlieť	3	2
čiernošatý	1	1	okrasa	2	1	točiť sa	3	1
čierny	3	7	okrášliť	1	1	tok	3	4
čistota	1	3	okúsiť	1	2	tón	4	4
cit	7	3	oltár	1	1	tôňa	1	1
cítiť	5	4	omdlieť	1	1	tráva	3	3
čln	1	1	omladlý	1	1	trávička	1	3
cloniť	1	1	on	4	1	treba	1	1
človek	2	4	ona	6	1	triasť sa	1	5
cnosť	1	2	oni	1	1	troje	1	2
ctnoty	1	1	opakovať	1	5	trpkosť	1	1
Čud	1	1	opera	1	2	túha	1	1
čudeso	1	1	opona	4	1	túliť	2	1

čudovať sa	1	1	orgán	1	4	tušenie	1	1
čušať	1	1	Orión	1	1	túženie	4	1
ďaleký	1	2	orlica	1	1	túžievať	1	1
dar	1	2	osada	1	4	túžiť	4	1
dať	11	10	osirelý	3	1	tvár	10	4
dať sa	1	5	oslava	1	2	tvárnosť	1	3
dávať	2	12	osláviť	3	2	tvoj	5	5
dávať sa	1	3	oslepiť	1	3	tvoriť	5	4
dávny	1	3	osnova	1	3	ty	3	4
dcéra	2	1	osoba	1	4	tyran	1	1
dejiny	1	1	osoh	1	1	ubiehať	1	1
deliť sa	2	4	ostať	2	5	ubiť	1	3
deň	2	3	ostýchavý	1	1	úbohý	1	2
deva	11	1	osud	3	2	ucho	2	3
devica	1	1	Otava	1	1	uchytiť	2	1
devin	1	1	otázka	2	2	učiť	3	5
Devín	3	1	otčina	5	1	udierať	1	1
diabelský	1	2	otcov	1	1	udusiť	1	5
dieťa	1	4	otočiť	3	3	uhádnuť	1	4
dieťatko	1	4	otrava	1	3	úkaz	1	1
dievčina	7	1	otráviť	1	4	ukázať	1	5
dievčinka	1	1	otrok	1	2	ukázať sa	1	3
dievka	1	3	otvárať	1	4	ukradnúť	1	1
div	1	2	otvoriť	4	4	ukrutný	1	2
dívať sa	14	3	oviať	2	1	ukryť	4	1
diviť sa	1	1	ozbrojiť	1	1	ukrývať sa	1	1
divný	1	1	ozdobiť	1	1	úloha	4	5
divodivý	1	1	ožiariť	1	3	umierať	8	1
divý	5	5	ožiť	3	3	úmor	1	2
doba	3	2	ozvať sa	1	4	umrieť	1	1
Dobroslava	2	1	ozvena	1	2	umučenie	2	1
dobrota	3	3	ozývať sa	8	4	umučiť	1	2
dobývať	1	2	padať	6	12	unášať	2	5
dohola	1	1	padnúť	4	12	uprosiť	1	1
dokázať	1	2	páliť	2	8	Ural	1	1
dokola	2	1	pamäť	1	6	určenie	1	2
dole	2	2	pamätať	1	3	určiť	2	3
dolina	7	1	pán	1	11	Urpín	1	1
dostať	5	9	panenský	1	2	usilovať sa	1	2

dostávať	1	8	pár	1	2	usmievať sa	2	1
dosti	1	1	Parom	2	1	ústa	1	2
dovoliť	1	1	pás	1	5	ustať	2	1
drahý (AJ)	5	3	pásť sa	1	2	ustískať	1	1
drahý (N)	4	1	pastier	3	1	ustlať	1	1
driemať	1	3	peklo	2	3	úsvit	2	1
drobný	1	2	peknosť	1	1	utekať	1	5
družica	2	2	pekný	1	3	utieknuť	1	1
družiť sa	1	3	pelešiť	1	1	utierať	1	2
dub	1	3	pena	1	2	útlosť	1	1
duch	4	5	perina	1	1	útly	2	2
dúha	3	2	piaty	1	1	uvädnúť	2	1
duma	1	1	piesenka	1	1	úžas	1	1
duša	15	5	pieť	1	1	uzavretý	1	4
duť	4	1	píjať	1	4	uzdravovať	1	1
dvere	1	1	píjavať	1	4	uzerať sa	1	1
dvoje	3	2	piť	1	4	úžiť sa	1	1
dvojiť	2	1	plachý	1	3	uznať	2	5
dvojiť sa	1	1	plakať	1	1	uzrieť	1	1
dvojne	1	1	plameň	4	1	vábiť	1	1
dvoriť si	3	1	planéta	1	2	Váh	2	1
dýchať	4	5	planý	1	4	val	1	1
Eol	1	1	plátno	1	3	Valhal	1	1
éter	1	2	plávať	5	5	valiť	1	2
fŕkať	1	4	plemä	1	1	valiť sa	2	3
Ganges	1	1	pleť	1	1	valný	1	2
hadov	1	1	pleva	3	1	vanúť	1	1
háj	4	1	plot	1	1	vanúť	1	1
haniť	1	1	pobúriť	1	2	variť	1	3
harmónia	1	1	pochvala	1	1	vaša	1	6
Himalája	1	1	pochybnosť	1	1	väz	1	2
hlad	1	3	pocítiť	3	3	večerný	1	1
hľadať	4	3	poctiť	1	1	večnosť	3	3
hľadieť	4	5	podať	3	5	večný	4	4
hladina	2	2	podávať	5	5	vedieť	1	6
hladiť	1	3	poddaný	1	1	vek	2	5
hlas	10	5	podieť sa	1	2	veleba	2	2
hlava	2	7	podlaha	1	1	velebný	3	2
hlboký	1	4	podlý	1	1	Velehrad	2	1

hložie	1	1	podnet	1	1	veličava	1	1
hnev	3	2	podoba	4	5	veľkocitý	2	1
hnusota	1	2	podopierať sa	1	1	veľkosť	4	5
hnutie	2	3	podpora	1	3	veľký	1	11
hodina	12	5	podstata	1	3	veniec	1	2
hodiť	3	2	pohár	2	4	verenie	1	1
hodiť si	1	1	pohľad	1	4	vernosť	6	1
hody	1	2	pohladiť	1	1	verný	5	3
hoja	1	1	pohnutie	1	1	veselý	5	2
hojiť	1	1	pohnutosť	1	1	viať	3	3
Hoľa	4	1	pohoda	1	2	viazať	2	8
holubinka	1	1	pohrávať	1	1	víchrica	2	1
holúbok	1	2	pohroma	1	1	vídavať	1	10
hora	14	2	Pohronie	1	1	vidieť	8	10
horieť	6	5	pohybovať	1	3	vidina	2	1
horievať	1	5	pohýnať	1	5	vienok	1	1
horliť	1	2	pohýnať sa	4	4	viera	4	3
horovať sa	1	1	poklad	3	2	Víla	9	1
horúci	2	2	pokoj	18	6	Víla-Marína	2	1
hospoda	1	3	pokojne	1	1	víno	1	1
Hospodin	2	1	pokonný	1	1	vinúť sa	2	3
hotový	1	5	pokora	2	1	viť	1	1
hoviadko	1	1	pokoriť sa	1	1	vítať	3	3
hovoriť	8	6	pokradnúť	1	1	víťaziť	1	2
hrad	2	1	pokrývať	1	3	vláda	5	4
hrať	4	6	pól	3	3	vládnuť	2	5
hrať sa	1	4	pole	4	7	vlaha	6	1
hrdosť	1	1	poletovať	1	2	vlákno	1	2
hrešiť	1	4	polkolo	1	1	vlas	4	3
hriech	1	3	poľúbok	1	1	vlások	2	2
hrkútať	1	1	polžitie	1	1	vlasť	2	2
hrmievať	1	3	pominúť sa	1	2	vlna	3	5
hrob	6	1	ponížiť	1	2	vlnka	1	5
hrobový	1	1	poobjímať	1	1	vlnobitie	1	1
hrom	4	1	popremáhať	2	1	vlúdiť	1	2
hromada	1	2	poprosiť	1	1	vnadný	1	1
hromoplesk	1	1	poroba	3	1	voda	11	4
Hron	6	1	porozdvojiť	1	1	vodička	1	4
hrozba	1	2	posadať si	1	1	vodiť	2	3

hroziť	1	3	posielať	2	2	vojvodiť	1	1
hrsť	1	2	poslať	2	2	vôl	1	1
hučanie	1	1	posmech	1	1	volať	10	7
hučať	2	2	posmievať sa	2	1	voňať	1	2
húsenka	1	1	postava	4	3	voňavý	1	1
hustý	1	2	posteľ	1	1	voziť sa	1	3
hviezda	1	3	posvätiť	1	2	vrah	3	1
hýbať sa	2	3	posvätný	2	2	vráta	1	1
hynúť	1	2	potkať	1	1	vrátiť sa	2	3
ideál	2	3	potok	2	1	vravieť	6	1
ihrať sa	1	1	potreba	5	4	vrelosť	1	1
iný	3	3	potvora	1	1	vrelý	2	2
istiť	1	1	povinný	1	2	vrenie	1	1
Itala	1	1	povstávať	2	1	vrkoč	1	1
ja	7	2	pozdraviť	3	3	všeobecnosť	1	1
jahoda	1	2	pozemský	1	2	všesvet	1	1
jaro	1	1	pozerať	1	3	vstať	1	2
jasne	1	1	požiar	3	1	vstávať	3	2
jasnosť	1	2	požívať	2	3	vtedy	1	2
jasný	1	8	pozlátiť	1	1	vtočiť sa	1	1
jastriť	1	1	poznať	1	7	vy	1	4
javiť sa	5	2	pozor	1	2	vycediť	1	1
Javorina	1	1	pozostať	1	1	vyčerpať	1	4
jed	3	2	požrať	1	2	východ	1	6
jediný	2	2	pozrieť	3	10	vychodiť	3	14
jedlina	1	1	praded	1	2	vychovať	1	3
jedno	1	11	prah	1	3	vyháňať	1	5
jednota	4	3	prameň	1	4	vyhnať	2	5
jedon	1	1	prapraprapraotec	1	1	vyhodiť sa	1	2
jeho	4	5	prask	1	2	vyhynúť	1	2
jelšina	1	1	pravdivý	1	2	vykradnúť sa	1	1
juh	1	2	práve	1	1	vykročiť	1	2
Júlia	1	1	pravica	1	3	vylepiť	1	4
kalich	2	3	pravý	1	6	výlev	3	1
kameň	1	3	prebiehať	1	6	vylúdiť	1	1
kamienok	1	3	prebiť sa	2	2	vymodliť	1	1
kar	1	1	prebrodiť	1	1	vymoriť	1	1
karmín	1	1	prebrodiť sa	1	1	vymrieť	1	1
kat	2	1	prebudiť	1	3	vynášať	1	6

kázať	2	3	prebývať	1	1	vypínať sa	1	5
kľačať	1	1	preč	1	4	vyplakať	1	2
klam	2	2	predmet	6	3	vyplniť	1	3
klebeta	6	1	predrahý	3	1	vypriasť	1	1
kliať	1	1	prehovoriť	2	5	výraz	1	3
kliatba	1	3	prehrešiť sa	1	1	vyrútiť sa	1	3
kloniť	1	1	prejatý	1	1	vyryť	1	3
kloniť sa	1	2	prejímať	1	3	výšava	2	1
kojiť	1	1	preletieť	1	6	výšina	2	2
koľaj	3	2	prelietať	1	6	vyšinúť sa	1	1
kolembať	1	1	prelievať	1	5	vysmiať	1	1
kolísať	1	3	preložiť	1	7	vysmiať sa	1	1
kolíska	4	2	premáhať	1	5	vysoko	1	6
kolovať	1	1	premena	1	1	vysoký	2	7
komár	1	1	premeniť	1	3	výsosť	2	1
Komárno	1	1	premeniť sa	1	1	vystaviť (to exhibit)	2	3
konať	1	2	premilý	2	1			
konečnosť	1	2	preplaviť sa	1	1	vystaviť (to build up)	1	1
koreň	2	3	prepletať sa	1	2			
koriť sa	1	1	prerieknuť	1	1	vystrieť	1	4
kosiť	1	2	prerodiť sa	1	1	výstrojiť	2	1
kosť	2	1	prerývať	1	2	vystúpiť	1	8
kraj	6	4	prestať	5	2	vysušiť	1	3
krajina	6	3	prestierať	1	2	vysvetliť	1	2
kráľ	1	4	prestol	1	1	vyvaliť sa	1	4
králica	2	1	prestrieť	1	2	vyvodiť	1	1
kráľovná	1	4	pretvoriť sa	1	1	vyžobrať	1	1
krása	12	2	prevážať	1	2	vyzrieť	1	2
krásne	1	2	previevať	5	2	vyzvať	1	1
kráž	2	1	prezrieť	1	3	vyzývať	1	1
krídlatý	1	1	prezývať	1	1	vzbúriť	1	3
krídlo	2	5	priateľ	3	3	vzdychať	2	3
krištáľ	1	2	pribyť	1	1	vziať	6	11
krivooký	2	1	príčina	1	2	vziať si	3	1
kríž	1	3	pridružiť	1	1	vzkriesenie	1	2
krížiť sa	1	3	priepasť	1	2	vzletieť	1	1
kŕmievať	1	1	priestor	2	2	vzlietnuť	1	2
kŕmiť	1	1	priestora	2	1	vzpínať sa	1	1
kročiť	2	1	prijať	2	10	vzývať	8	2

krok	1	4	prijímať	4	10	zábava	1	2
krútiť sa	3	4	prikovať	1	1	zábavný	1	2
kryštál	2	3	prikročiť	2	2	zabiť	1	5
kryť	2	3	primrieť	4	1	zablúdiť	3	2
kryť sa	3	1	pripínať	2	1	zabudnúť	1	6
kúpavať sa	1	2	pripojiť	1	2	zabudnutie	2	2
kúr	1	1	prirásť	1	1	začať	1	3
kus	1	5	príroda	9	3	zachodiť	1	6
kvákať	1	1	prirovnať	1	1	zachovať	1	3
kvet	7	4	prísaha	1	1	začínať	1	3
kvitnúť	1	3	prísnosť	1	1	záclona	1	1
kypieť	1	3	prísny	1	3	zacloniť	1	2
kývať	2	3	prísť	1	20	zadláviť	1	1
kývať sa	1	1	pristrojiť	1	2	zahaliť sa	1	1
labuť	1	1	pritisnúť sa	1	1	zahnať	1	4
ľad	2	2	privodiť	2	1	zahorieť	3	4
lahoda	1	1	prosba	1	1	záhrada	1	1
ľalia	2	1	prosiť	2	2	zahryznúť	1	2
lampada	1	1	prsia	1	4	záhuba	5	1
lapiť	1	1	prstenka	1	1	zahynúť	1	2
láska	6	3	psohlavý	1	1	zajať	1	3
láskavý	2	1	psota	3	1	zajatý	1	1
let	4	3	púčkový	1	1	zájsť	1	6
letieť	14	9	purpura	1	1	zakázať	1	1
leto	1	1	pustatina	1	1	zakliaty	2	1
ležať	2	7	pustiť sa	1	5	zákon	5	5
liať	2	4	puto	2	2	zakrútiť	1	4
liať sa	1	2	rad	2	4	zakrývať	1	3
líce	6	2	rád	3	5	zakvetlý	1	1
lícomilý	1	1	rada	2	3	zakvitnutý	1	1
lietať	7	5	radosť	3	2	zaletieť	1	3
ligotať sa	1	1	raj	1	3	zálety	1	2
lipa	1	5	rameno	1	3	zaliať	1	3
lipový	1	1	raniť	2	1	zalkať	1	1
lkať	1	1	ranný	1	1	žaloba	2	2
ľúbiť	4	3	ráno	1	1	založiť	1	9
ľúbiť sa	1	2	ráz	2	1	zaľúbiť sa	1	3
ľúbosť	12	2	reč	1	7	zamieriť	1	2
Lučatín	1	1	reťaz	1	2	zamĺknuť	1	1

lúčiť sa	1	1	rieka	4	1	zamračila sa	1	2
ľúto	1	4	Rím	1	1	zanoriť sa	2	1
ľútosť	1	2	rod	1	6	zapadnúť	1	8
ľutovať	1	3	rodina	9	4	zápal	3	3
machnatý	1	2	rodinný	2	2	zaplakať	1	1
magnetik	1	1	rodiť sa	5	2	zápona	1	2
mak	1	2	rojiť sa	1	2	zaprisahať	1	2
malebný	1	1	rok	1	3	zardieť sa	1	1
málo	1	1	roniť	1	1	zarechtať sa	1	1
malý	5	8	rosa	2	2	zárod	1	2
mámenie	1	2	rosiť	2	1	žartovať	2	2
mámivý	1	3	rovnosť	1	1	zašatiť sa	1	2
mamľas	1	1	rovný	1	6	zasľúbiť	1	1
Marína	21	1	rozbiť	1	2	zasmiať sa	1	1
mariť	2	2	rozbroj	3	1	zasmútiť	2	1
márny	1	2	rozchod	2	4	zasnívať sa	1	1
mať (N)	1	2	rozdrapiť	1	3	zaspať	1	2
mať (V)	9	11	rozdvojiť	1	2	zaspievať	1	2
mater	1	2	rozhnať	1	2	zastať	9	7
matka	2	5	rozhnať sa	1	2	zastať	2	7
mdloba	1	1	rozjariť sa	1	1	zastaviť	2	3
medzera	1	2	rozkladať	1	7	zastaviť sa	1	3
meniť	2	3	rozkladať sa	1	4	zastrieť	4	4
meniť sa	1	3	rozklásť	1	1	zastrieť sa	1	2
meno	1	3	rozkvitnúť	1	3	zašumieť	1	1
menovať	2	3	rozliať	1	3	zašusťať	1	1
mesiac	1	3	rozliať sa	1	2	zasvätený	1	2
Mesiáda	1	1	rozlietnuť sa	1	1	zasvätiť	1	4
milá	6	2	rozlievať	2	3	zasvietiť	1	2
milenec	1	2	rozložiť	1	7	zato	1	1
milenka	5	2	rozložiť sa	1	4	zatriasť	1	1
milený	1	1	rozlúčenie	1	1	zatúžiť	1	1
milión	1	2	rozlúčiť	1	1	zaviať	5	4
milosť	2	5	rozlúčiť sa	1	2	závidieť	1	1
milostivý	1	2	rozmarín	1	1	zavierať	3	1
milovať	2	4	rozmetať	2	1	závistlivý	1	1
milý (AJ)	15	4	rozopäť	1	3	zaviť	4	2
milý (N)	3	2	rozosiať	1	2	závoj	1	1
mizerný	3	3	rozpomienka	1	1	závora	1	2

miznúť	1	2	rozpomínať sa	1	1	zavrieť	3	7
mladosť	2	2	rozrušiť	1	2	zavrieť sa	2	3
mladucha	3	1	rozryť	1	1	zažať sa	1	1
mladý	10	3	rozrývať	1	1	zazdať sa	1	1
mlčať	2	3	rozsiať	1	2	zážívný	1	2
mnohotvárny	1	1	rozsievať	1	2	zaznieť	4	1
moc	5	6	rozsúdiť	1	2	zaznievať	2	1
mocnár	1	2	roztúžený	1	2	zázrak	2	2
mohyla	2	2	roztvoriť	1	3	zazrieť	1	2
môj	20	5	rozum	1	2	zazvoniť	2	2
Mojmírove	1	1	rozvíjať sa	1	3	zbadať	1	1
more	6	2	rozvinúť	1	4	zbaviť	1	3
moriť	1	1	rozviť sa	3	2	zbiť	1	3
možnosť	1	3	rozvliecť	1	2	zbĺknuť	1	1
mrak	4	3	rúbať	1	4	zboriť	2	1
mramor	1	1	rubín	2	1	zbroj	1	2
mráz	2	3	rúhať sa	1	1	zdanie	1	2
mreža	1	1	ruka	2	4	zdesiť sa	1	1
muka	1	1	rušiť	5	3	zdravý	1	3
mumlavý	1	2	rútiť sa	1	3	zduriť	1	2
musieť	2	7	ruža	1	3	zdvojiť	1	1
mútiť sa	1	2	ružička	1	3	zdýmať	1	3
my	3	4	ružový	2	3	zelený	1	4
nabrať	2	4	ryť	2	3	zem	8	5
nachýliť sa	1	1	sa	3	5	zemeplaz	2	1
nadávať	1	2	sad	1	2	zemský	1	1
nádej	2	2	sadnúť	1	8	zemšťan	1	1
nádeja	4	1	sadnúť si	1	8	žena	2	2
nadšenie	4	1	sádok	1	1	zhniť	1	1
nadšený	2	1	Sahara	2	1	zhojiť	1	1
nadýmať	1	1	sám	1	5	zhora	1	2
nadýmať sa	1	1	samota	2	2	zhorieť	2	2
náhoda	1	1	satan	1	1	zhubiť	2	1
náhrada	1	2	satanský	1	1	zhýbať sa	1	1
nahý	1	2	satira	1	2	zhynúť	1	1
naliať	1	3	šatiť	1	1	žiadať	3	3
námesačník	1	1	šaty	3	2	žiadny	1	1
napadať	1	1	Sáva	1	1	žiadúci	1	1
napínať	1	3	scendžať	1	1	žiaľ	2	1

napiť sa	1	3	schladnúť	2	3	žialiť	2	1
naplašiť	1	1	schovať	1	1	žiara	1	1
nárek	2	1	schytiť	4	2	žiariť	2	3
nariekať	5	2	schytiť sa	1	3	získať	1	5
národ	3	4	sčiastky	1	1	ziskriť sa	2	2
náručie	4	1	sedieť	1	8	žiť	11	11
náš	1	7	šedivý	1	3	žitie	5	4
nasýtiť	1	3	sen	3	4	život	7	12
naveky	2	2	sever	2	2	živý	3	9
nazmar	1	1	siať	1	1	zjasniť sa	1	1
nazývať	1	1	sila	3	5	zjavenie	1	2
nebáť sa	2	3	Sion	1	1	zjaviť sa	10	2
nebeský	1	2	široký	1	4	zjavovať sa	1	2
nebešťan	1	1	širošíry	1	1	zjednotiť	1	1
neblažiť	1	1	šírosť	2	2	zlato	2	4
nebo	10	2	šíry	1	1	zlatý	10	7
nebodajný	1	1	Sitno	2	1	zletieť	3	3
nebyť	8	7	sivý	1	1	zlietnuť	1	1
nechať	1	10	skala	11	2	zlomiť	1	4
nechcieť	2	4	skalina	1	1	zlosť	1	1
nečistý	1	3	skalný	1	2	zlosyn	1	1
necítiť	3	4	skamenieť	1	2	zložiť	2	10
nedať	1	10	skaza	2	1	zmámený	1	1
nedávať	1	11	sklátiť	1	1	zmeniť sa	2	1
nedbať	1	3	sklepenie	1	1	zmierenie	1	1
nedláviť	1	3	skloniť	1	1	zmizeť	1	3
nedočkavý	1	1	skloniť sa	2	1	zmiznúť	1	3
nedostať	1	9	skočiť	1	5	zmožený	1	1
nekaliť	1	3	škoda	1	2	zmútiť	1	4
neľúbiť	2	3	škola	1	6	zmyť	1	2
nelúčiť sa	1	1	škrekľavý	1	2	znamenať	3	2
nemať	1	11	skryť sa	6	3	známy	1	3
nemeniť sa	1	3	skrývať	1	3	znemieť	1	2
nemota	1	2	škúliť	1	4	zničiť	2	1
nemý	2	2	skúpy	1	2	zobudiť	1	2
nenie	1	1	skúsiť	1	3	zočiť	2	1
neoberať	1	3	skvitnúť	3	2	zodierať	1	4
neobjímať	1	1	slabosť	1	2	zohnúť sa	1	1
neočariť	1	1	sladiť sa	1	1	zora	3	1

neodbiť	1	4	sladkosť	1	2	zore	1	1
neodkliať	1	1	sladký	2	4	zoriť	1	1
neohlušiť	1	2	slasť	5	1	zornica	2	2
neotvoriť	1	4	sláva	17	4	zorný	1	1
neozývať sa	1	4	Sláva	3	1	zostarať sa	2	2
nepohnúť	1	5	slávievať	1	3	zostať	1	1
nepokoj	1	3	sláviť	1	3	zostávať	1	1
nepostaviť	1	10	slávny	1	4	zožierať	1	5
nepotrebný	1	1	slepý	1	5	zrada	1	1
nepovädnúť	1	2	sloboda	9	5	zradne	1	1
nepovedať	1	8	slobodne	1	1	zrak	7	2
nepoznať	1	7	slovo	3	3	zrastený	1	1
nepravý	1	1	sľub	1	1	zrážať	1	7
neprebiť	1	3	sľúbiť	2	2	zrkadlo	2	3
neprechodiť sa	1	1	slúchať	1	2	zrodiť	3	2
neprestrašiť	1	1	sluha	1	2	zrušiť	1	2
nepreväžiť	1	4	slušať	1	1	žubrienka	1	1
neprosiť	1	1	slúžiť	5	8	zutekať	1	1
neraniť	1	1	smútiť	2	1	zvábiť	1	1
nerásť	1	6	smútkový	1	1	zvädnúť	2	2
nerozboriť	1	1	smutný	2	2	zváľať	1	2
nerozdrážiť	1	1	sňať	2	1	zväzok	1	4
nerozdvojiť	1	2	sneh	1	2	zveličenie	1	1
nerozprávať	1	3	snem	4	1	zveriť	2	3
neškodiť	1	1	sniť	2	1	zviazať	1	6
neskúsiť	2	3	snívať	8	3	zvlnený	1	3
nesláviť	1	3	snívať sa	1	1	zvon	4	2
nespievať	5	6	snuť sa	1	1	zvoniť	3	6
nespravodlivý	1	1	Sodoma	2	1	zvrieť	1	1
nešťastie	2	2	sokol	1	1	zvuk	1	2
nešťastný	1	4	spanilý	4	1	zvýšiť	1	4

Appendix 2

Frequencies and number of meanings of rhyme words in *Morava*

Word	Freq.	No. of Meanings
blýskať	1	2
čas	1	9
človek	1	4
deň	1	3
Kykymora	1	1
Morava	1	1
nádeja	1	1
nemota	1	2
oko	1	2
opona	1	1
pazúr	1	2
prestať	1	2
purpura	1	1
rana	1	4
rieka	1	1
rozbroj	1	1
rozprávať	1	3
skala	1	2
skálie	1	1
skloniť sa	1	1
sláva	1	4
slepcov	1	1
slota	1	2
strana	1	9
svoj	1	6
tieň	1	4
tôňa	1	1
trojhlavý	1	1
utískať	1	1
valiť sa	1	3
viať	1	3
vlaha	1	1
výstraha	1	1
zápas	1	1
zatriasať	1	1
zazvoniť	1	2
zbojcov	1	1
zdať sa	1	3
zdýmať sa	1	1
zhora	1	2

Glottometrics 49, 2020, 32-51

Quantitative Analysis of Academic Writing as to Informality and Vocabulary Features

Ziqi Liu[1], *Haitao Liu*[2]

Abstract. What matters for a learner of English for academic purposes is to possess the ability to present results and achievements in international top journals. The ability is related to degrees of informality, vocabulary richness, and lexical complexity in academic writing. This study takes Chinese master degree candidates and advanced writers as research objects and concentrates on two research questions: (1) To what extent do Chinese master degree candidates and advanced writers differ in the use of informal features in their writings? (2) In what ways do Chinese master degree candidates and advanced writers differ in vocabulary choices? The results are based on studying two datasets of the research objects. However, our results show that there is a complex picture for each informality indicator. Finally yet importantly, advanced writers show a higher level of vocabulary richness and complexity.

Keywords: informality features, vocabulary richness and complexity, Chinese master degree candidates, advanced writers, academic writing, EFL.

1. Introduction

Avoiding informality is necessary and essential for learners of English for academic purposes. Academic writing is characterized as an impersonal and objective reporting on independent and external reality (Lee, Bychkovska, & Maxwell, 2019; Hyland, 2001a). Thus, to avoid informality in academic writing is a key factor. Furthermore, academic writing is not just about the results, it is also relevant to the representation of writers (Hyland, 2002). Then, an important factor is how to employ vocabulary, and what words should be selected to convey the study content. Thus, exploring the gap in informality features, lexical richness, and complexity between "novices" and "experts" is indispensable.

As for the method, a quantitative approach should have a firm place and wide application in the study of academic writing. It can be employed to process a great amount of material not only with a lot of diverse features, but also in a short time. Furthermore, more details of datasets are acquired by exploiting the quantitative method. What is more, it is feasible to state the frequency of each informality feature and the figures of vocabulary richness and complexity indexes. Based on those data, further and more accurate analysis about the comparison is conducted.

[1] Department of Linguistics, Zhejiang University, Hangzhou, China.

[2] Institute of Quantitative Linguistics, Beijing Language and Culture University, Beijing, China; Department of Linguistics, Zhejiang University, Hangzhou, China. Correspondence to: Haitao Liu. Email address: htliu@163.com, ORCID-No.: https://orcid.org/0000-0003-1724-4418.

In recent decades, some studies have employed different indexes to compare informality features among diverse texts. The study by Petch-Tyson (1998) is carried out on writings of EFL students from different backgrounds, either language or cultural ones, including French, Dutch, Swedish, and Finnish. Aijmer (2002) concentrates on the situation of modality in Swedish learners' written interlanguage. It is difficult for second language students to express a suitable degree of doubt and certainty. Thus, Hyland and Milton (1997) study qualification and certainty in the writings of L1 and L2 students. Cobb (2003) explores the Québec learner corpus. As for sentence-initial *and* and sentence-initial *but*, the study by Bell (2007) is based on selections from 11 academic journals containing the domains of science, the humanities, and social science. Wang (2016) studies grammatical colloquial features through theses of EFL learners. In order to investigate the trend of informality, Hyland and Jiang (2017) examine 10 informal features (first-person pronouns, unattended anaphoric pronouns, split infinitives, sentence-initial conjunctions or conjunctive adverbs, sentence-final preposition, listing expressions, second-person pronouns/determiners, contractions, direct questions, and exclamations) across four disciplines, which are applied linguistics, sociology, electrical engineering, and biology. In the discipline of applied linguistics, Alipour and Nooreddinmoosa (2018) also investigate informality features. Besides, Lee et al.'s contribution (2019) is to compare informality features in the writing of L1 and L2 undergraduate students.

According to Fang and Liu (2015), the study of lexical richness was founded by Chotlos and Yule (Chotlos, 1944; Yule, 1944). It is complex either in linguistic or in mathematical aspects (Wimmer & Altmann, 1999). Lexical richness measurement belongs to one of the most traditional domains in quantitative linguistics (Kubát & Milička, 2013). The reason for employing it lies in the fact different groups of people use vocabulary with specific features. As for lexical diversity, Wen's research (2006) represents that the mean value of vocabulary richness of written English is higher than that of spoken English. In written language, writers have more time to avoid repeating some words. Thus, higher repeat rate and lower lexical richness are more likely to occur in colloquial English. In the study of Li and Liu (2019), they propose that written English and spoken English are two main types of style, and compared with written English, there exist informality features in spoken English according to Hickey (2014). Thus, vocabulary diversity is associated with informality features.

Vocabulary richness can be employed to investigate stylistic features (Smith & Kelly, 2002), to analyze different translation works (Fang & Liu, 2015), to explore genre analysis (Kubát & Milička, 2013) and authorship attribution (Jamak, Savatić, & Can, 2012; Hoover, 2003). As for lexical complexity – another indicator of writing style –, if there are more complex words, it means that the text is more sophisticated (Dai & Liu, 2019).

However, few studies compare the gap between Chinese master degree candidates and advanced writers based on both informality features (first/second-person pronouns, sentence-initial conjunctions / conjunctive adverbs, listing expressions, and modal verbs), and on vocabulary richness / complexity. In order to fill in the gap and to help Chinese English learners publish research articles in international top journals, this paper will employ the combination of the two aspects to study the following research questions:

(1) To what extent do Chinese master degree candidates and advanced writers differ in the use of informal features in their writings?

(2) In what ways do Chinese master degree candidates and advanced writers differ in vocabulary choices?

The arrangement of this paper goes as follows. In the second section, the information about two self-built datasets and methodology is introduced. The third section is the results and discussion of the study, which relates to presentation and analysis of informality features and vocabulary richness / complexity of the datasets. In the final section, a conclusion is presented.

2. Methodology

2.1 Description of Material

In order to explore the gap in informality and vocabulary richness / complexity, two datasets, Chinese Master Thesis (CMT) and International Research Article (IRA), are established. They contain abstracts of Chinese master theses and of international research articles. An abstract is essential for academic writing. It summarizes the major aspects of the paper, which are introduction, methodology, results, and discussion. Besides, the abstract includes the research background and research questions, contains experimental design and methods used, and includes key results and their interpretations. The research of abstracts is also divergent. Abstracts are important materials in many studies concerning, for instance, publication (De Bruin, Treccani, & Sala, 2014; Scherer, Dickersin, & Langenberg, 1994; Snedeker, Totton, & Sargeant, 2010) or academic literacy practices (Starke and Bailer, 2019). Given the wide application of abstracts of research articles, they can also be employed to study the degree of informality and vocabulary features.

IRA contains abstracts of articles from 2014 to 2018 of three top journals in the domain of linguistics, which are *Journal of Memory and Language* (5-Year Impact Factor = 5.763), *Applied Linguistics* (5-Year Impact Factor = 4.516), and *Journal of Second Language Writing* (5-Year Impact Factor = 4.177). There are 75 abstracts in total (5 per journal each year). Besides, the amount of tokens for IRA is 13,555 and that of types is 2,570. To balance with IRA, CMT comprises 45 abstracts of Chinese master theses from the domain of Foreign Linguistics and Applied Linguistics in the same 5 years from three universities, which are ZJ (Zhejiang University), DL (Dalian Maritime University), and HN (Henan Normal University). ZJ belonged to Project 985[3] and Project 211[4]. DL was one member of Project 211. HN did not belong to either projects.

There are 15 abstracts for ZJ, 20 for DL, and 10 for HN. The tokens of CMT are 17,666 and the types are 2,507. Both the number of tokens and types have been acquired by software, QUITA (Kubát, Matlach, & Čech, 2014). The total of tokens of the two datasets is 31,221. In the study of Kalantari and Gholami (2017), 18,751 running words in the corpus are employed to investigate the lexical complexity development. Thus,

[3] Project 985 in China aims to construct world-class universities.

[4] Project 211 in China aims to strengthen about 100 institutions of higher education and key disciplines.

the amount of tokens of datasets in this study seems appropriate. Besides, the details of texts in CMT and IRA are listed in the appendix.

Table 1

Descriptions of CMT and IRA

Text	Types	Tokens
CMT	2,507	17,666
IRA	2,570	13,555

2.2 Data Analysis

2.2.1 Informality Features

In order to explore informality features of the two datasets, this research adopts an approach which is based on the revised version of other studies, which include Hyland and Jiang (2017), Lee et al. (2019), Aijmer (2002), and Petch-Tyson (1998). In the study of Hyland and Jiang (2017), first-person pronouns, second-person pronouns, unattended reference, and sentence-initial conjunctions / conjunctive adverbs are important indexes to indicate informality.

Next, academic writing should be semantically clear. If needless words are omitted, it will benefit achieving that goal. Employing unattended reference appropriately will make the expressions more concentrated, economical, and concise. Thus, unattended reference is not adopted as a feature of informality in this paper.

Last but not the least, academic writings prefer concrete and specific expressions. However, listing items are usual in the process of writing with vague and abstract meanings. Furthermore, it is also easily neglected. Thus, listing expression is taken into account as a feature of informality in the study.

All informality indexes employed in this paper to evaluate different degrees of informality in CMT and IRA are shown in Table 2.

Table 2

Description of informality features indexes

Category	Details
First-person pronouns	I, me, my, mine, we, us, our, ours
Second-personal pronouns	you, your
Sentence-initial conjunctions / Conjunctive adverbs	and, but, or, so, yet, again, also, besides, however, indeed, still, thus
Listing expressions	and so forth, and so on, etc.
Modal verbs	can, may, might, will, must, would, could, shall, should, ought to, have (got) to

To investigate whether the difference in each informality feature is significant or not, log-likelihood (*LL*) value is counted by the calculator (http://ucrel.lancs.ac.uk/llwizard.html) (Lee et al., 2019). At 5% level, $LL \geq 3.84$ means $p < 0.05$; at 1% level, $LL \geq 6.63$ is significant for $p < 0.01$; at 0.1% level, $LL \geq 10.83$ is equal to $p < 0.001$; at 0.01% level, $LL \geq 15.13$ represents $p < 0.0001$. As stated in Lee et al. (2019), *ELL* measure (Johnston, Berry, & Mielke, 2006), which represents effect size of log-likelihood measure, is also contained in the calculator. Besides, through the software AntConc (Anthony, 2011), the frequency of each informality feature in Table 2 is obtained.

2.2.2 Vocabulary Features

To study the gap of vocabulary features, which are richness and complexity, in CMT and IRA, different indexes (TTR, h-point, R_1, Repeat Rate, Entropy, and Average Tokens Length) are employed in this research.

TTR (V/N; the ratio of different words to all words) is an indicator of testing vocabulary richness (Yoon, 2017). Next, repeat rate (RR) and entropy are also indicators of vocabulary diversity, which are both based on the probability of occurrences of words in the text. In detail, the smaller the repeat rate, the greater the vocabulary richness; on the contrary, the greater the entropy, the greater the richness (Popescu, Čech, & Altmann, 2011).

The h-point is also calculated on the basis of word frequency (Dai & Liu, 2019). It is the point where the rank is equal to its frequency. Then,

$$r = f(r)$$

is applied to this situation. If there is no exact place like this, two neighbouring points will be adopted, which have $f(i)$ and $f(j)$. Under these circumstances,

$$f(i) > r_i \text{ and } f(j) < r_j,$$

and generally $r_i + 1 = r_j$ (Popescu et al., 2011). Here comes the formula of h-point:

(1) $$h = \frac{f(i) \times r_j - f(j) \times r_i}{r_j - r_i + f(i) - f(j)}.$$

The h-point is a critical point for the rank-frequency distribution of words in a text. Autosemantic words tend to appear after the h-point. In contrast, synsemantic words appear before the h-point. Hence, the h-point is an indicator for vocabulary richness.

$F(h)$ is the cumulative probability of words with the order from 1 to the h-point. With h and $F(h)$, another vocabulary richness index, R_1, has been proposed. R_1 is defined as follows:

(2) $$R_1 = 1 - \left(F(h) - \frac{h^2}{2N}\right).$$

TTR, h-point, R_1, repeat rate, and entropy are all the indicators to explore vocabulary richness. As for lexical complexity, word length is a common index (Dai & Liu, 2019). The larger the word length is, the more complex the text is. Thus, average tokens length is employed to investigate lexical complexity in this research. It is the mean of all the tokens lengths in the whole text. Those indicators of vocabulary richness and lexical complexity are measured by QUITA (Kubát et al., 2014).

3. Results and Discussion

3.1 Gap in Informality Features between CMT and IRA

Table 3 lists the overall frequencies and descriptions of five informality features. As shown, it is rather complicated to interpret these results. First-person pronouns are more likely used in IRA. However, sentence-initial conjunctions / conjunctive adverbs, listing expressions, and modal verbs occur more frequently in CMT.

For CMT and IRA, there is a significant difference in four indicators, except the index of second-person pronouns. Second-person pronouns (*you*, *your*) represent an obvious way to refer to readers as general or individual referents (Hyland, 2005). They are also the most visible acknowledgements of the reader's presence (Hyland, 2001b). Besides, there is a high percentage of occurrences of second-person pronouns in the texts of L2 learners (Petch-Tyson, 1998). Furthermore, in the study of Lee et al. (2019), second-person pronouns are also numerous in COLTE, the corpus of L2 learners. However, Hyland's research (2005) proposes that these reader pronouns (*you* and *your*) occur rarely in the student corpus. As for Table 3, the frequency of second-person pronouns in both databases is 0. It indicates that there may be a low frequency of *you* and *your* in abstracts, and even in the whole academic writing. For CMT, the writers are perhaps willing to present themselves in a relative junior status compared with the teachers, supervisors, and readers (Hyland, 2005). Thus, they try to avoid using second-person pronouns. As for the advanced writers or experts, informal features are also not suitable in their studies. Thus, the writers in CMT and IRA are likely not to use the second-person pronouns.

Table 3

Overall frequency and description of informality features

	CMT	IRA	*LL*	*ELL*	Significant
First-person pronouns	9	83	89.82	0.00078	√
Second-person pronouns	0	0	0	0	×
Sentence-initial conjunctions / conjunctive adverbs	70	26	10.96	0.00009	√
Listing expressions	4	0	4.56	-5.79	√
Modal verbs	117	55	9.45	0.00007	√

Note: *LL* = log-likelihood value; *ELL* = effect size for log likelihood; √ = $p < 0.05$; × = $p > 0.05$.

3.1.1 First-person Pronouns

As to Table 3, the frequency of first-person pronouns in CMT is 9 and that in IRA is 83. Besides, the log-likelihood value yields 89.82 (p < 0.0001), which means there is a significant difference; the effect size for log-likelihood is 0.00078.

What is more, the first-person pronoun is the only indicator that occurs more frequently in IRA than in CMT. In details, there is no occurrence of four pronouns (*me*, *my*, *mine*, *ours*) in both datasets. As a consequence, the difference between them is not significant. Compared to zero frequency of *us* in CMT, it occurs only once in IRA. As for *I*, *we*, and *our*, there is an obviously significant difference. Those data are presented in Table 4.

Table 4

Overall frequency and description of first-person pronouns

	CMT	IRA	*LL*	*ELL*	Significant
I	0	5	8.34	0.00034	√
me	0	0	0	0	×
my	0	0	0	0	×
mine	0	0	0	0	×
we	6	62	69.7	0.00066	√
us	0	1	1.67	-0.00006	×
our	3	15	12.23	0.00019	√
ours	0	0	0	0	×
total	9	83	89.82	0.00078	√

Note: *LL* = log-likelihood value; *ELL* = effect size for log likelihood; √ = p < 0.05; × = p > 0.05.

The results of previous studies for distribution of first-person pronouns are not unequivocal. Petch-Tyson's study (1998) reveals that non-native speakers of English adopt first-person pronouns more frequently than native speakers. Nevertheless, some studies show the opposite situation. According to Hyland (2002), experts or professional writers use more first-person pronouns than students. Besides, L1 writers are more likely to intervene with self-mentions (Lee & Deakin, 2016). Lee et al. (2019) also support the claim that L1 writers employ first-person pronouns/determiners more than ESL students. The study in this paper is in line with the second point – that advanced writers use them more in IRA. In the study by Leedham and Fernandez-Parra (2017), they find out that there is more occurrence of *we* for L1 Chinese and L1 Greek students than for L1 English students, and less frequency of *I* for L1 Chinese and Greek students than for L1 English students. However, in Table 4, *I*, *we*, and even *our* are used more frequently in IRA than in CMT.

First-person pronouns are considered to be a typical informality marker (Hyland & Jiang, 2017). Arguments in academic writings should be proposed in the most convincing way. Besides, acceptability, certainty, and plausibility of research require different and complex features, which include strong evidence, originality, and innovation of study, and an authoritative professional personality (Hyland & Jiang, 2017). Thus, an independent identity and the writer's voice need to be established. In the study of Hyland (2001a), employing first-person pronouns is a way to build and project a personal standing and authority. In addition, it is also a function to distinguish the writers from others. Thus, intervening with first-person pronouns appropriately benefits Chinese students in the constantly changing and competitive circumstances. Word choice also reveals the writers' social and psychological factors (Hyland, 2002); because of cultural background, some writers are more likely to avoid using first-person pronouns or to show modesty. Given the results of first-person pronouns, especially for *we*, it is essential to study further whether to make them the indicators of informality, or not.

3.1.2 Modal Verbs

As shown in Table 5, there is a significant difference in the modal verbs as a whole. The log-likelihood value is 9.45 ($p < 0.01$), with the effect size of 0.00007. Among them, significant difference only exists in four verbs, which are *may*, *will*, *could*, and *should*. Unlike *will*, *could*, and *should*, *may* is used more frequently in IRA rather than in CMT.

Table 5

Overall frequency and description of modal verbs

	CMT	IRA	*LL*	*ELL*	Significant
can	48	23	3.61	0.00003	×
may	14	22	4.54	0.00005	√
might	4	1	1.22	0.00005	×
will	21	1	17.45	0.00025	√
must	5	1	1.96	0.00007	×
would	3	4	0.53	0.00002	×
could	9	1	5.42	0.00012	√
shall	0	0	0	0	×
should	13	0	14.81	4.46	√
have (got) to	0	2	3.34	-0.00076	×
ought to	0	0	0	0	×
total	117	55	9.45	0.00007	√

Note: *LL* = log-likelihood value; *ELL* = effect size for log likelihood; √ = $p < 0.05$; × = $p > 0.05$.

Will is a predictive modal and is employed to predict future events with some certainty (Grant & Ginther, 2000; Aijmer, 2002). Besides, *may* is in the category of possibility modals with a lower certainty (Grant & Ginther, 2000; Aijmer, 2002; Hinkel, 2009). As for *could*, it is also a possibility modal with the meaning of probability (Grant & Ginther, 2000; Aijmer, 2002). According to Kennedy (1998), *may* occurs less in spoken corpus than in written corpus (879 versus 1,323); however, *could* (2,000 versus 1,744) and *will* (4,286 versus 2,804) are more highly employed in the spoken corpus. Hence, the lower occurrence of *may* in CMT represents higher informality. Besides, more uses of *will* and *could* also illustrate the lower formality of CMT. What is more, *should*, belonging to the obligation and necessity group, represents some actions with the meaning of desire and suggestion (Grant & Ginther, 2000; Aijmer, 2002). Biber et al. (2002) suggest that although obligation modals are usually suppressed, they are also exploited to express the meaning of personal obligation. Besides, some writings of non-native speakers seem brusque, dogmatic, too direct, and too tentative (Hyland & Milton, 1997). Thus, *should* is used less by advanced writers in IRA.

3.1.3 Sentence-initial Conjunctions / Conjunctive Adverbs

Alipour and Nooreddinmoosa (2018) illustrate that sentence-initial conjunctions are the most frequently used among those informality feature indexes in both native and non-native articles. As of Table 6, the significant difference lies in the total of sentence-initial conjunctions and conjunctive adverbs (*LL* = 10.96, $p < 0.001$, *ELL* = 0.00009) in CMT and IRA. More connectors in CMT may be a result of instructions. The writers are encouraged to use them to convey logic of academic writing and to show the connection between the preceding content and the following one. From this perspective, it means that the degree of informality for CMT is higher than for IRA. Besides, there exists significant difference in three main indicators, which are sentence-initial *and*, *so* and *besides*.

Table 6

Overall frequency and description of sentence-initial conjunctions / conjunctive adverbs

	CMT	IRA	*LL*	*ELL*	Significant
And	22	1	18.50	0.00026	√
But	1	0	1.14	-9.21	×
Or	0	0	0	0	×
So	5	0	5.69	-4.65	√
Yet	0	1	1.67	-0.00006	×
Again	0	0	0	0	×
Also	1	0	1.14	-9.21	×
Besides	7	0	7.97	-2.38	√
However	29	20	0.14	0	×
Indeed	0	0	0	0	×
Still	0	0	0	0	×
Thus	5	4	0	0	×
total	70	26	10.96	0.00009	√

Note: *LL* = log-likelihood value; *ELL* = effect size for log likelihood; √ = $p < 0.05$; × = $p > 0.05$.

As shown in Table 6, sentence-initial *and*, which is second to *however*, is still highly employed in CMT. Bell (2007) also proposes three main roles of sentence-initial *and*, which are (i) to indicate the last item within the whole list; (ii) to develop arguments further; (iii) to represent shifts in authorial perspectives. In the study of Bell

(2007), sentence-initial *and* and sentence-initial *but* are most frequently used with additive and contrastive meanings respectively. In Table 6, sentence-initial *and* is still ranking first in its semantic group. However, sentence-initial *however* becomes the first rather than *but* for writers in both CMT and IRA. Furthermore, sentence-initial *however* is the most frequently adopted in both datasets, CMT and IRA. According to Hyland and Jiang (2017), the increases of sentence-initial *however* as well as declines of sentence-initial *but* and sentence-initial *and* also occur in the domain of applied linguistics and sociology. Some studies also prove that sentence-initial *however* is the most frequent used item of sentence-initial conjunctions and conjunctive adverbs (Lee et al., 2019; Alipour & Nooreddinmoosa, 2018).

3.1.4. Listing Expressions

Listing expression, a common index, is another type of informality features. As shown in Table 7, there is only one significant difference in the group. The log-likelihood value for the whole is 4.56 ($p < 0.05$), and the effect size is -5.79. However, no significant difference exists for individual listing expressions.

Table 7
Overall frequency and description of listing expressions

	CMT	IRA	*LL*	*ELL*	Significant
and so on	1	0	1.14	-9.21	×
and so forth	0	0	0	0	×
etc	3	0	3.42	-6.93	×
total	4	0	4.56	-5.79	√

Note: *LL* = log-likelihood value; *ELL* = effect size for log likelihood; √ = $p < 0.05$; × = $p > 0.05$.

Listing expressions in both datasets occur at a much lower frequency than the other four informality features. Compared with four occurrences of listing expressions in CMT, there is no hit in IRA. It may be due to the fact that advanced writers are aware of vagueness of listing expressions (Lee et al., 2019).

3.2 Gap in Vocabulary Richness and Complexity between CMT and IRA

As shown in Table 8, except the h-point and *RR*, values of vocabulary richness indicators – which are TTR, entropy, and R_1 – are higher in IRA than in CMT. The values of h-point and *RR* are higher in CMT than in IRA. All the data represent that there is more diversified and colourful vocabulary in advanced writers' material (IRA).

TTR is the type-token ratio. Compared with more tokens and fewer types in CMT, there are fewer tokens and more types in IRA. As to entropy and R_1, direct indicators of lexical richness, advanced writers have more word choices. The higher value of *RR* in CMT represents its lower lexical diversity. Besides, IRA (lower h-point) are more likely to possess more autosemantic words, which tend to come after h-point, and higher vocabulary richness.

Table 8

Description of vocabulary richness indexes

	CMT	IRA
TTR	0.141911	0.189598
h-Point	46	40
Entropy	8.902671	9.257164
R_1	0.635515	0.683807
RR	0.013138	0.008741

Note: CMT – types are 2,507; tokens are 17,666. IRA – types are 2,570; tokens are 13,555.

Average tokens length is an approach to test lexical sophistication approximately. It is seen in Table 9 that advanced writers in IRA are more likely to employ words with more complexity. Thus, a gap exists in lexical richness and complexity between master degree candidates and advanced writers.

Table 9

Description of vocabulary complexity indexes

	CMT	IRA
Average Tokens Length	5.455734	5.710144

Note: CMT: Types are 2,507; tokens are 17,666. IRA: Types are 2,570; tokens are 13,555.

4. Conclusions

This study employs two self-built datasets (CMT and IRA) to explore two research questions.

(1) To what extent do Chinese master degree candidates and advanced writers differ in the use of informal features in their writings?

(2) In what ways do Chinese master degree candidates and advanced writers differ in vocabulary choices?

In order to respond to the first research question, five informality features indexes (first-person pronouns, second-person pronouns, sentence-initial conjunctions / conjunctive adverbs, listing expressions, and modal verbs) – the choice based on the previous studies (Hyland & Jiang, 2017; Lee et al., 2019; Aijmer, 2002; Petch-Tyson, 1998) – are employed. The research is carried out by AntConc (Anthony, 2011), log-likelihood value and effect size calculator, and QUITA (Kubát et al., 2014). Details for each informal indicator provide a complex picture. On the one hand, advanced writers overuse first-person pronouns to express their identities and stances. On the other hand, Chinese master degree candidates frequently employ more modal verbs, sentence-initial conjunctions / conjunctive adverbs, and listing expressions. However, there is no occurrence of second-person pronouns in either group.

To answer the second research question, six indexes (type-token ratio, h-point, R_1, repeat rate, entropy, average tokens length) are selected to capture lexical diversity and vocabulary sophistication; the values are counted by the QUITA (Kubát et al., 2014) software. As for lexical richness, advanced writers possess higher type-token ratio, R_1, and entropy as well as lower h-point and repeat rate. It means that Chinese master degree candidates show a lower vocabulary diversity. Besides, experts in IRA also have a higher average tokens length. According to this result, Chinese master degree students are more likely to employ shorter words with less lexical sophistication than advanced writers.

Acknowledgements

This work is partly supported by the National Social Science Foundation of China (Grant No.17AYY021) and the Special Project for Humanities and Social Science, Zhejiang University (*Quantitative Analysis of Sentence Hierarchical Structure in English*).

References

Anthony, L. (2011). AntConc (Version 3.2. 4w). Tokyo: Waseda University. Available from http://www.laurenceanthony.net/software.

Aijmer, K. (2002). Modality in advanced Swedish learners' written inter-language. In: S. Granger, J. Hung & S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins Publishing Company, 57–76.

Alipour, M., & Nooreddinmoosa, M. (2018). Informality in Applied Linguistics Research Articles: Comparing Native and Non-Native Writings. *Eurasian Journal of Applied Linguistics* 4(2), 349–373.

Bell, D. (2007). Sentence-Initial *And* and *But* in Academic Writing. *Pragmatics* 17(2), 183–201.

Biber, D., Conrad, S., & Leech, G. (2002). *Longman Student Grammar of Spoken and Written English*. Harlow: Longman.

Chotlos, J. W. (1944). IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs* 56(2), 75–111.

Cobb, T. (2003). Analyzing Late Interlanguage with Learner Corpora: Québec Replications of three European Studies. *The Canadian Modern Language Review* 59(3), 393–424.

Dai, Z., & Liu, H. (2019). Quantitative Analysis of Queen Elizabeth II's and American Presidents' Christmas Messages over 50 Years (1967–2018). *Glottometrics* 45, 63–88.

De Bruin, A., Treccani, B., & Della Sala, S. (2015). Cognitive Advantage in Bilingualism: An Example of Publication Bias? *Psychological science* 26(1), 99–107.

Fang, Y., & Liu, H. (2015). Comparison of vocabulary richness in two translated Hongloumeng. *Glottometrics* 31, 54–75.

Grant, L., & Ginther, A. (2000). Using Computer-Tagged Linguistic Features to Describe L2 Writing Differences. *Journal of Second Language Writing* 9(2), 123–145.

Hickey, R. (2014). *A Dictionary of Varieties of English*. Hoboken: Wiley-Blackwell.

Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics* 41(4), 667–683.

Hoover, D. L. (2003). Another Perspective on Vocabulary Richness. *Computers and the Humanities* 37, 151–178.

Hyland, K. (2001a). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes* 20(3), 207–226.

Hyland, K. (2001b). Bringing in the Reader Addressee Features in Academic Articles. *Written Communication* 18(4), 549–574.

Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics* 34(8), 1091–1112.

Hyland, K. (2005). Representing readers in writing: Student and expert practices. *Linguistics and Education* 16(4), 363–377.

Hyland, K., & Jiang, F. (2017). Is academic writing becoming more informal? *English for Specific Purposes* 45, 40–51.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing* 6(2), 183–205.

Jamak, A., Savatić, A., & Can, M. (2012). Principal Component Analysis for Authorship Attribution. *Business Systems Research* 3(2), 49–56.

Johnston, J. E., Berry, K. J., & Mielke Jr, P. W. (2006). Measures of Effect Size for Chi-Squared and Likelihood-Ratio Goodness-of-Fit Tests. *Perceptual and Motor Skills* 103(2), 412–414.

Kalantari, R., & Gholami, J. (2017). Lexical Complexity Development from Dynamic Systems Theory Perspective: Lexical Density, Diversity, and Sophistication. *International Journal of Instruction* 10(4), 1–18.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Longman.

Kubát, M., Matlach, V., & Čech, R. (2014). *QUITA. Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag.

Kubát, M., & Milička, J. (2013). Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics* 20(4), 339–349.

Lee, J. J., Bychkovska, T., Maxwell, J. D. (2019). Breaking the rules? A corpus-based comparison of informal features in L1 and L2 undergraduate student writing. *System* 80, 143–153.

Lee, J. J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing* 33, 21–34.

Leedham, M., & Fernández-Parra, M. (2017). Recounting and reflecting: The use of first person pronouns in Chinese, Greek and British students' assignments in engineering. *Journal of English for Academic Purposes* 26, 66–77.

Li, T., & Liu, X. (2019). Ji yu yu liao ku gao zhong sheng ying yu shu mian yu kou yu hua te zheng yan jiu [A corpus-based study on the colloquial features in English writing produced by senior high students]. *Basic Foreign Language Education* 21(1), 3–11.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In: S. Granger (ed.), *Learner English on Computer*. London: Longman, 107–118.

Popescu, I.-I., Čech, R., & Altmann, G. (2011). *The Lambda-structure of Texts*. Lüdenscheid: RAM-Verlag.

Scherer, R. W., Dickersin, K., & Langenberg, P. (1994). Full publication of results initially presented in abstracts: A meta-analysis. *Jama* 272(2), 158–162.

Smith, J. A., & Kelly, C. (2002). Stylistic Constancy and Change across Literary Corpora: Using Measures of Lexical Richness to Date Works. *Computers and the Humanities* 36, 411–430.

Snedeker, K. G., Totton, S. C., & Sargeant, J. M. (2010). Analysis of trends in the full publication of papers from conference abstracts involving pre-harvest or abattoir-level interventions against foodborne pathogens. *Preventive Veterinary Medicine* 95(1–2), 1–9.

Starke, M. D. D. J., & Bailer, C. (2019). Práticas de letramentos acadêmicos de alunos do Pibid interdisciplinar linguagens-Furb. *Revista EntreLínguas* 5(1), 195–209.

Wang, N. (2016). Investigating Grammatical Colloquial Features in EFL Learners' Theses by Chinese English Learners. *International Journal of English Linguistics* 6(6), 138–146.

Wen, Q. (2006). Ying yu zhuan ye xue sheng shi yong kou yu – bi yu ci hui de cha yi [Vocabulary variation across speech and writing produced by English majors]. *Foreign Languages and Their Teaching* 7, 9–13.

Wimmer, G., & Altmann, G. (1999). Review Article: On Vocabulary Richness. *Journal of Quantitative Linguistics* 6(1), 1–9.

Yoon, H. J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System* 66, 130–141.

Yule, G. U. (1944). *A Statistical Study of Literary Vocabulary*. Cambridge: University Press.

Appendix: Texts Information

Texts in IRA

Text	Title
1	Unconventional Word Segmentation in Emerging Bilingual Students' Writing: A Longitudinal Analysis
2	Critical Analysis of CLIL: Taking Stock and Looking Forward
3	Discipline and Level Specificity in University Students' Written Vocabulary
4	An Investigation into Metaphor Use at Different Levels of Second Language Writing
5	Dynamics of Complexity and Accuracy: A Longitudinal Case Study of Advanced Untutored Development
6	Assessing Lexical Proficiency Using Analytic Ratings: A Case for Collocation Accuracy
7	Involvement in University Classroom Discourse: Register Variation and Interactivity
8	The Theoretical Research Article as a Reflection of Disciplinary Practices: The Case of Pure Mathematics
9	Towards a Theory of Diagnosis in Second and Foreign Language Assessment: Insights from Professional Practice Across Diverse Fields
10	Marking Importance in Lectures: Interactive and Textual Orientation
11	Teacher Trainers' Beliefs About Feedback on Teaching Practice: Negotiating the Tensions Between Authoritativeness and Dialogic Space
12	An Activity-Theoretic Study of Agency and Identity in the Study Abroad Experiences of a Lesbian Nontraditional Learner of Korean
13	The Effects of Complexity, Accuracy, and Fluency on Communicative Adequacy in Oral Task Performance
14	Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels
15	Implicit and Explicit Cognitive Processes in Incidental Vocabulary Acquisition
16	Individual Differences in Early Language Learning: A Study of English Learners of French
17	Exploring the Role of Phraseological Knowledge in Foreign Language Reading
18	Comprehension and Knowledge Components That Predict L2 Reading: A Latent-Trait Approach
19	A Longitudinal Study on the Impact of CLIL on Affective Factors
20	The Impact of Out-of-School Factors on Motivation to Learn English: Self-discrepancies, Beliefs, and Experiences of Self-authenticity
21	Fitting in or Standing out? A Conflict of Belonging and Identity in Intercultural Polite Talk at Work

22 Multilingual Children between Real and Imaginary Worlds: Language Play as Resignifying Practice

23 Evidence for the Adoption of the Flemma as an Appropriate Word Counting Unit

24 Building a Trustworthy Corporate Identity: A Corpus-Based Analysis of Stance in Annual and Corporate Social Responsibility Reports

25 Towards Ecological Validity in Research into Input-based Practice: Form Spotting Can Be as Beneficial as Form-meaning Practice

26 Effects of acoustic-phonetic detail on cross-language speech production

27 What can we learn from learning models about sensitivity to letter-order in visual word recognition?

28 *Nomina sunt consequentia rerum* – Sound–shape correspondences with every-day objects figures

29 Variables and similarity in linguistic generalization: Evidence from inflectional classes in Portuguese

30 The “sense boost” to dative priming: Evidence for sense-specific verb-structure links

31 Early semantic activation in a semantic categorization task with masked primes: Cascaded or not?

32 Examining the effects of active versus inactive bilingualism on executive control in a carefully matched non-immigrant sample

33 Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing

34 The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language

35 The effect of plausibility on eye movements in reading: Testing E-Z Reader’s null predictions

36 The persistence of syntactic priming revisited

37 Structural priming in artificial languages and the regularisation of unpredictable variation

38 Primed codeswitching in spontaneous bilingual dialogue

39 Unifying structural priming effects on syntactic choices and timing of sentence generation

40 The independence of syntactic processing in Mandarin: Evidence from structural priming

41 Phrase frequency effects in free recall: Evidence for redintegration

42 Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers

43 Distributional learning of subcategories in an artificial grammar: Category generalization and subcategory restrictions

44 Serial position, output order, and list length effects for words presented on smartphones over very long intervals

45 The first- and second-language age of acquisition effect in first- and second-language book reading

46 Testing enhances memory for context

47 Voluntary language switching: When and why do bilinguals switch between their languages?

48 Listener sensitivity to probabilistic conditioning of sociolinguistic variables: The case of (ING)

49 How does foveal processing difficulty affect parafoveal processing during reading?

50 Semantic diversity, frequency and the development of lexical quality in children’s word reading

51 Quantifying the development of phraseological competence in L2 English writing: An automated approach

52 Conceptualizing and measuring short-term changes in L2 writing complexity

53 Exploring multiple profiles of L2 writing using multi-dimensional analysis

54 Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners

55 L2 student–U.S. professor interactions through disciplinary writing assignments: An activity theory perspective

56 The effects of cognitive task complexity on writing complexity

57 What our students tell us: Perceptions of three multilingual students on their academic writing in first year

58 Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables

59 “We’re drifting into strange territory here”: What think-aloud protocols reveal about convenience editing

60 Exploring changes in FL writers’ meaning-making choices in summary writing: A systemic functional approach

61 The relationship between lexical sophistication and independent and source-based writing

62 Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings

63 Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays

64 The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality

65 Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings

66 Motivation and feedback: How implicit theories of intelligence predict L2 writers’ motivation and feedback orientation

67 Source text use by undergraduate post-novice L2 writers in disciplinary assignments: Progress and ongoing challenges

68 Using mind maps to reveal and develop genre knowledge in a graduate writing course

69	Emergent arguments: A functional approach to analyzing student challenges with the argument genre
70	Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis
71	Conceptualizations of language errors, standards, norms and nativeness in English for research publication purposes: An analysis of journal submission guidelines
72	An analysis of grammatical patterns in generation 1.5, L1 and L2 students' writings: A replication study
73	Balancing stability and flexibility in genre-based writing instruction: A case study of a novice L2 writing teacher
74	Articulating struggle: ESL students' perceived obstacles to success in a community college writing class
75	Synchronous and asynchronous teacher electronic feedback and learner uptake in ESL composition

Texts in CMT

Text	Title
1	A Corpus-based Study of English Verb Patterns in Marine Engineering English
2	A Corpus-based Study on Adjective Complementation
3	A Study of the Effect of Instruction under SCT on the Reading Achievement
4	Stylistic Analysis on Language Characteristics of Maritime Oral English
5	A Corpus-based Study on OVER in Maritime News English from the Cognitive Perspective
6	A Corpus-based Panchronic Study on Semi-auxiliaries
7	A Corpus-based Lexical Study in *the International Aeronautical and Maritime Search and Rescue Manual*
8	A Corpus-based Analysis of Stylistics of Headlines of Maritime News
9	The Study on Polysemous Word PUT from Cognitive Perspective
10	A Study on Three Metafunctions in *Marine News* Texts
11	A Corpus-based Study on Collocations of Key Words in Nautical English
12	Case Studies on the Effects of Text Summarization on Argumentation Writing Qualities of EFL Learners at Different Proficiency Levels
13	An Investigation into the Relationship between Junior High School Students' Foreign Language Anxiety, Emotional Intelligence and English Achievement
14	Semantic Features of Evaluative *It*-Clauses in the Research Articles by Chinese Writers
15	A Corpus-based Study on the Use of Shell Nouns in Marine Accident Investigation Report
16	A Study on Chinese College Students' Use of *Of*-Clusters and *Of*-Errors

17	A Corpus-based Comparative Study on *Turn* Verbs in MEC and BNC
18	A Case Study on the Impacts of Comprehensive Corrective Feedback on EFL Learners’ Written Accuracy
19	The Effect of Tasks on Collaborative Dialogue – An Empirical Study of English Majors at Dalian Maritime University
20	A Corpus-based Study on Nautical Lexis in Nautical Fiction
21	A New Analysis of *THERE* Sentences – from the Perspective of Copy Movement
22	A Contrastive Analysis of Hedges in English News from Chinese and American Newspapers
23	A Study of English Middle Construction Acquisition by Chinese EFL Learners
24	A Study on Characteristics of Contrastive Discourse Markers by Chinese English Majors
25	An Empirical Study on the Acquisition on English Verb Raising by Chinese College Students
26	An Empirical Study on the Acquisition of Wh-questions by Chinese EFL Learners
27	A Study on Chinese EFL Learners’ Article Acquisition from the Perspective of Syntax-pragmatics Interface
28	A Contrastive Study of Thematic Progression Patterns in English and Chinese Fairy Tales
29	A Corpus-based Study on the Use of Stance Adverbs in Chinese Learners’ Academic Writing
30	A Study on the Developmental Features of Lexical Bundles in Chinese English Learners' Academic Writing
31	Vocabulary Input in EFL Middle School Textbooks in China – A Corpus-based Study of Frequency, Complexity and Distribution
32	A Corpus-based Approach to the Interaction of English Verb Patterns with *it* and Registers
33	A Corpus-based Analysis of Recent Changes in American English Perfect Construction
34	Chinese Scholars' Perceptions of Writing for International Publication: An Applied Linguistics Perspective
35	Contextual Effects on the Processing Mechanism of Chinese 3VO Metaphor: An ERP Study
36	An Empirical Study on the Developmental Differences of Chinese EFL Learners' Implicit and Explicit Grammatical Knowledge
37	A Dependency Treebank-based Research on the Syntactic Complexity in Chinese EFL Learners’ Writings: A Developmental Perspective
38	Promotional Functions of Modal Verbs in the Introduction Sections of International Journal Articles: A Corpus-Based Analysis
39	Semantic and Affective Processing of Emotional Words in L1 and L2 in Unbalanced Bilinguals

40	Dependency Distance Differences across Interpreting Types
41	The Validity and Utility of the Component Unit in Chinese Sentence Chunking
42	A Study of Two Types of Writing Tasks in EFL Testing: Writing-only Tasks and Continuation Tasks
43	On Passive Voice Attrition among Non-English Majors from the Perspective of Dynamic Systems Theory
44	A Multidimensional Approach to Temporal Distance in Language
45	A Study on Demotivation of High School EFL Learners – A Case Study in a High School in Zhejiang Province

Glottometrics 49, 2020, 52–65

School and Gender in Numbers: A Stylometric Insight into the Lexis of Teenagers' Description Essays

Michal Místecký[1], *Lucie Radková*[2]

Abstract. The goal of the paper is to make use of four quantitative indicators (MATTR, ATL, Q, and VD) to study vocabulary richness, lexis complexity, text activity, and syntactic complexity of Czech schoolchildren's writing tasks. The corpus comprises 60 texts written by elementary-school and secondary-school pupils, distributed equally according to the gender and the education level (30 each); the genre of the task was description ("my bedroom" for the elementary schoolers, and "class/school of the future" for the secondary ones). The research is carried out in three distinct comparisons (schools, genders, and the mixture of both), and the results are interpreted with the assistance of a pedagogical professional. At the end of the study, a detailed summary of the outcomes is provided.

Keywords: MATTR, ATL, Q, VD, stylometry, quantitative linguistics, pedagogy, school, gender, Czech.

1. Introduction

Recently, a lot of studies have appeared focusing on quantitative investigation of various linguistic discourses. With the literary and politics-oriented studies leading the way (cf. Andreev et al. 2018; Místecký 2018; Dai, Liu 2019), there are other fields in which a stylometric analysis can bear needed fruit (David et al. 2014; Čech 2016). In the present paper, the sphere of Czech schoolchildren's writing tasks will be researched, with the goal to enrich both didactic scholarship, and to prove the usability of the quantitative methods in the domain.

The school writing has been given much attention in the last years (cf. Čechová at al. 2008, Holubová 2014, Štěpáník, Holanová 2017, Rysová 2017), which culminated in devoting an entire section in the new Czech handbook of stylistics (Hoffmannová et al. 2016) to the subject. However, to our knowledge, no paper has been published yet to study the matter from the viewpoint of quantitative measurement. This is why this article will try to fill the gap in the research, and may become a pioneering piece for other analyses to come.

[1] University of Ostrava, Ostrava, the Czech Republic; e-mail: mmistecky@seznam.cz.

[2] University of Ostrava, Ostrava, the Czech Republic; e-mail: lucie.radkova@osu.cz.

2. Methods

Out of the numberless methods used in stylometry, four have been selected to start the investigation off (MATTR, ATL, Q, and VD). The set is a combination of the pragmatic approach – all the indexes can be computed automatically, with no manual work needed –, and the endeavour to take various stylistic factors into account. Moreover, all the calculations have been proved to be independent of text length, and also of each other (cf. Zörnig, Místecký 2018).

First, MATTR (Moving-Average Type-Token Ratio) has been used, as it seems to be the most effective and widespread tool of assessing vocabulary richness (cf. Covington, McFall 2010). Its formula, based on TTR (Type-Token Ratio), but taking into consideration the division of a text into moveable sections (windows), is as follows –

$$MATTR(L) = \frac{\sum_{i=1}^{N-L} V_i}{L(N - L + 1)}.$$

In the formula, L stands for the number of tokens in one window, V_i for the number of types in one window, and N for the total of the tokens in the text. Given the length of the studied tasks, the window size in the present research was set at 30. The basic unit of the research is, due to the workings of the software, the word-form.

The count will be exemplified upon the following micro-text:

Tato škola je velmi moderní. V budově školy se nachází více než sto učeben.[3]

The excerpt comprises 14 tokens; arbitrarily, the size of the window will be set at 10 tokens. It means that there are, in total, five windows (including words 1–10, words 2–11, words 3–12, words 4–13, and words 5–14). The count reads –

$$MATTR(10) = \frac{9 + 9 + 10 + 10 + 10}{10 * (14 - 10 + 1)} = 0.96.$$

The overall MATTR-calculated vocabulary richness of the excerpt is 0.96.

Second, the texts were investigated on the basis of the average tokens length (ATL). This indicator is supposed to provide some information on the complexity of the vocabulary the producer of a text tends to use. Its formula is –

$$ATL = \frac{1}{N} \sum_{i=1}^{N} p_i ;$$

p_i stands for the number of graphemes in a word i, and N for the total of the words in the text. It is to be noted that such a count may be problematic in languages with no script, or in those where the written form does not correspond much with the spoken one (e.g., English and French).

For the sake of an example, let us have a Czech sentence

Moje škola budoucnosti by měla být veliká.[4]

[3] "This school is very modern. In the school building, there are more than one hundred classrooms."
[4] "My school of the future should be big."

The ATL value of its words is calculated as –

$$ATL = \frac{4+5+11+2+4+3+6}{7} = 5\,.$$

This means that the average length of the tokens in the example is 5 graphemes.

Next, we will focus on measuring to what extent a text is story-oriented, or, inversely, description-based. To this end, we will make use of an index originally devised by Busemann (1925), and later on employed in stylometry (cf. Andreev, Místecký, Altmann 2018). Activity (Q), as it is called, is a ratio of the number of the verbs in the text (V) and the total of the adjectives (A) and the verbs in it; formally –

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

The count will be exemplified upon a sample task (p_M_1). In it, there are 13 verbs and 8 adjectives; the calculation thus yields –

$$Q = \frac{13}{13+8} = 0.6190\,.$$

If $Q > 0.5$, the text may be considered active; if $Q < 0.5$, it is taken as descriptive. In our case, the text is active.

The last index to be counted is verb distances (VD). It is a simple indicator of syntactic complexity of a text, which takes into account the number of words to be found in between two verbs. Mathematically –

$$VD = \frac{1}{D}\sum_{i=1}^{D} d_i\,,$$

D signifying the number of the distances between the verbs, and d_i the number of the words between the verbs.

The count will be presented upon the sample sentence

Vchod vás nejdříve skenuje a poté se rozjede pás, který vás doveze do šatny.[5]

Here, there are three verbs ("skenuje", "rozjede", and "doveze"), which accounts for two distances; the formula thus calculates –

$$VD = \frac{3+3}{2} = 3\,.$$

The average verb distance in the given sentence is 3 words.

To conclude, in order to be able to compare the results of text groups, we will employ the statistical u-test, which is a traditional tool in quantitative linguistics (cf. Kubát 2016). Its formula reads –

[5] "First, the entrance will scan you, and then, a line will start moving, taking you to your locker room".

$$u = \frac{|\overline{X_1} - \overline{X_2}|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}};$$

X_1, X_2 stand for the average values of the two sets of values, s_1, s_2 for the standard deviations, and n_1, n_2 for the numbers of texts in the two groups. The standard deviations are counted on the basis of the given values. If $u > 1.96$, the difference between the two sets is considered statistically significant.

3. Material

The study material comprises 60 texts in total; 30 texts were written by pupils at elementary school, another 30 by the grammar-school ones. The age of the participants was 12–13, which corresponds to the "sixth grade" of the elementary school (and the first grade of the Czech upper elementary education), and the first grade (called "prima") of the eighth-year grammar secondary school. The pupils at the latter have passed a written entrance exam. The gender was represented equally, each group consisting of 15 boys and 15 girls. The genre of the written task was description; the elementary-school pupils were supposed to describe their bedrooms, whereas the grammar-school ones focused on their class of the future, or on the school of the future.

Before the very task, both groups were given instructions on how to write the description; they were confronted with various examples, and were asked to produce sample writings on their own. The attention was paid to the choice of suitable parts-of-speech, with adjectives (including those used in recommended comparisons) prevailing over the number of verbs, and to keeping the unified structure of the description (the left-right, centre-periphery, or bottom-up directions, etc.); the pupils were told not to repeat words, especially verbs "be" and "have". As to the personal intake of the teacher, some elementary-school boys followed the option to describe their bedrooms from unusual viewpoints (e.g., the cell phone, the second person plural, the things in the bedroom, etc.).

Each of the 60 texts is tagged according to whether it was written by an elementary-school pupil ("e"), or a grammar-school one ("g"); next, the gender is indicated, "M" standing for the males, "F" for the females; and finally, it is allocated a number, the boys occupying the upper first half (1–15), and the girls the lower one (16–30). Examples will be presented in the forthcoming section.

As to the interpretations, we have made use, besides our own pedagogical experience, of the insights provided by PhDr. Věra Podhorná, psychologist and the head of the Advisory Centre of Pedagogy and Psychology at Karviná, Czechia.

4. Results

The results will be commented upon in various constellations. First, the exhaustive list of all the outcomes will be presented (see Tables 1a–b). Next, three sets of text groups will be contrasted to each other (schools, genders, and the combination of the two). To conclude, the indexes will be investigated as to their capacity to provide statistically significant outcomes in comparing the studied sets. All the figures are rounded to the nearest hundredth.

Tables 1a–b
The results of the counts for all the samples

	MATTR	**ATL**	**Q**	**VD**
e_M_1	0.90	4.18	0.62	4.75
e_M_2	0.82	4.06	0.50	5.72
e_M_3	0.85	4.12	0.47	4.96
e_M_4	0.87	4.22	0.69	4.93
e_M_5	0.90	4.18	0.85	4.31
e_M_6	0.81	4.30	0.72	4.97
e_M_7	0.89	4.14	0.52	4.39
e_M_8	0.85	4.31	0.59	4.83
e_M_9	0.93	4.10	0.64	3.89
e_M_10	0.91	4.31	0.76	5.00
e_M_11	0.88	4.34	0.79	5.86
e_M_12	0.91	4.70	0.52	4.50
e_M_13	0.90	4.42	0.56	5.56
e_M_14	0.80	4.00	0.67	4.06
e_M_15	0.85	4.43	0.63	6.80
e_F_16	0.87	4.25	0.22	5.55
e_F_17	0.85	4.23	0.45	5.41
e_F_18	0.88	4.10	0.53	5.84
e_F_19	0.87	4.34	0.51	5.27
e_F_20	0.88	3.94	0.64	5.24
e_F_21	0.88	4.34	0.42	4.82
e_F_22	0.87	3.87	0.67	5.41
e_F_23	0.89	4.20	0.61	4.59
e_F_24	0.86	3.98	0.53	5.14
e_F_25	0.86	3.97	0.72	5.20
e_F_26	0.87	4.06	0.74	5.00
e_F_27	0.87	4.00	0.44	5.18
e_F_28	0.86	4.00	0.71	4.17
e_F_29	0.90	4.48	0.34	8.56
e_F_30	0.87	4.32	0.44	5.33

	MATTR	**ATL**	**Q**	**VD**
g_M_1	0.86	4.70	0.39	5.72
g_M_2	0.87	4.95	0.41	6.25
g_M_3	0.93	5.11	0.38	6.43
g_M_4	0.85	4.62	0.50	3.89
g_M_5	0.94	5.03	0.54	5.37
g_M_6	0.90	4.75	0.51	5.64

g_M_7	0.94	4.76	0.39	6.51
g_M_8	0.88	5.01	0.45	6.20
g_M_9	0.92	4.79	0.44	5.70
g_M_10	0.90	4.89	0.38	6.73
g_M_11	0.91	4.62	0.38	6.95
g_M_12	0.86	4.49	0.29	6.17
g_M_13	0.94	5.10	0.58	4.77
g_M_14	0.87	4.67	0.32	3.70
g_M_15	0.93	4.86	0.50	4.40
g_F_16	0.95	4.92	0.45	5.82
g_F_17	0.94	4.78	0.27	8.67
g_F_18	0.91	4.62	0.46	6.21
g_F_19	0.93	5.55	0.35	6.25
g_F_20	0.87	4.78	0.32	8.89
g_F_21	0.91	4.76	0.40	5.41
g_F_22	0.88	4.98	0.30	7.19
g_F_23	0.93	5.08	0.53	7.61
g_F_24	0.94	4.93	0.59	5.42
g_F_25	0.90	4.72	0.41	6.60
g_F_26	0.92	4.68	0.50	5.95
g_F_27	0.93	4.94	0.42	6.97
g_F_28	0.91	5.05	0.46	7.25
g_F_29	0.91	4.74	0.43	5.38
g_F_30	0.90	4.35	0.53	3.94

4.1 Elementary School vs. Grammar School

In this part, we are going to confront the outcomes of the pupils on the basis of the type of school they attend. The results are presented in Table 2; here, the averages of the elementary-school and grammar-school values are listed, altogether with the u-test calculations. The statistically significant results are marked with asterisks.

Table 2
The results of the school confrontation

	Elementary School	**Grammar School**	**u-test**
MATTR	0.87	0.90	5.02*
ATL	4.20	4.84	12.20*
Q	0.58	0.43	5.24*
VD	5.17	6.07	3.28*

It is visible that the schools significantly differ in all the indicators. In activity, the grammar-school pupils manifest, on average, descriptiveness ($Q < 0.5$), which is in line with the requirements of the genre they have produced; on the other hand, the elementary-school participants tend towards activity, even though the average is not very high above 0.5. It is thus to be inferred that the grammar-school attendants respect the pre-defined norms of the genre, whilst elementary-school children may not go by them so strictly.

It is probable that the high level of descriptiveness is linked to the difference in verb distances, too. The multitude of adjectives prolongs the sentences, raising the complexity of their structures; this is also supported by the use of comparisons and examples. It is not be forgotten, though, that the grammar-school pupils wrote about the class/school of the future; this topic necessitates more explanations, which may have complicated the used phrasing even further.

To conclude, there is a significant difference in the average length of tokens, elementary-school children using, on average, shorter words than their grammar-school peers. This is also connected to the same situation in the sphere of the MATTR-measured vocabulary richness. Both indicate that grammar-school people may be more sophisticated in the use of lexis than the elementary schoolers. It is an open question whether the factor behind these results is the intellectual capacity of the children, or whether they are attributable to the motivated family environments. Another possible explanation stems from, once again, the difference in the topics – the sci-fi-like description of the school/classroom of the future may require the use of a lot of loanwords (names of appliances, specialized vocabulary, etc.), these being longer than the basic words of the Slavic origin.

4.2 Boys vs. Girls

Regardless of the schools, the gender will be investigated in this subchapter. The general results are listed in Table 3.

Table 3
The results of the gender confrontation

	Boys	**Girls**	**u-test**
MATTR	0.88	0.89	1.10
ATL	4.54	4.50	0.41
Q	0.53	0.48	1.53
VD	5.30	5.94	2.27*

The results confirm the idea that gender itself is not a sufficient discriminant of style; only one difference – the one in verb distances – is statistically significant. We may presuppose variation in the stylistic manners of boys and girls at different schools; the average values are thus not very indicative of general trends. For instance, if the standard deviations of the activity values are counted, we arrive at the interval of 0.39–0.67 for the boys, and 0.35–0.61 for the girls; it means that the two genders manifest, on average, the same figures. As to verb distances, however, girls tend to be more syntactically complex, which may hint at the fact that they may respect the standards of the genre more than the boys, since the principles of the description writing (use of adjectives, exemplifications, comparisons, a certain amount of precision, etc.) favour employment of complicated phrase structures. Nonetheless, a detailed viewpoint may shed more light on this interpretation.

4.3 Combining Factors

In the present research, the corpus will be divided into four groups, with respect to both gender and school. The elementary-school boys ("e_M"), the elementary-school girls ("e_F"), the grammar-school boys ("g_M"), and the grammar-school girls ("g_F") will be treated separately. For each index, the counted values of the u-test will be summed up, presented, and visualised in the forms of scatter plots. This is a well-established procedure in statistics-driven research (cf. Kubát, 2016).

First, the vocabulary richness results will be commented upon.

Table 4

MATTR: the values of u-test of the studied text groups

MATTR	e_M	e_F	g_M
g_F	5.74*	10.00*	2.40*
g_M	3.24*	4.44*	
e_F	0.31		

Table 5

The average MATTRs and the sums of the u-test values for the studied text groups

	MATTR	**u-test Sum**
e_M	0.87	9.30
e_F	0.87	14.75
g_M	0.90	10.09
g_F	0.92	18.15

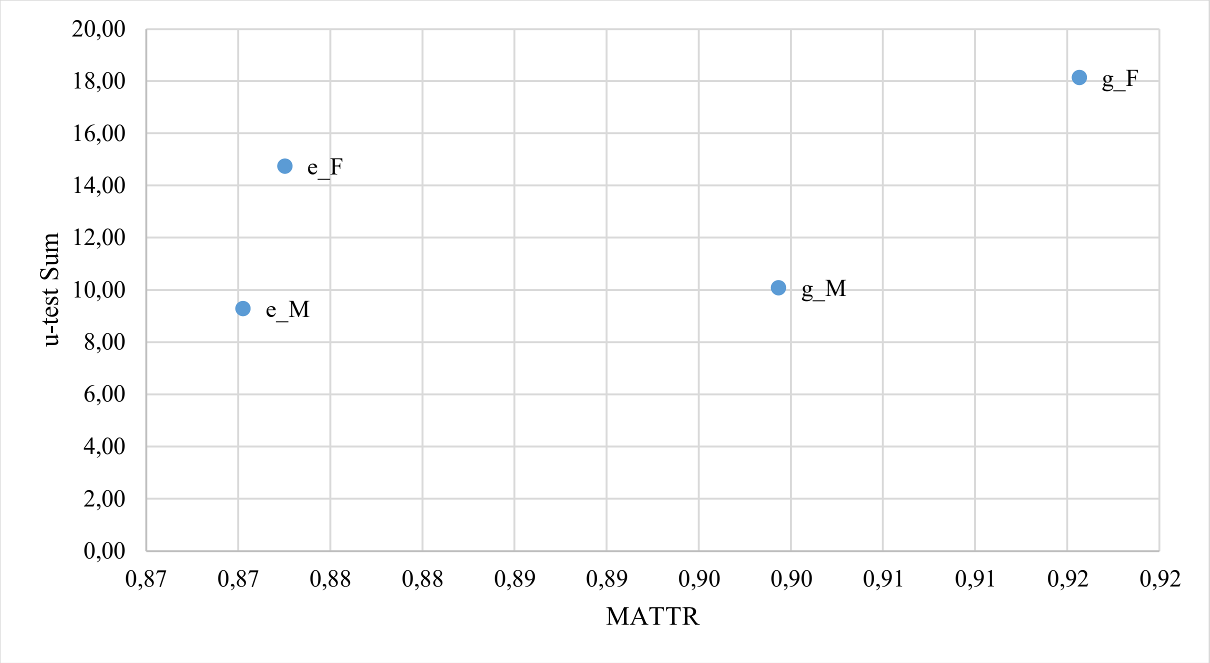

Figure 1. The scatter plot of the MATTR values for the studied text groups.

The values of the MATTR seem to respect the school division, providing one aspect as a premium – the significant difference between the grammar-school boys and girls. The values of the g_F group seem to be high very consistently, as the increased value of the u-test also means a low figure of the standard deviation (see the corresponding formula). The wide vocabulary range of the secondary-school girls is probably due to the aforementioned respect of theirs to the genre requirements. The use of adjectives and the emphasis the Czech stylistic tradition puts on lexical variety may have been the determining factors behind these figures. It is to be noted that the elementary-school girls do also manifest a slightly higher (and more consistent) score than the one of the boys, without opening such a gap as the grammar-school goers do.

Second, the outcomes for ATL will be interpreted.

Table 6

ATL: the values of u-test of the studied text groups

ATL	e_M	e_F	g_M
g_F	7.56*	8.99*	0.44
g_M	8.74*	10.46*	
e_F	1.84		

Table 7

The average ATLs and the sums of the u-test values for the studied text groups

	ATL	**u-test Sum**
e_M	4.25	18.14
e_F	4.14	21.29
g_M	4.82	19.64
g_F	4.86	17.00

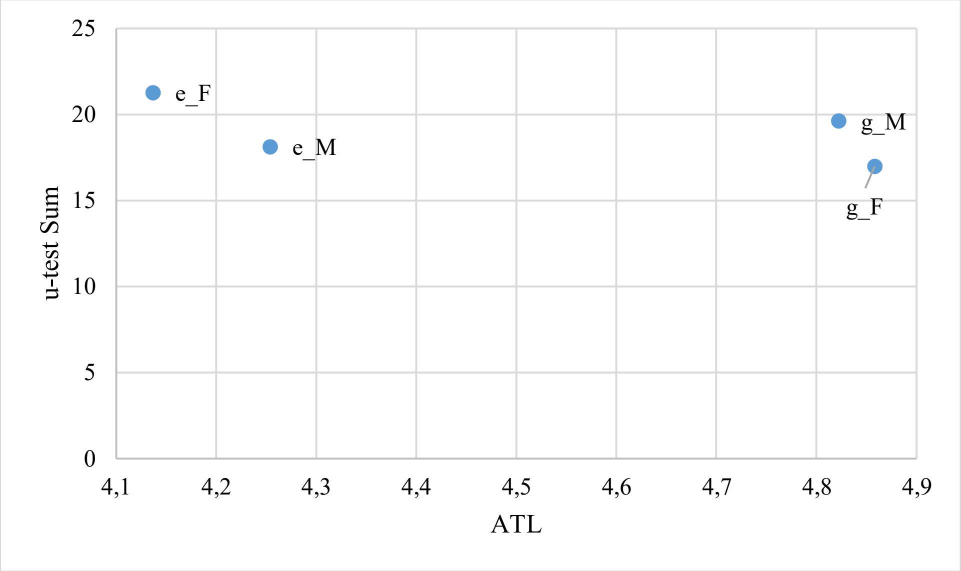

Figure 2. The scatter plot of the ATL values for the studied text groups.

In a way, the ATL situation is the easiest to grasp interpretatively. There are two distinct groups – the elementary-school participants, who manifest lower ATL values, and the grammar-school pupils, who boast higher numbers. The clear-cut situation points at the type of school as the decisive factor for the use of vocabulary; moreover, a textbook may also play part, as both the institutions use different ones. This is corroborated by the fact that the differences in the values of the same-school gender groups are not significant when compared to each other (e.g., "e_M" to "e_F"; see Table 6).

Third, an interpretation of the activity figures will be provided.

Table 8

Q: the values of u-test of the studied text groups

Q	e_M	e_F	g_M
g_F	5.76*	2.38*	0.10
g_M	5.91*	2.37*	
e_F	2.22*		

Table 9

The average Qs and the sums of the u-test values for the studied text groups

	Q	**u-test Sum**
e_M	0.64	13.89
e_F	0.53	6.97
g_M	0.43	8.38
g_F	0.43	8.24

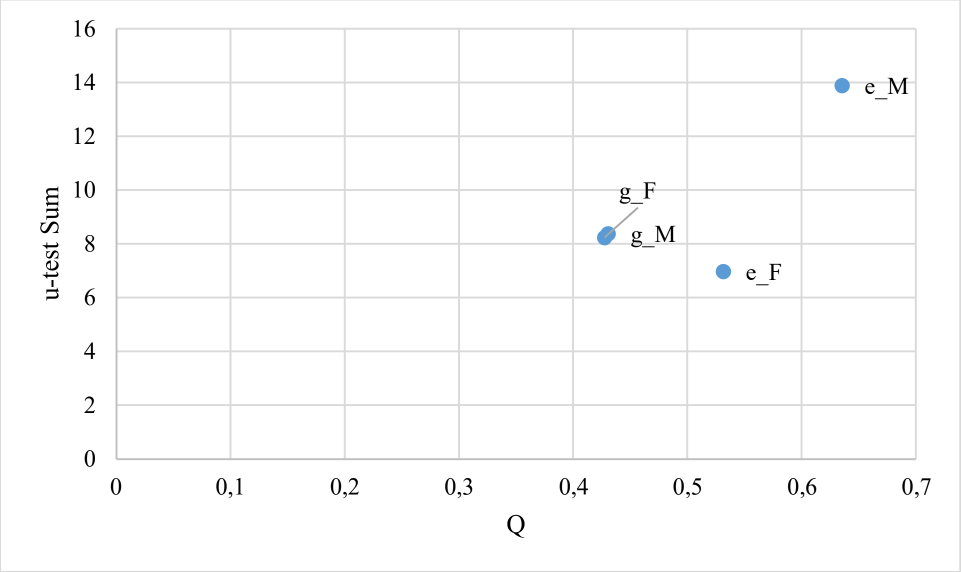

Figure 5. The scatter plot of the Q values for the studied text groups.

The activity values have produced an idiosyncratic picture. The values of the grammar-school people almost overlap; there is some variation to their data (the u-test sums being moderate), but in general, they represent the descriptive part of the corpus. The situation in the elementary-school results is more complicated: the girls seem to show uneven activity figures, whereas in the boys' essays, verbs prevail – considerably and steadily – over adjectives. This is a crucial outcome, as it may hint at the fact that the elementary-school males do not follow the principles of writing descriptions very much, as they prefer telling a "story of description" over the description itself. The reasons behind this outcome (e.g., lack of abstract thinking, a sort of disrespect to rules, etc.) will be more likely uncovered as soon as more research has been done in the field. Be that as it may, their descriptions present the outlier of the activity research.

Fourth and last, we will have a look at the verb distances values.

Table 10
VD: the values of u-test of the studied text groups

VD	e_M	e_F	g_M
g_F	4.07*	2.77*	2.12*
g_M	2.08*	0.71	
e_F	1.34		

Table 11
The average VDs and the sums of the u-test values for the studied text groups

	VD	**u-test Sum**
e_M	4.97	7.49
e_F	5.38	4.82
g_M	5.63	4.91
g_F	6.50	8.95

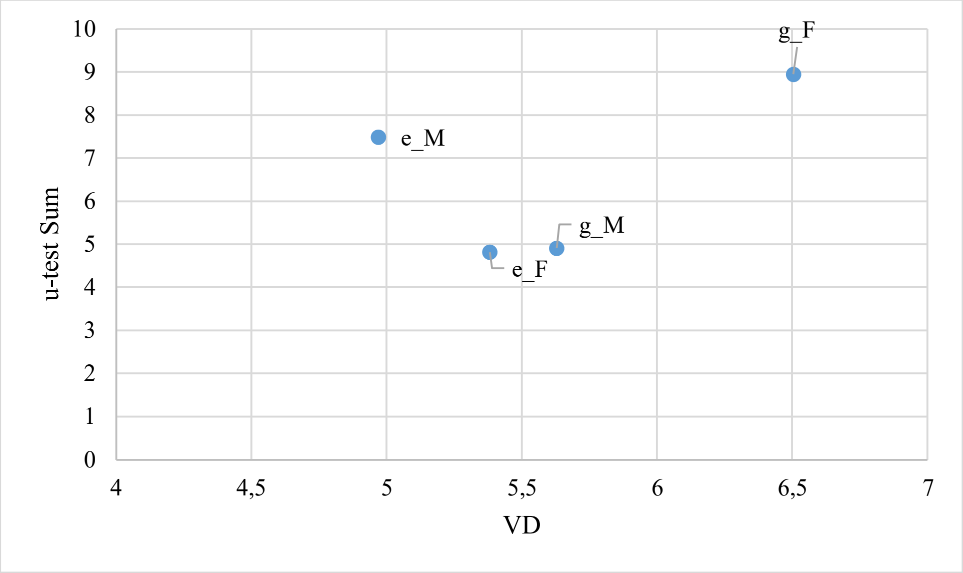

Figure 6. The scatter plot of the VD values for the studied text groups.

The results of VD do not seem to follow either of the factors; the samples are grouped haphazardly, in a constellation that has not been spotted in the preceding parts of the research. There are two outliers: the elementary-school boys, who use shorter verb distances, and the grammar-school girls, who (very consistently) boast complex syntactic structures. The results of the girls were awaited, as VD was the only index in which the boys and the girls significantly differed on a general basis. A more amount of variation is present in the elementary-school girls and the grammar-school boys, who, on average, manifest moderate numbers. The outcomes show that the biggest difference is between the syntax of e_M and g_F, or that there may be other factors than school and gender that determine the level of sentence complexity.

4.4 Research on Statistical Significance

Throughout the research, we have tested the statistical significance of the obtained results; however, these counts can also be regarded from the viewpoint of the individual indexes. Each of them has been used to calculate eight tests – one in the cross-school comparison, one in the cross-gender one, and six in the combined investigation. Table 12 sums up the proportion of the statistically significant tests to the totals of them for each index; it seems that they can be divided into two groups, according to the decreasing values.

As to our research, the texts appear to differ mostly on the grounds of activity and vocabulary richness; especially the former is surprising, as the description is a rather prescriptive genre when it comes to the use of adjectives and verbs. The differences indicate variated approaches to the principles of writing, and may be connected to the psychology of gender as well.

Table 12
The indexes from the viewpoint of statistical significance

	Number of significant differences	Proportion
Q	6	75.00%
MATTR	6	75.00%
ATL	5	62.50%
VD	5	62.50%

To conclude, the statistical significance will be investigated from the perspective of the studied groups. Each category (e.g., the elementary-school boys) has undergone three comparisons with the others per index; there being four indexes, this accounts for 12 comparisons in total. The upcoming table shows the percentage of those of these tests which yielded significant values.

Table 13
The categories from the viewpoint of statistical significance

	Number of significant differences	Proportion
e_M	9	75%
e_F	8	66.67%
g_M	9	75%
g_F	10	83.33%

The table ranks the text groups according to the amount of their differences from each other; in general, this is almost the same (9 ± 0.7), with slightly elevated numbers in case of the grammar-school children (they show 9.5 significant differences on average, compared to 8.5

differences of the elementary-school pupils). This can be explained by the special situation of the grammar-school girls, who occupy outlying positions in MATTR and VD. Their style of writing can thus be deemed the most discernible.

5. Conclusions

The summarizing remarks will be presented in the following points.

1. It has been found out that there are considerable differences between the researched schools. The statistical significance appeared in the values of the indexes of MATTR, ATL, Q, and VD. We may thus conclude that in general, the grammar-school children have more complex vocabulary and syntax, use longer words, and tend to be more descriptive.

2. As to the gender comparison, the girls have been proved to be more syntactically complex than the boys. This may be due to the attention they pay to observing the principles of the genre (exemplifications, use of adjectives, precision, etc.).

3. Concerning the research with the combined factors, each category will be treated separately.

a) The elementary-school boys tend to use shorter words, and less complex vocabulary and syntax; on the other hand, they prefer using verbal over adjectival description. This contradicts the expectations of the genre, and may have various reasons, which will be studied further.

b) The elementary-school girls, too, limit themselves to shorter words, and simple lexis and sentence structure. As to activity, they manifest middle values with a lot of variation.

c) The grammar-school boys, on the other hand, tend to score high in the length of words and vocabulary richness, though in the latter, the figures do variate. They are never outliers and mostly team up with various categories, sharing lower activity values with the grammar-school girls and middle figures in verb distances with the elementary-school girls.

d) The grammar-school girls share the employment of long words with the grammar-school boys, surpassing them, however, on the grounds of vocabulary richness. Given their low values of activity and an outstanding figure in verb distances, they seem to stick to the rules of the genre most firmly. Moreover, the high scores of the u-test sums in MATTR, ATL, and VD show compactness of the measured values.

4. Regarding the effectivity of the indexes, activity and MATTR are of the highest discriminatory value. Furthermore, the grammar-school groups display more statistically significant differences than the elementary-school ones, this being due to the specific style used by the grammar-school girls (see 3d).

Finally, it has to be stated that all the outcomes and their interpretations are but the first attempts to use stylometry in pedagogy; more research would be needed to come up with general results.

References

Andreev, S., Místecký, M., Altmann, G. (2018). *Sonnets: Quantitative Inquiries.* Lüdenscheid: RAM-Verlag.

Busemann, A. (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik. Sprachstatistische Untersuchungen*. Jena: Verlag von Gustav Fischer.

Covington, M. A., McFall, J. D. (2010). Cutting the Gordian Knot: The Moving Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100.

Čech, R. (2016). *Tematická koncentrace textu v češtině*. Praha: ÚFAL.

Čechová, M., Krčmová, M., Minářová, E. (2008). *Současná stylistika.* Praha: Lidové noviny.

Dai, Z., Liu, H. (2019). Quantitative Analysis of Queen Elizabeth II and American Presidents' Christmas Messages Over 50 Years (1967-2018). *Glottometrics*, 45, 63–88.

David, J., Davidová Glogarová, J., Radková, L., Šústková, H., Čech, R. (2014). *Slovo a text v historickém kontextu: perspektivy historickosémantické analýzy jazyka*. Brno: Host.

Hoffmannová, J., Homoláč, J., Chvalovská, E., Jílková, L., Kaderka, P., Mareš, P., Mrázková, K. (2016). *Stylistika mluvené a psané češtiny*. Praha: Academia.

Holubová, P. (2014). *Komparativní analýza jazykové a komunikační kompetence v psaných komunikátech studentů 4. ročníku SŠ. Hodnocení maturitních slohových prací ve vztahu k jejich jazykové úrovni.* Praha: Univerzita Karlova v Praze.

Kubát, M. (2016). *Kvantitativní analýza žánrů*. Ostrava: Ostravská univerzita.

Místecký, M. (2018). Belza Chains in Machar's *Letní sonety*. *Glottometrics*, 41, 46–56.

Popescu, I.-I. (2007). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.). *Exact methods in the study of language and text.* Berlin / New York: Mouton de Gruyter, 557–567.

Rysová, K. (2017). Charakteristika výsledků písemných maturitních prací z českého jazyka. *Český jazyk a literatura*, 68(4), 157–168.

Štěpáník, S., Holanová, R. (2017). K jazykovým a stylistickým dovednostem budoucích češtinářů. *Český jazyk a literatura*, 68(5), 230–239.

Zörnig, P., Místecký, M. (2018). Quantifying the Importance of Stylometric Indicators: A Principal Component Approach to Czech Sonnets. *Glottometrics*, 43, 11–30.

Glottometrics 49, 2020, 66–86

Tools for Semi-Automatic Analysis of Sound Correspondences: The *soundcorrs* Package for *R*

Kamil Stachowski[1]

Abstract. *soundcorrs* is a small R library of functions intended to facilitate computer-aided analysis of sound correspondences between languages. It is not designed to draw its own conclusions, merely to automate labour-intensive tasks and furnish the linguist with sifted and processed data for him or her to interpret. To make use of its basic functionality, *soundcorrs* requires only a very rudimentary knowledge of R, and no understanding of statistics at all. More advanced functions can be accessed and more involved results obtained after only a brief course of the two.

Keywords: software, sound correspondences, loanword adaptation, letter-grapheme-phone correspondences, phonology.

1 Introduction

The traditional method of investigating sound correspondences between languages is to spend long hours filling shoeboxes with flashcards, and then longer hours still excavating from them pairs of words which exemplify the currently discussed problem. The *soundcorrs* library can help reduce the time and effort involved in this kind of analysis. By design, it does not attempt to draw any conclusions on its own, a relatively recent practice which the more traditionally-minded linguists find well-justified reasons to distrust; it merely automates the management of data, acting more as a secretary than an assistant.

It is a library for R, which means that, although effort has been made to render it as easy to use as possible, some rudimentary knowledge of the language and the environment is required. The necessary topics include: basic data structures, function invocation, and variable assignment. A degree of understanding of regular expressions is highly recommended, and any additional knowledge can surely be put to good use, too. In the least effort scenario, simply repeating the example (sec. 4)] should prove useful, hopefully as an incentive to embark on a more in-depth exploration.

The *soundcorrs* library exports a number of functions, some serving an analytic purpose, others being just convenient helpers. The former are discussed in the sections below; the latter are only mentioned, but I trust that the documentation provided in the package will prove

[1] Jagiellonian University, ul. Gołębia 24, PL – 31-007 Cracow, Poland; https://orcid.org/0000-0002-5909-035X; kamil.stachowski@gmail.com.

sufficient for a researcher to utilize them. As a quick reference, the analytic functions are listed in tab. 1, organized by their output.

Tab.1.
A quick reference to *soundcorrs*' most important functions. For a full list, see the vignette (run `vignette ("soundcorrs")`).

Output	**Details**	**Function**	**Section**
contingency table	segment-to-segment	`summary()`	3.1
contingency table	correspondence-to-correspondence	`table()`	3.2
contingency table	correspondence-to-metadata	`table()`	3.2
contingency table	*n*-gram-to-*n*-gram	`ngrams()`	4
contingency tables	all, as a list	`allTables()`	3.2
fitting	one dataset, multiple models	`multiFit()`	3.3
fitting	multiple datasets, multiple models	`fitTable()`	3.3
n-grams	table with counts	`ngrams()`	3.3
pairs	single, unformatted	`findPairs()`	3.1
pairs and tables	all, formatted	`allPairs()`	3.1
segments	in relation to a correspondence	`findSegments()`	3.2

The present paper has been written primarily with loanword adaptation in mind. However, *soundcorrs* can be also used to explore other topics, both in purely qualitative linguistics (e.g., sound correspondences between related languages, morphological correspondences), as well as in more quantitatively-oriented research (e.g., grapheme-phoneme correspondences; Altmann/Fengxiang, 2008).

The organization of this paper is as follows: in sec. 2, data preparation; in sec. 3, a discussion of the most important of *soundcorrs*' functions, ordered from the more qualitative to the more quantitative approach to research; in sec. 4, a simple sample session with *soundcorrs*, and in sec. 5, a brief discussion of the errors and warnings issued by *soundcorrs*, as well as a caveat concerning encoding.

This paper describes *soundcorrs* as of version 0.1.1.

2 Preparation

Before work with *soundcorrs* can begin, the user needs to prepare a definition of the transcription or transcriptions in which the data are recorded, and the data themselves. Especially the latter can be a lengthy process, but it is unfortunately unavoidable. Both are described separately in subsections below.

One remark, however, is common to them, and it concerns encoding. The recommended choice is UTF-8. It has not been found to cause any issues under BSD, Linux, and macOS, but it has under Windows, and for this reason it is suggested that Windows users limit their transcriptions to plain ASCII. This is a harsh restriction; hopefully, future versions of *soundcorrs* will be able to do without it.

2.1 Transcription

There are two reasons why *soundcorrs* needs to know about the transcription in which the data are recorded. Firstly, without this knowledge, traditional linguistic regular expressions ("wildcards") would not be possible; and secondly, it allows to involve phonetics and other aspects in the analysis performed using *soundcorrs*.

The transcription is stored in a tsv file in the form of a table with two or three columns, as shown in Fig. 1, and can be read using the `read.transcription()` function. The only required argument is the name of the file; optionally, custom names for columns can also be provided. The return value is an object of class `transcription`.

```
GRAPHEME    VALUE                      META
b           cons,stop,lab,vd           b
p           cons,stop,lab,vl,mark1     p
f           cons,fric,lab,vl,mark1     f
B           cons,stop,lab              [pb]
P           mark1                      [pf]
-           NULL                       -
```

Fig. 1. Sample transcription file.

The first column, `GRAPHEME`, contains a list of graphemes. It is recommended that the transcription only employ single characters, as multigraphs may cause *soundcorrs* to yield unexpected and incorrect results, even if they are always isolated into separate segments. This restriction applies especially to metacharacters (“wildcards”). If a character missing from the Unicode is necessary, e.g. *b* with acute, it is generally recommended that it be not composed using a combining diacritical mark, but rather replaced with another single character, such as Б, ב, ب, etc. Characters used by R as metacharacters in regular expressions (`. + * ^ \ $ ? | ( ) [ ] { }`) cannot be used as graphemes. When reading a dataset, *soundcorrs* will warn about segments not covered by the provided transcription.

The second column, `VALUE`, contains comma-separated (without spaces) features of individual graphemes: phonetic attributes, formant frequencies, etc. The intention behind this column is to help analyze phonetics, but it is not actually necessary that the features be phonetic. They can be thought of as markers or labels required to generate the `META` column. Grapheme(s) which represent “linguistic zero” should be given the value of `NULL`. Such graphemes will be ignored, among others, by the function `findPairs()` if the argument `exact` is set to `FALSE` (which is the default).

The third column, `META`, is optional. If it is not given in the transcription file, *soundcorrs* will generate it automatically based on the `VALUE` column (the recommended method) – but if it is there, *soundcorrs* will not check its accordance with the `VALUE` column. The `META` column can be used to extend R’s in-built set of metacharacters to include symbols which are conventionally used in linguistics. Technically, the values in this column are regular expressions which are substituted for the characters from the `GRAPHEME` column when a query with `findPairs()`is performed. Graphemes which are not meant to be used as metacharacters, including the linguistic zero, should be simply repeated; those that are need to be expanded to an enumeration. For example, if ‹N› is to represent ‘any nasal consonant’, it should be expanded to `[mḿnńŋ]`, or whatever other set of nasal consonants is available in the given transcription. When the `META` column is generated automatically, the expansion will include all graphemes with values that form a superset of the value of the given grapheme; e.g. if `N` is given the value of `cons,nasal`, it will be expanded to an enumeration of all graphemes which contain both `cons` and `nasal` in the `VALUE` column.

Phonetic analysis is the primary intended use for a transcription, but it can as well be used for morphology, and perhaps other angles of inquiry as well. The general rule is that transcription should match segmentation (see sec. 2.2): if the latter is phonetic, so should the transcription be; if it is morphological, so should the transcription be.

2.2 Data

The same as the transcription, the data are stored in text files in the form of tables. They can be in what will be referred to as the *long format* or the *wide format*. The "long format" is a table with at least two columns: `ALIGNED` and `LANGUAGE`, and each entry occupying its own row, as in Fig. 2a. The "wide format" is perhaps less convenient for people, but it is used internally and required by *soundcorrs*. In it, corresponding pairs / triples / … of words are placed in a single row which, therefore, contains at least two columns, each holding words from a single language: `ALIGNED.x` and `ALIGNED.y`, or `LATIN` and `GERMAN`, etc. – as in Fig. 2b.

Data frames can be converted from one format to the other using functions `long2wide()` and `wide2long()`. Partial conversion is also possible, for metadata which are more conveniently viewed as describing entire pairs than individual words. For this purpose, `long2wide()`'s argument `skip` is used.

(a) The "long format".

```
LANGUAGE    WORD        ALIGNED

Latin       mūsica      m|ū|s|i|k|a
English     music       m|jū|z|i|k|-
German      Musik       m|u|z|ī|k|-
Polish      muzyka      m|u|z|y|k|a

Latin       prōvincia   p|r|ō|v|i|n|s|i|a
English     province    p|r|ɒ|v|i|n|s|-|-
German      Provinz     p|r|o|v|i|n|c|-|-
Polish      provincja   p|r|o|v|i|n|c|j|a
```

(b) The "wide format".

```
WORD.LAT    ALIGNED.LAT          WORD.ENG    ALIGNED.ENG
mūsica      m|ū|s|i|k|a          music       m|jū|z|i|k|-
prōvincia   p|r|ō|v|i|n|s|ia     provnice    p|r|ɒ|v|i|n|s|-|-

      WORD.GER  ALIGNED.GER        WORD.POL    ALIGNED.POL
      Musik     m|u|z|ī|k|-        muzyka      m|u|z|y|k|a
      Provinz   p|r|o|v|i|n|c|-    prowincja   p|r|o|v|i|n|c|j|a
```

Fig. 2. Sample data file.

The `ALIGNED` column contains words divided into segments using a fixed separator (`"|"`, by default). Segments can be simply single graphemes, but in some cases it may be more useful to separate entire morphemes or affixes into individual segments. It is necessary that all words in each pair / triple /… have the same number of segments and that the corresponding segments are aligned, though each segment can be composed of a different number of characters. Linguistic zeros (see sec. 2.1) can be used to create empty segments that are required to preserve the alignment. For example, if the German word *Junker* 'landowner' was rendered in Polish as *junkier* 'Prussian landowner, …' (de Vincenz − Hentschel, 2010), this can be encoded, e.g., as `j|u|n|k|e|r : j|u|n|ḱ|e|r`, but also `j|u|n|k|-|e|r : j|u|n|k|j|e|r` etc., depending on the preferred phonological interpretation. Segmentation and alignment can be either performed manually, or using one of the automated tools, such as *alineR*, *LingPy*, or *PyAline* (Downey – Sun − Norquest, 2017; List – Greenhill − Forkel, 2018; Huff, 2010). It is, however, recommended that their results be thoroughly inspected by a human, if for no other reason than to allow the researcher to acquaint themselves with the material and its specificity. *soundcorrs* only offers a very simple function `addSeparators()`, which intersperses a vector of words with a separator character, providing a convenient starting point for manual alignment. As was mentioned in sec. 2.1, segmentation does not necessarily need to be phonetic; it can follow morphology, or any other kind of boundaries.

The second column, `LANGUAGE`, contains the name of the language from which the given word has been taken.

Data files can contain any number of additional columns, e.g., for comments, references to sources, etc. Single rows in them can be hidden from *soundcorrs* by placing a number sign (#) at the beginning of the line. A `soundcorrs` object can also be subsetted using the function `subset.soundcorrs()`.

To allow a greater degree of flexibility, data from various languages are read in individually, using the `read.scOne()` function, which requires four arguments: the path to the file, the name of the language, the name of the column with the aligned words, and the path to the transcription file. If the segment separator is different from the default `"|"`, it should also be specified. Objects returned by `read.scOne()` can then be merged into a single `soundcorrs` object (see sec. 4). It is perfectly possible to create a `soundcorrs` object out of data for a single language, read the data out of a single file, only using different columns as the designated 'aligned' column.

It is not advised to store data from different languages in separate files, but it is possible. In such case, care needs to be taken to avoid conflicts between column names, and it is required that words from each language are recorded in the same order, or that each file contains a column with matching IDs (this column must have the same name in all the files). It is still recommended that the final, merged dataset be inspected before analysis.

3 Analysis

The *soundcorrs* library provides tools for linguistic analysis on the spectrum ranging from the traditional, purely qualitative approach to the more recent, statistical and wholly quantitative perspective. For easier orientation, they have been divided here into three uneven groups: the qualitative approach in sec. 3.1, the intermediate one in 3.2, and the quantitative one in 3.3.

Only the more important functions (cf. Tab. 1) are discussed in detail. Each description in sec. 3.1 and 3.2 is followed by a very basic example. The functions discussed in 3.3 cannot be illustrated so succinctly. Examples of their usage, as well as more complex examples of functions explained in sec. 3.1 and 3.2, are included within a sample *soundcorrs* session, which is presented in sec. 4. Further examples, as well as the documentation of all of *soundcorrs*' functions are available in the vignette (run `vignette("soundcorrs")`), and through the in-built help (run `?NameOfTheFunction`).

3.1 The qualitative approach

Let us begin with three entirely qualitative functions: `findPairs()`, which looks for examples of specific sound correspondences, `summary()`, which describes a single language or a dataset as a whole, and `allPairs()`, which produces a formatted listing of the entire dataset.

*

In practice, `findPairs()` is perhaps the most frequently used of all *soundcorrs* functions . What it does is it searches – in a dataset of two languages – for pairs of words which exhibit a specific sound correspondence. The function has three obligatory arguments: `data`, which is the dataset to be searched (a `soundcorrs` object), and `x` and `y`, which are the sounds to look for. It also has two optional arguments: `exact`, and `cols`.

The function operates in two modes: the exact mode (when the argument `exact` is set to `TRUE`), and the inexact one (when it is set to `FALSE`, or just omitted). In the exact mode, the comparison is performed on a strict segment-to-segment basis: `x` must occupy exactly the same segments as `y`, and both must be the entire segments. In the inexact mode, `x` is allowed to start or end one segment earlier or later than `y`, and they can be just parts of the specific segments. Let us use the following pair as an example: (`a|bc`, `ab|c`). In the exact mode, such a pair will only be matched in two cases (not counting queries with regular expressions): if `x=="a"` and `y=="ab"` – or if `x=="bc"` and `y=="c"`. The inexact mode, being more liberal, will also match this pair when `x=="a"` and `y=="a"`, when `x=="ab"` and `y=="a"`, and in several more cases. In addition, in the exact mode, linguistic zeros count as any other character would, while in the inexact mode, they are entirely disregarded.

In both modes, both `x` and `y` can be regular expressions: as provided by R (the default, 'extended' type, not Perl-like), or as defined in the transcription (see sec. 2.1). For example, to find all cases where *a* : *a* or *e*, one might run `findPairs (data, "a", "[ae]")`, and to find all cases of diphthongization in general, one might first define `V` to represent 'any vowel or semivowel', and then run `findPairs(data, "V", "VV")`. It should be noted that searching is performed on whole words, so, e.g., `findPairs(data, "a", ".*")` will cause *a* in the first language to be compared to the entire word in the second language – and will therefore only rarely return a match. To find all pairs in which one of the words contains a specific segment, regardless of what it corresponds to in the other word, `x` or `y` should be an empty string; for example, to find all pairs in which there is an *a* in the first language, without checking what it corresponds to in the second, one needs to run `findPairs(data, "a", "")`.

The function that translates metacharacters, as defined in the transcription, to regular expressions can also be used outside of `findPairs()`. It is called `expandMeta()`, and it takes two arguments: a `transcription` object, and the string to be transcribed. This first argument is necessary, but should it prove cumbersome in practice, a wrapper function can be defined as follows (assuming `ipa` is a `transcription` object): `expandIpa <- function(x) expandMeta(ipa, x)`.

The core of the return value of `findPairs()` is a subset of the provided data frame, which contains the matching pairs. By default, it is limited to only two columns which hold the aligned words, but this can be customized using the `cols` argument. It needs to be emphasized that pairs are only included once in the result, even when the specified correspondence appears multiple times in them (as, e.g., *lo* : *lo* in the German word *Haplologie* < Greek/Latin). For this reason, the number of rows in the result may be lower than the total number of occurrences of the given sound correspondence. The latter can be obtained using the function `summary()` [see below].

Technically, the return value of `findPairs()` is a list of class `df.FindPairs`. The mentioned subset is stored in the field named `data`; the `found` field holds a data frame with the exact positions of the matching segments; and the field named `which` is a vector of logical values which can be used to turn the output of `findPairs()` into a new `soundcorrs` object, as shown in sec. 4.

Example:

```
findPairs (sampleSoundCorrsData.abc, "a", "[ou]")
```

will look for all pairs in which L1 *a* : L2 *o* or L2 *u*, and print:

```
  ALIGNED.L1 ALIGNED.L2
3      a|b|c      o|b|c
4    a|b|a|c    u|w|u|c
```

*

The `summary()` function provides a more general overview by producing a contingency table of all the segment correspondences attested in a dataset. The default layout is with segments from the first language (= *L1*) in rows, and segments from the second language (= *L2*) in columns. The values represent in how many words the given correspondence occurs in the dataset. Thus, for example, to see all the renderings that L1 *a* has in L2, one would either issue `summary(data)` and look for the row named *a*, or run `summary(data)["a",]` and have only this row printed. And conversely, to see all the L1 sounds which yield L2 *a*, one would run `summary(data)[,"a"]`. (Assuming that *a* is always separated into an individual segment.)

The direction of the table can be modified using the argument `direction`. The default value is `1`, which is the "*x* yields *y*"-perspective, while `2` stands for "*y* stems from *x*". Another argument `summary()` can take is `unit`. This defines whether the values in the table represent the number of times, or the number of words in which the given correspondence occurs, i.e. whether *lo* : *lo* in *Haplologie* is counted twice or once, respectively. The accepted values are: `"o"`, `"occ"`, `"occurrence"`, `"occurrences"`, and `"w"`, `"wor"`, `"word"`, `"words"` (the default). Lastly, the argument `count` determines whether values in the table are absolute or relative – with relation to the entire row. The values it accepts are: `"a"`, `"abs"`, `"absolute"`, and `"r"`, `"rel"`, `"relative"`.

Example:

```
summary (sampleSoundCorrsData.abc)
```

will print a contingency table of segment-to-segment correspondences:

```
   L2
L1   a b c ə o u w
  - 0 0 0 2 0 0 0
  a 4 0 0 0 1 1 0
  b 0 5 0 0 0 0 1
  c 0 0 6 0 0 0 0
```

*

The function `allPairs()` combines `findPairs()` with `summary()` – and a little automation, in order to produce a nicely formatted digest of the entire dataset. Its output is very similar to the material part of many a work dealing with loanword adaptation or sound correspondences in general, such as Pekaçar (2006) or Pomorska (2018). It is divided into sections, one for each segment attested in either L1 (the default), or L2. Sections open with a table with counts of all the renderings of the given segment; they are followed by subsections, each listing all the pairs with the given rendering.

Like in `summary()`, the perspective can be switched by changing the `direction` argument from the default 1 ("*x* → *y*") to 2 ("*y* ← *x*"). Likewise, values in tables can represent either the number of words, or the number of occurrences of the given correspondence, and can either be absolute or relative (arguments unit and count). The listings of pairs are taken from the output of findPairs(), which means that by default, they are the aligned columns ALIGNED.x and ALIGNED.y, containing both linguistic zeros and separators (see sec. 2 above). While both can be easily removed in any text editor or word processor, it may be more convenient to use the cols argument to fine-tune the output of allPairs(), in the same way as it is used with findPairs(). By default, allPairs() will print to the screen, but it can be made to write to a file by setting the file argument to the desired path in place of NULL.

The output of allPairs() is formatted by a specialized function, defined through the formatter argument. The *soundcorrs* library provides three such functions: formatter.none(), which does almost no formatting at all (the default), formatter.html(), which outputs HTML code, and formatter.latex(), which returns LaTeX code. For users of LibreOffice, Microsoft Word, or another word processor, HTML may prove the most convenient option, as it can be opened in any web browser and simply copied to the processor without losing the formatting.

The provided formatters are not customizable, but it is not too difficult to write a custom one using one of them as a template (see sec. 4). Such a function needs to take at least three arguments: what, x, and direction. The last one is simply 1 or 2, the middle one is the data sent by allPairs(), and what defines the type of x as "section", "subsection", "table", or "data.frame". Additional arguments can also be used, and will be sent to the formatter function directly from the call to allPairs().

Example:

```
allPairs (sampleSoundCorrsData.abc)
```

will print all segment-to-segment correspondences, with only the most basic formatting:

```
section [1] "-"
table   ə
table   2
subsection   [1] "-" "ə"
data.frame     ALIGNED.L1 ALIGNED.L2
data.frame   5    a|b|c|-    a|b|c|ə
data.frame   6  a|b|a|c|-  a|b|a|c|ə
```

etc.

3.2 The intermediate approach

Next, functions which stand halfway between qualitative and quantitative linguistics – i.e., those which use statistical methodology, but return results which describe qualitative features – will be presented. Four functions are discussed in this subsection: table(), which builds contingency tables of sound correspondences; findSegments(), which creates metadata for use with table(); binTable(), which collapses tables to single correspondences; and allTables(), which acts as a wrapper for the two, with additional functionality.

*

First, the table() function will be discussed; as its name suggests, it generates contingency tables. Unlike summary(), however, it cross-tabulates not segments, but correspondences – with themselves or with metadata.

The default mode is the former. It is invoked by setting the argument column to NULL.

The output is a table in which both rows and columns are sound correspondences, and the values represent the number, or the percentage of times or of words in which they co-occur. The names of rows and columns are composed from segments and separated with an underscore, so, e.g., L1 *a* : L2 *e* would be notated `a_e`.

The other mode is invoked when the argument `column` is set to the name of one of the columns in the dataset. (As was mentioned in 2.2, the data are internally stored in the "wide format", i.e., with suffixes appended to column names, unless the `skip` argument was used with `read.soundcorrs()`.) The output in this mode is very similar, only columns hold the metadata from the indicated column, instead of correspondences. For `table()`, it is irrelevant whether the metadata are numeric, or categorical.

As in `summary()` [sec. 3.1 above], the units and the direction can be changed using the arguments `unit` and `direction`. They accept the same values, and the defaults are likewise `"w"` and `1`. Similarly, `table()` also takes the argument `count`, which accepts the same values and defaults to `"a"`. The difference is that its mode of operation with `summary()` was only a special case. As a general rule, the table is always divided into blocks: in the external mode, those blocks are made up of rows which share the same initial segment; in the internal mode, they are the intersection of rows which share the same initial segment, and of columns which share the same initial segment. For example, in `summary()`, each row summed up to 1; in the external mode of `table()`, all rows beginning with `"a_"` will sum up to 1, as will all rows beginning with `"b_"`, `"c_"`, etc.; on the other hand, in the internal mode of `table()`, the rectangle made of all rows beginning with `"a_"` and all columns beginning with `"b_"` will sum up to 1, while rows beginning with `"a_"` in their entirety will sum up to more – if, that is, L2 has more segments than just *b*. The reason for the distinction between absolute and relative counting is more clear with `table()` than it was with `summary()`. On their own, absolute numbers can be very misleading; for example, if it is found that L1 *a* : L2 *e* co-occurs multiple times with L1 *o* : *ö* but only rarely with L1 *u* : L2 *ü*, the reason may be that whatever palatalizing factor was present in those words, it does not affect *u*; but it may also be that *a* and *u* almost never appear in the same words in L1. With relative counting, the output may contain empty places. This means that the two segments just never appear together; their relative frequency is 0/0, which R represents as `NaN` and, in a table, prints as an empty space.

It should be noted that in the correspondence-to-correspondence mode, co-occurrence with itself is also counted. Values in tables produced in this mode may thus not be immediately understandable, especially if `unit` is set to `"o"`, and one or more of the correspondences appear multiple times in a single word (as in *Haplologie*). Let us consider two pairs of words: L1 *abc* : L2 *abc*, and L1 *aba* : L2 *aba*. In *abc*, we have three combinations: (*a*:*a*, *b*:*b*), (*a*:*a*, *c*:*c*), and (*b*:*b*, *c*:*c*). As shown in Fig. 3a, this would be represented as a 3×3 table (to accommodate all the three different combinations), filled with 1's (because each combination only appears once). With *aba*, the situation is in essence the same. We have three combinations: (*a*:*a*, *b*:*b*), (*a*:*a*, *a*:*a*), and (*b*:*b*, *a*:*a*), and would likewise use a 3×3 table filled with 1's. But here, because the first and last row and column are the same, we can simplify the table and combine the two *a*-rows and the two *a*-columns by simply adding them together, as is demonstrated in Fig. 3b. Hence, there are two co-occurrences of L1 *a* : L2 *a* with L1 *b* : *b* – (a_1 : a_1, *b* : *b*), and (*b* : *b*, a_2 : a_2) – and hence, there are four co-occurrences between L1 *a* : L2 *a* and itself.

(a) L1 *abc* : *abc*

	a:*a*	*b*:*b*	*c*:*c*
a:*a*	1	1	1
b:*b*	1	1	1
c:*c*	1	1	1

(b) L1 *aba* : *aba*

	a:*a*	*b*:*b*	*a*:*a*
a:*a*	1	1	1
b:*b*	1	1	1
a:*a*	1	1	1

→

	a:*a*	*b*:*b*	*a*:*a*
a:*a*	2	2	2
b:*b*	1	1	1

→

	a:*a*	*b*:*b*
a:*a*	4	2
b:*b*	2	1

Fig. 3. Counting the number of co-occurrences with `table()`.

Example:

```
table (sampleSoundCorrsData.abc)
```

will print a correspondence-to-correspondence contingency table:

```
        L1→L2
L1→L2  -_ə a_a a_o a_u b_b b_w c_c
  -_ə   2   2   0   0   2   0   2
  a_a   2   4   0   0   4   0   4
  a_o   0   0   1   0   1   0   1
  a_u   0   0   0   1   0   1   1
  b_b   2   4   1   0   5   0   5
  b_w   0   0   0   1   0   1   1
  c_c   2   4   1   1   5   1   6
```

*

The metadata used with `table()` can be virtually anything, including phonetics. The *soundcorrs* library provides the function `findSegments()` to help make use of this kind of data. In short, it generates a list of segments preceding or following the matches found by `findPairs()` [sec. 3.1]. It takes four arguments: `data`, `x`, and `y` – just like `findPairs()`, and, in addition, `segment`, which determines which segment to extract, in relation to the segments which realize the L1 *x* : L2 *y* correspondence. For example, to extract the segments which directly precede L1 *a* : L2 *e*, one would run `findSegments(data,"a","e",-1)`.

The output of `findSegments()` is a list of two vectors: one for segments taken from L1, and the other for those from L2. Both vectors are of the same length as the original dataset so as to be easily attachable to it: `data.new <- cbind(data, BEFORE.A.E=findSegments(data,"a","e",-1)$L1)`. Naturally, not every pair in the dataset must necessarily realize the *a* : *e* correspondence; those that do not are represented as `NA`'s. The lists produced by `findSegments()` can also be translated into phonetics using the function `char2value()`; an example of this is given in sec. 4.

Example:

```
findSegments (sampleSoundCorrsData.abc, "a", "u", 1)
```

will find segments that directly follow the L1 *a* : L2 *u* correspondence (cf. the example for `findPairs()` in sec. 3.1):

```
$L1
[1] NA     NA     NA     "c,b" NA     NA
$L2
[1] NA     NA     NA     "c,w" NA     NA
```

*

One of the possible uses for a contingency table is a test of independence. However, some of the most popular ones require that the sample be relatively large, and in linguistics, such data may not always available. The `binTable()` function attempts to alleviate this problem by selecting from a table only those rows and columns which are to be investigated, and combining (summing) all the others so that the initial table is reduce, as illustrated in Fig. 4. `binTable()` takes three arguments: `x`, which is the table or matrix to be collapsed, and `row` and `col`, which are the numbers of the rows and columns that are to be spared. The latter two can be single integers, or vectors of integers.

A side effect of this procedure is that the binned table contains one comparison where previously there were many, which makes it possible to ask more specific questions. One just needs to be careful to make sure that binning of given rows or columns makes sense from the linguistic point of view. For example, it may be reasonable to wish to compare how often L1 *a* : L2 *e* coincides with L1 *o* : L2 *ö*, versus all the other possible renderings of L1 *a* and *o*, but it will take quite specific circumstances to justify cross-tabulating these two correspondences against, e.g., all the other correspondences combined, including all the consonants, suffixes, and whatever else may have been separated into its own segment in the dataset.

	o:*o*	*o*:*ö*
a:*a*	10	1
a:*e*	2	10
a:*o*	3	0

→

	o:*o*	non-*o*:*o*
a:*a*	10	1
non-*a*:*a*	5	10

…

	o:*ö*	*non-o:ö*
a:*o*	0	3
non-*a*:*o*	11	12

Fig. 4. Binning of a 2×3 table into 2×2 tables.

Example:

```
binTable (table(sampleSoundCorrsData.abc), row=7, col=6)
```

will collapse the table that we saw above in the example for `table()` to its last but one cell:

```
          b_w non-b_w
c_c        1       19
non-c_c    2       46
```

*

The `allTables()` function automates the use of `table()` and `binTable()`, and generates a list of all contingency tables for the given dataset. The result has a form that can be easier to read for a person, but its primary intended use is to make easier the application of tests of independence – for which task the function `lapplyTest()` [see below] can be employed.

The function takes several arguments: `data` is, as usual, the dataset; `column`, `unit`, `count`, and `direction` work as with `table()` above; in addition, `bin` determines whether to bin through all the produced tables (defaults to `TRUE`), or whether to content itself with slicing the general contingency table (as produced by `table()`) into blocks devoted to single segments.

The return value of `allTables()` is a list containing all the resulting tables. It is named using the same logic as with `table()`. Specifically, if `bin` is `FALSE`, the names will be simply the segments attested in L1 or L2, depending on the value of `direction`; and when `bin` is `TRUE`, they will be composed of correspondences and values taken from `column` or, if that is `NULL`, correspondences again, all separated by underscores. For example, `allTables(data,"DIALECT")$a_e_D1` will hold the table for L1 *a* : L2 *e* with dialect D1, and `allTables(data)$a_e_o_ö` for L1 *a* : L2 *e* with L1 *o* : L2 *ö*. (Or with the languages swapped, if `direction=2`.) Cross-tabulations of correspondences with themselves are skipped, that is, e.g., L1 *a* : L2 *a* would be compared with the correspondences of L1 *b*, *c*, etc., but not with those of L1 *a* itself, which is why, e.g., the field `$a_a_a_o` (L1 *a* : L2 *a* × L1 *a* : L2 *o*) will be missing from the result.

Example:

```
allTables (sampleSoundCorrsData.abc)
```

will generate a binned table for each cell of the table we saw in the example for `table()` above – except, as explained above, for the mutually exclusive ones:

```
$`-_ə_a_a`
     a_a non-a_a
-_ə   2       4
$`-_ə_a_o`
     a_o non-a_o
-_ə   0       6
```

etc.

*

Another function is `lapplyTest()`, which applies a function to a list. The main difference between it and regular `lapply()` is the handling of warnings and errors. Its main intended use is to apply a test of independence to a list of contingency tables, such as produced by `allTables()`.

This function can take two or more arguments: `x`, which is the list of tables, `fun`, which determines what function is to be applied (the default is `chisq.test`), as well as all additional arguments to that function.

The return value is a list of the outputs of `fun`. It is of the `list.lapplyTest` class, so it can be passed to `summary()` to be turned into a brief overview of the results. In the report printed by `lapplyTest()`, only results below a specific *p*-value are included, with the default being 0.05. (It is for this reason that the return value of `fun` must contain an element named `p.value`, as it is from this field that the *p*-values are extracted. If the desired function has an incompatible return value, it will be necessary to write a wrapper around it). An exclamation mark at the beginning of a line in the output means that `fun` returned a warning. The specific message is attached to the given element of the list as an attribute named `"warning"`. If `fun` returned an error, the return value is a list with an attribute `"error"`.

The list returned by `lapplyTest()` preserves the naming scheme of the list passed to it as `x` so, for example, in order to see the result of the χ^2 test applied to the contingency table of L1 *a* : L2 *e* and L1 *o* : L2 *ö*, one can run

```
lapplyTest(allTables(data))$a_e_o_ö,
```

and to see if it produced a warning –

```
attr(lapplyTest(allTables(data))$a_e_o_ö, "warning").
```

In the case of the default chi-squared test, the message “Chi-squared approximation may be incorrect” often indicates insufficient data, and may be helped by the application of binning in the `allTables()` function.

With `allTables()` and `lapplyTest()`, as opposed to `table()` above, attention needs to be paid to whether the chosen test is compatible with the metadata (i.e., the values of the columns). Contingency tables are primarily used for categorical data, such as the names of the consultants who provided the given pronunciations, or the dialects they spoke. Numeric data, such as the year when the given word was recorded, may require an entirely different approach, perhaps one from beyond what is offered directly by *soundcorrs*.

Example:

```
res <- lapplyTest (allTables(sampleSoundCorrsData.abc))
```

will apply the χ^2 test to all the tables we saw in the example for `allTables()` above, and store the result in a variable called `res`. Then

```
summary (res)
```

will display a brief summary:

```
Total results: 34; with p-value ≤ 0.05: 7.
! -:ə with a:o: p-value = 0.014
! -:ə with a:u: p-value = 0.014
! -:ə with b:w: p-value = 0.014
  c:c with -:ə: p-value = 0.008
  c:c with a:o: p-value = 0.001
  c:c with a:u: p-value = 0.001
  c:c with b:w: p-value = 0.001
```

and

```
res$c_c_b_w
```

will display the result for the table we saw in the example for binTable() above:

```
   Chi-squared test for given probabilities
data:  tab
X-squared = 10.286, df = 1, p-value = 0.001341
```

3.3 The quantitative approach

Lastly, let us examine the three functions that will be mostly useful to quantitative linguists: `ngrams()`, which produces a table with counts of *n*-grams, and two functions which fit multiple models to one or more datasets: `multiFit()` and `fitTable()`.

*

Let us begin with `ngrams()`, a simple function that extracts *n*-grams or, more accurately, *n*-segments. The first argument is a `scOne` object (not a `soundcorrs` one, as is the case with nearly all the functions discussed here); the second is `n`, the length of subsequences to be extracted. Its default value is `1`, in which case `ngrams()` produces simply a frequency list of all segments; if it is larger than the number of segments in one of the words, that word is ignored in the final calculation. The third argument is `zeros`, which determines whether linguistic zeros (see sec. 2.1) are to be included (defaults to `TRUE`); the fourth is `as.table`, about which see below.

The return value of `ngrams()` are absolute counts of *n*-grams in the data for one language. The default format is a table. A table is legible, and it can be converted quite easily into a data frame with ranks (see the vignette; `vignette("soundcorrs")`), but it cannot be cross-tabulated with another language. For this purpose, the `as.table` argument can be set to `FALSE` to make `ngrams()` return the result in the form of a list; see an example of this in sec. 4. Note that cross-tabulation is only possible when the lists for both languages have the matching number of *n*-grams for each word – which is an alignment that setting the argument `zeros` to `FALSE` may destroy.

*

The function `multiFit()` fits multiple models to a single dataset. The first argument is a list of models. Each of its elements needs to contain two named fields: `formula`, and `start`. The latter contains the starting estimates for the fitting function. It is possible to include several sets of estimates, but even when there is only one, `start` needs to be a list of lists [e.g., `list(list(a=1))`]. The second argument is the dataset, in the form of a data frame or a list, or potentially any other that the fitting function accepts (see below). The column names in the dataset must correspond to the names given in the formulae in `models`. The third argument is the fitting function. It defaults to R's built-in `nls()`, but functions from external packages, such as `nlsLM` (Elzhov et al., 2016), might prove to be more convenient, especially when it comes to the accuracy of the starting estimates. Lastly, `multiFit()` can take some additional arguments and pass them to the fitting function.

The return value of `multiFit()` is a list with the outputs of the fitting function. When fitting failed to produce a result, and `multiFit()` suppresses the printing of errors, the value is `NA`, and the error or the warning are attached to it as attributes (in the same way that `lapplyTest()` does, see sec. 3.2). Technically, the output is of the `list.multiFit` class, so that it can be passed as an argument to `summary()` to produce a table for a more convenient comparison of the results. The metric can be set using the argument `metric`; the available options are: `"aic"`, `"bic"`, `"rss"` (the default), and `"sigma"`.

*

Similarly to `multiFit()`, `fitTable()` fits multiple models, the difference being that it fits them to multiple datasets. The first argument, `models`, is the same. The second, `data`, requires a matrix or a table, such as the ones produced by `summary.soundcorrs()` or `table.soundcorrs()` [sec. 3.1 and 3.2, respectively]. The third argument is `margin`, and it is nearly the same as with `apply()`: `1` for rows, or `2` for columns. The fourth argument, `conv`, is a function that will be applied to `data` in order to turn individual rows or columns (i.e., vectors) into data frames. The *soundcorrs* library offers three such functions: `vec2df.id` (only adds a column of subsequent numbers starting with 1; the default choice), `vec2df.hist` (creates a data frame from midpoints and counts extracted from a histogram), and `vec2df.rank` (sorts the data and adds a column with ranks). A custom function can be defined quite easily; it needs to take exactly one argument, a numeric vector, and return whatever format the fitting function can accept. The three converters provided by *soundcorrs* return data frames with two columns named X and Y. The custom function does not have to follow this convention, but the names must correspond to the variables given in the formulae in the models argument. To this, additional arguments can be added, which fitTable() will pass on to multiFit() [this includes the fitting function], and which multiFit() will then pass on to the fitting function.

The return value of fitTable() is a nested list with the outputs of the fitting function, or NA's. As with multiFit(), its class is list.multiFit, so it can be passed to summary() to generate a table for convenient comparison.

4 Example

Now, let us put all of the above into practice. The *soundcorrs* package contains two sample transcription files and three sample datasets. The transcription files are named trans-common.tsv and trans-ipa.tsv, and contain parts of the common tradition of linguistic transcriptions (as used in the Americanist phonetic notation, the Finno-Ugric transcription, and most others) and of the International Phonetic Alphabet. Neither are full, as they are intended only as samples, based on which users will be able to craft a set specifically to their needs. The sample datasets are data-abc.tsv, data-capitals.tsv, and data-ie.tsv. The first is entirely fabricated; the second contains the names of EU capitals in German, Polish, and Spanish[2] (linguistically, of course, it has no reason to be, for methodological reasons; it is only meant to serve as an example that stands on the common ground of a highly specialized field); lastly, the third contains a dozen words showcasing the Grimm's and Verner's laws (adapted from Campbell, 2013: 136f). Here, we will mostly use "abc", and leave the other two for the user to explore.

The three datasets are preloaded as sampleSoundCorrsData.abc, sampleSoundCorrsData.capitals, and sampleSoundCorrsData.ie, but here, we will read them from files. Let us assume that R and *soundcorrs* are installed, and begin by loading *soundcorrs* and the data. The paths to the sample files can be found using system.file().

```
# Load soundcorrs.
#    The warning is correct, but no cause for alarm.
library (soundcorrs)

# Find the path to a sample transcription.
path.trans.com <- system.file ("extdata", "trans-common.tsv",
      package="soundcorrs")
path.trans.ipa <- system.file ("extdata", "trans-ipa.tsv",
      package="soundcorrs")

# Find the paths to the two sample datasets.
path.abc <- system.file ("extdata", "data-abc.tsv",
      package="soundcorrs")
path.ie <- system.file ("extdata", "data-ie.tsv",
      package="soundcorrs")

# The "ie" set is in the wide format, it can be read as it is.
```

[2] I would like to express my gratitude to José Andrés Alonso de la Fuente, Ph.D. (Cracow, Poland), for his help with the Spanish data.

```
#     Different languages can use different transcriptions.
#     Regarding the warnings, see sec. 5.
d.ie.lat <- read.scOne (path.ie, "Lat", "LATIN", path.trans.com)
d.ie.eng <- read.scOne (path.ie, "Eng", "ENGLISH", path.trans.ipa)
d.ie <- soundcorrs (d.ie.lat, d.ie.eng)

# The "abc" set needs to be first converted to the wide format.
#     The "ID" column refers to pairs as a whole,
#     so it will not be converted.
tmp <- long2wide (read.table(path.abc,header=T), skip=c("ID"))
d.abc.l1 <- scOne (tmp, "L1", "ALIGNED.L1",
      read.transcription(path.trans.com))
d.abc.l2 <- scOne (tmp, "L2", "ALIGNED.L2",
      read.transcription(path.trans.com))
d.abc <- soundcorrs (d.abc.l1, d.abc.l2)
```

The calls to `read.scOne()` and `scOne()` will cause *soundcorrs* to show warnings about certain segments in both datasets not being covered by the transcription (cf. sec. 5). Since we will not be performing an in-depth phonetic analysis here, these warnings can be safely ignored. To inspect the loaded data, one can simply run `d.abc` and `d.ie`, which will print a brief summary, or to see all the individual examples, `d.abc$data`, and `d.ie$data`.

Thus prepared, let us proceed to simulate a brief working session with *soundcorrs*, including all the analytic functions, though not in the same order in which they were discussed in sec. 3.

```
# First let us prepare the material part of the paper,
#     printing all words in the appropriate orthography
#     rather than in the working, segmented form,
#     and format the output in HTML.
allPairs (d.abc, file="~/Desktop/abc.html",
      cols=c("ORTHOGRAPHY.L1","ORTHOGRAPHY.L2"),
      formatter=formatter.html)

# Now, let us see a general overview of the "abc" dataset as a whole.
summary (d.abc)

# Does counting words and occurrences make
#     a considerable difference?
summary (d.abc, unit="w") # same as above, since "w" is the default
summary (d.abc, unit="o")

# Let us take a closer look at "a" because this seems to be
#     the most complex one.
summary (d.abc, unit="o") ["a", ]

# Does it seem that the rendering of "a" may be
#     tied to some piece of metadata?
#     Let us create a convenience variable for column names.
myCols <- c("ALIGNED.L1","ALIGNED.L2","DIALECT.L2")

# Let us see the rounded correlates.
findPairs (d.abc, "a", "O", cols=myCols)

# Do there appear to be any regularities?
```

```
table (d.abc, unit="occ")
round (table (d.abc, unit="occ", count="rel"), 3)

# Perhaps “a” with “b”. Let us only see this part of the table.
tab <- table (d.abc, unit="occ")
rows <- which (rownames(tab) %hasPrefix% "a")-->
cols <- which (colnames(tab) %hasPrefix% "b")-->
tab [rows, cols]

# The number of examples is low, but the result seems promising.
chisq.test (tab[rows,cols])

# Will we be able to find any more?
tabs <- allTables (d.abc)
chisq <- lapplyTest (tabs)
summary (chisq)

# Unfortunately, L1 - (= linguistic zero) and L1 c
#     only correspond to L2 ə and c, so these results
#     provide little insight.
tabs$`-_ə_a_o`
tabs$`c_c_-_ə`

# Considering that this is only an example, and the number of
#     examples is very low, let us be generous.
chisq <- lapplyTest (tabs)
summary (chisq, p.value=0.3)
chisq$a_u_b_w
attr (chisq$a_u_b_w, "warning")
tabs$a_u_b_w

# Now let us look at the cases of apocope.
#     ‘exact’ must be explicitly set to ‘TRUE’ because in the
#     default inexact mode, linguistic zeros (here ‘-’) are ignored.
findPairs (d.abc, "-", "", cols=myCols)
findPairs (d.abc, "-", "", cols=myCols, exact=T)

# Does it appear in all the “southern” pairs?
apocopes <- subset (d.abc, findPairs(d.abc,"-","",exact=T)$which)
southern <- subset (d.abc, DIALECT.L2=="south")
identical (apocopes, southern)

# Is there any particular environment in which it appears?
#     “NA”’s mean that the word does not exemplify the given
#     correspondence – which cannot be the case here, since
#     we are using the “apocopes” set –, or that the segment that
#     is looked for falls outside of the word – as with the second
#     command here.
findSegments (apocopes, "-", "", -1)
findSegments (apocopes, "-", "", +1)

# The output can be easily turned into its phonetic value.
before.apocope <- findSegments (apocopes, "-", "", -1)
char2value (apocopes, "L1", before.apocope$L1)

# Most correspondences in “d.abc” seem to be quite one-sided.
```

```
summary (d.abc)

# Let us see if they follow a simple power law.
models <- list (
"model A" = list( formula="Y ~ X^a", start=list(list(a=-1))),
"model B" = list( formula="Y ~ a*X^b", start=list(list(a=1,b=-1))))
fit <- fitTable (models, summary(d.abc), 1, vec2df.rank)
summary (fit)

# Now, let us see if any patterns can be found in n-grams.
bigrams.l1 <- ngrams (d.abc.l1, n=2, zeros=T, as.table=F)
bigrams.l2 <- ngrams (d.abc.l2, n=2, zeros=T, as.table=F)
table (unlist(bigrams.l1), unlist(bigrams.l2))
```

Several more examples, together with an alternative explanation of the workings of each function, can be found in *soundcorrs*' vignette, which is accessible via `vignette("soundcorrs")`. Further details about each function are available in the package documentation, which can be accessed via `?findPairs`, etc.

5 Errors and how to solve them

Below is a near-comprehensive list of warning and error messages displayed by *soundcorrs* (abbreviated to *W* and *E*, respectively), together with brief explanations of possible causes and recommended solutions.

One issue that may appear without a message is encoding. Tests under BSD, Linux, and macOS did not reveal any problems with UTF-8, but they did under Windows. Some issues could be helped using `iconv()` to convert data from UTF-8 to UTF-8 (sic!), but other problems proved to be more resilient. Since no complete solution could be found, and a partial one would be misleading, *soundcorrs* does not contain any mechanism at all to remedy this situation. It is recommended that under Windows, only plain ASCII characters be used. This is quite unfortunate, and a priority for future versions of *soundcorrs*.

(E) At least two "scOne" objects are required. A `soundcorrs` object can only be created for two or more languages. It is perfectly acceptable, however, to pass data from one language, just with different 'aligned' columns and different names, as different 'languages'.

(E) Differing column names for different suffixes. The columns defined for one language do not match the columns defined for the other language. See sec. 2.2 and inspect the data file for typos in column names.

(E) Differing number of *X*. The data from two languages do not match. This error may occur when one converts between the "long" and "wide formats", or when one combines `scOne` objects into a `soundcorrs` object. If `X` is `columns` or `entries`, see sec. 2.2; if it is `segments`, check the specified lines in the data file and make sure that both words in the pair are divided into the same number of segments.

(E) Differing values between columns specified in "skip". A column listed in the `skip` argument of the `long2wide()` function must have identical values for each pair / triple / … of words. See 2.2 and inspect the data file for any mismatches.

(E) Extended regular expressions metacharacters are used as graphemes: … Characters `. + * ^ \ $ ? | ( ) [ ] { }` have special meanings in *R*'s extended regular expressions, and they cannot be used as graphemes. See sec. 2.1 and replace them with other characters in the transcription file.

(E) Extended regular expressions metacharacters are used in the data: … Characters `. + * ^ \ $ ? | ( ) [ ] { }` have special meanings in *R*'s extended regular expressions, and they cannot be used in the column with segmented words. See sec. 2.1 and replace them with other characters in the data file.

(E) Incompatible datasets. Perhaps conflicting column names? This error is shown when the user attempts to combine into a `soundcorrs` object two or more `scOne` objects with incompatible data in them. The reasons can be varied, e.g., mismatched column names or differing numbers of examples in the datasets. It is usually easier to spot inconsistencies when all the data are stored in the same file.

(E) It is required that 0 < *row* ≤ nrow(*x*) and 0 < *col* < ncol(*x*). This error is shown by `binTable()` when the argument `row` or `col` is beyond the limits of the specified table `x`. The number of rows and columns of `x` can be checked with `nrow(x)` and `ncol(x)`, respectively.

(E) Linguistic zero is not defined in the transcription. Some of *soundcorrs* functions, e.g. `findPairs()`, can only work correctly if the transcription defines a symbol to denote linguistic zero. Sec. 2.1 explains how to do it.

(E) Linguistic zeros must be separate segments: … In general, *soundcorrs* allows multiple characters within a single segment, but the character used to denote linguistic zero is an exception. Linguistic zeros can only fulfil their intended role in *soundcorrs* when they are isolated into separate segments of their own.

(E) Multiple definitions for graphemes: … The same character is defined more than once in the transcription file. The specified graphemes should be checked and redundant definitions removed.

(E) One or more column names are missing from *X*. (Variants: Column *Y* is missing from *X*.) The column names given as the argument to the function cannot be found in the given dataset. Inspect the line for typos, or check the available columns by running `colnames(X$data)`.

(E) The specified "language" is missing from *X*. The value of the argument `language` to the function `char2value()` is not compatible with the dataset *X*. To check the available names, simply run `X`.

(E) This function does not know how to handle an object of class *X*. The given function has only been defined for objects of the `scOne` or `soundcorrs` classes. If the user would like to define it for a different class, this should not collide with the working of *soundcorrs*.

(E) Transcription metacharacters are used in the data: … Characters defined in the transcription as metacharacters cannot be used in the column with segmented words. See sec. 2.1 and replace them with other characters in the data file.

(E) *X* cannot be empty string or NA. Argument `X` needs to be given a concrete value in order for the function to be able to perform its function. The user is directed to the documentation of the specific function (run `?NameOfTheFunction`).

(E) *X* must be exactly one column name. Only one column can be designated as the one that holds the segmented words. It is perfectly possible to have in a dataset multiple columns with segmented words, potentially each segmented according to a different set of rules, but they must be all read into separate variables.

(E) *X* must be *Y*. (Variants: *X* must be of class *Y*, *X* must refer to *Y*.) Argument `X` can only take values of certain type or from a limited range. That range may be specified in the warning message in full or, in the 'refer'-variant, in short. The user is directed to the documentation of the specific function (run `?NameOfTheFunction`).

(W) Missing the metacharacters column. The "*X*" column was generated. Leaving the generation of metacharacters to *soundcorrs* is the recommended method, but at the same time, it is always safest to check manually any content that has been generated automatically. Transcription is easy to find and explore – it is simply a data frame passed as the output of the `read.transcription()` function.

(W) The following object is masked from 'package:base': table. This should not interfere with the working of `table()` for all applications beyond *soundcorrs*.

(W) The following segments are not covered by the transcription: … The data contain segments that are not listed in the transcription. However, not all tasks that can be performed with *soundcorrs* require the transcription. As long as it is not explicitly made use of, this warning can be safely ignored. See 2.1.

(W) This function only supports two languages. As of version 0.1, some functions provided by *soundcorrs* only accept datasets of no more than two languages. This is planned to change in future releases.

6 Summary

The paper presents an R library by the name of *soundcorrs*. The goal of this package is to facilitate the analysis of sound correspondences between languages by automating the most tedious of tasks involved in this kind of investigation. In opposition to cladistics and other computerized methods that have gained a degree of popularity among linguists in recent years, *soundcorrs* is not intended to produce any conclusions, merely to extract information from a dataset while leaving interpretation entirely to the user.

This paper discusses in some detail the most important functions, the applications of which range from a purely qualitative approach to a more quantitatively-oriented one. It also presents a sample session with *soundcorrs*, and explains the meaning of warnings and errors issued by *soundcorrs*.

Plans for future versions of *soundcorrs* include: solution of the problem of encoding under Windows, implementation of support for multiple languages in those functions which currently only accept pairs of languages, addition of functions to simulate phonetic changes, and more. Users are asked to address all requests, as well as bug reports, to this author. If *soundcorrs* is used in published research, please cite this paper.

References

Altmann, G. & Fengxiang, F. (2008). *Analyses of Script. Properties of Characters and Writing Systems* (= *Quantitative Linguistics* 63). Berlin – New York: Mouton de Gruyter.

Campbell, L. (2013). *Historical Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.

Downey, S. S. & Sun, G. & Norquest, P. (2017). alineR: an R Package for Optimizing Feature-Weighted Alignments and Linguistic Distances. *The R Journal* 9(1), 138–152.

Elzhov, T. V. & Mullen, K. M. & Spiess, A.-N. & Bolker, B. (2016). *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds*. https://CRAN.R-project.org/package=minpack.lm.

Huff, P. (2010). *PyAline*. http://pyaline.sourceforge.net.

List, J.-M. & Greenhill, S. & Forkel, R. (2018). *LingPy. A Python library for historical linguistics*. http://lingpy.org.

Pekaçar, Ç. (2006). Kumuk Türkçesine Arapça ve Farsçadan Geçen Kelimelerdeki Ses Olayları. *Selçuk Üniversitesi Türkiyat Araştırmaları Dergisi* 19, 53–71.

Pomorska, M. (2018). *Russian Loanwords in the Chulym Turkic Dialects*. Kraków: Księgarnia Akademicka.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

de Vincenz, A. & Hentschel, G. (2010). *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache* (= *Studia Slavica Oldenburgensia* 20). Oldenburg: BIS-Verlag.

Glottometrics 49, 2020, 87–97

Fitting the Distribution of the Syllabic Types in Different Positions of Verse

Vadim Andreev[1]

Abstract. The article is devoted to the study of distribution of syllabic types in three long poems by A. Pushkin, one of the founders of Russian literature. Contrary to the usually practised "linear" approach, where syllabic types are viewed as a consistent sequence in which the position of every syllabic type in a poetic line is not taken into account, the present research is focused on the "vertical" arrangement of syllables in poetic texts. In this case, the sequence under study includes syllables which occur in the same position in different lines of a poem. To reveal the peculiarities of such distributions, the Zipf-Alekseev function, which gives a good fit, and repeat rate indicator are used.

Keywords: Zipf-Alekseev function, repeat rate indicator, syllabic types, distribution, long poems.

The study of sequences of different types of syllables in poetry has demonstrated an evident order in their distribution, proving that the distribution of frequencies of syllabic types in the text is not random (Zörnig, et al. 2019).

Such research of regularities in the distribution of syllabic types has been mainly focused on the "horizontal" arrangement of syllables. This means that types of syllables were counted from the beginning of the poem to its end; the researcher successively moves from line to line without paying attention to the metric positions in these lines. In other words, all verse lines are viewed as a single sequence beginning with the first syllable in the poem and ending with its last syllable.

One of the most important peculiarities of poetic text (verse) is its double nature: its structure is organized not only in the linear, horizontal direction, as it is usual in prose, but also "vertically". Verse text is divided into lines, which have similar features. These features, when repeated, ensure a certain resemblance of verse lines. Among features supporting vertical relations between poetic lines – such as rhyme, poetic syntax (syntagmatic pauses, enjambments, syntactic links), assonance, etc. –, the most powerful and effective means of creating such similarity are metre and rhythm[2]. This raises the question of finding out if there is any regularity in the distribution of syllabic types in the same rhythmic positions in different lines.

For the purposes of the present investigation of 8–9 metric positions in the iambic tetrameter, we shall distinguish strong positions (ictuses), on which the stress should fall, and weak positions (metrically unstressed).

[1] Smolensk State University, ul. Przhevalskogo, 4. Smolensk 214000. RF. E-mail: vadim.andreev@ymail.com.

[2] By metre, we understand a syllabic pattern of a line which is characterized by the number of syllables and regularity of stressed positions (ictuses), whereas rhythm is a concrete realization of metre in a line with possible deviations from its metric scheme.

The database includes 3 long poems by A. Pushkin: *Graf Nulin* ("Earl Nulin"), *Ruslan i Lyudmila* ("Ruslan and Ludmila"), *Mednij Vsadnik* ("The Bronze Horseman"). Out of each of these poems, 200 lines were taken.

In the samples, the following 17 types of syllables were found (V – vowel, C – Consonant): V, VC, CV, VCC, CVC, CCV, VCCC, CVCC, CCVC, CCCV, CVCCC, CCVCC, CCCVC, CCCCV, CVCCCC, CCCVCC, CCCCVC.

To illustrate the horizontal and vertical distributions, the first four lines from the long poem *Graf Nulin* are taken. In the coded form, the lines can be represented as it is done in Table 1. Columns are the consecutive number of a syllable in the poetic line (its position in the line), lines of the table represent the first four poetic lines from the poem.

Table 1

Syllabic types in the positions of the first 4 lines
of the long poem *Graf Nulin*

	Position in the line								
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
Line 1	CV	CV	CV	CV	CV	CV	CCV	CVC	
Line 2	CCV	CV	CV	CVC	CV	CVC	V	CV	CVC
Line 3	CVC	CCVC	VC	CV	CV	CVC	CV	CVC	
Line 4	CVC	CV	CV	CCV	CV	CVC	CV	CCV	CVC

The count of all the types in these four lines (horizontal dimension) gives the following frequencies (down-ranked): CV = 17, CVC = 10, CCV = 4, V, CCVC and VC = 1 each. The vertical count of the types within 9 separately taken positions brings about the following frequencies (Table 2).

Table 2

Vertical distributions of syllabic types in the first 4 poetic lines of *Graf Nulin*

	Position in the line								
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
CV	1	3	3	2	4	1	2	1	0
CVC	2	0	0	1	0	3	0	2	2
CCV	1	0	0	1	0	0	1	1	0
V	0	0	0	0	0	0	1	0	0
VC	0	0	1	0	0	0	0	0	0
CCVC	0	1	0	0	0	0	0	0	0

The 9th position is observed only when the author used the feminine rhyme. In this extract, it happens twice. If one is interested to study the actual end of all the lines, they should combine the syllables in the 8th position in masculine rhymes and the 9th-position syllables in feminine rhymes. Thus, in our example, the last syllable count will give the following: CVC – 4 cases (2 in masculine lines and 2 in feminine lines).

The analysis was carried out in the following way. Firstly, the horizontal dimension was studied, and then, it was compared to the vertical one.

The distribution of frequencies is fitted by the Zipf-Alekseev function (Hřebíček, 2002):

$$(1) \qquad f_x = f_1 x^{a+b*\ln x} \,,$$

where f_1 is the frequency of syllables in $x = 1$, a and b – parameters, x – the frequency of syllabic types.

The results of the fitting are shown in Table 3. It presents the observed frequencies of the types in different metrical positions and theoretically expected frequencies, which are calculated according to the Zipf-Alekseev function. The frequencies are down-ranked.

Table 3

Frequencies of the syllabic types in metric positions of *Graf Nulin* and fitting the Zipf-Alekseev function to the sample

Types	Observed	Expected
CV	766	766.00
CVC	542	524.40
CCV	136	194.76
CCVC	102	68.37
V	82	24.93
VC	45	9.62
CCCVC	11	3.94
CVCC	8	1.70
CCCV	8	0.77
a = 0.65, b = -1.73, R^2 = 0.984		

As seen in the table, the result of the fitting is very satisfactory (98.40 %). The same holds for two other long poems (Table 4).

Table 4

Frequencies of the syllabic types in metric positions of *Ruslan i Ludmila* and *Medniy Vsadnik* and fitting the Zipf-Alekseev function to the samples

Ruslan i Ludmila			*Medniy Vsadnik*		
Types	Observed	Expected	Types	Observed	Expected
CV	718	718.00	CV	750	750.00
CVC	571	556.23	CVC	523	506.49
CCV	148	200.02	CCV	163	209.35
CCVC	104	65.48	V	97	83.47
V	64	22.02	CCVC	78	34.60
VC	54	7.82	VC	40	15.12
CCCVC	14	2.94	CVCC	14	6.96
CVCC	12	1.17	CCCV	14	3.36

CCCV	12	0.49	CCCVC	11	1.69
CCVCC	1	0.21	CCVCC	2	0.88
CCCCV	1	0.10	CCCVCC	1	0.48
CVCCC	1	0.05	CCCCVC	1	0.27
a = 0.99, b = -1.96, R^2 = 0.986			a = 0.45, b = -1.47, R^2 = 0.992		

At the second stage of analysis, syllabic types in the same metrical position in all the lines were counted (vertical dimension). As has been mentioned above, there are 8 positions in all the lines of the analyzed long poems and an additional one – 9th position – in those lines which have feminine rhymes.

Table 5 presents the observed and theoretically expected (according to the Zipf-Alekseev function) frequencies of the syllabic types in *Graf Nulin*.

Table 5

Fitting the Zipf-Alekseev function to *Graf Nulin* (vertical dimension)

Rank	Position 1		Position 2		Position 3		Position 4	
	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp
1	69	69.00	86	86.00	112	112.00	82	82.00
2	35	41.56	83	82.80	47	44.90	72	71.02
3	31	28.02	11	12.90	14	19.63	22	25.64
4	26	20.27	8	1.52	11	9.58	12	8.19
5	21	15.38	5	0.18	8	5.09	4	2.66
6	14	12.08	4	0.02	6	2.89	3	0.91
7	3	9.74	2	0.00	1	1.74	3	0.33
8	1	8.01	1	0.00	1	1.09	1	0.13
9			1	0.00			1	0,05
	a = -0.58 b = -0.22 R^2 = 0.934		a = 2.80 b = 4.12 R^2 = 0.991		a = -0.86 b = -0.66 R^2 = 0.994		a = 1.25 b = -2.10 R^2 = 0.995	

Rank	Position 5		Position 6		Position 7		Position 8	
	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp
1	88	88.00	94	94.00	108	108.00	83	83.00
2	76	75.48	64	62.47	39	42.35	77	75.98
3	12	15.60	13	19.36	28	21.73	19	23.65
4	11	2.62	11	5.55	13	12.82	13	6.27
5	5	0.45	10	1.66	8	8.26	4	1.68
6	5	0.08	5	0.53	3	5.66	3	0.48

7	2	0.02	1	0.18	1	4.05	1	0.14
8	1	0.00	1	0.07				
9			1	0.03				
	a = 2.09 b = -3.34 R^2 = 0.985		a = 0.86 b = -2.09 R^2 = 0.982		a = -1.16 b = -0.27 R^2 = 0.992		a = 1.61 b = -2.50 R^2 = 0.990	

Rank	Position 9		All final syllables	
	Obs	Exp	Obs	Exp
1	53	53.00	104	104.00
2	44	44.00	72	71.74
3	3	3.00	14	15.66
4			6	2.97
5			4	0.58

	a = 3.74
	b = -5.78
	R^2 = 1.000
			a = 1.49
			b = -2.93
			R^2 = 0.997	

The long poem *Graf Nulin* was written by Pushkin in 1825. It is possible to compare these results with the results of the similar analysis of two poems of the same author – one, written earlier, in 1818–1820 (*Ruslan i Ludmila*), and the other, written later, in 1833 (*Medniy Vsadnik*).

The results of the fitting of the Zipf-Alekseev function to the two samples are presented in Tables 6 and 7.

Table 6

Fitting the Zipf-Alekseev function to *Ruslan i Ludmila*

Position in the line	*a*	*b*	R^2
1	-0.62	-0.24	0.949
2	1.57	-2.32	0.994
3	0.84	-2.18	0.990
4	0.76	-1.46	0.988
5	1.33	-2.36	0.980
6	0.62	-1.75	0.977
7	0.24	-1.26	0.997
8	2.16	-3.34	0.987
9	5.23	-8.10	1.000
All final syllables	2.35	-4.15	0.999
Horizontal dimension	0.99	-1.96	0.986

Table 7
Fitting the Zipf-Alekseev function to *Medniy Vsadnik*

Position in the line	*a*	*b*	R^2
1	-0.08	-0.49	0.944
2	1.08	-2.21	0.987
3	0.42	-1.42	0.999
4	1.24	-2.06	0.982
5	0.19	-1.67	0.990
6	1.38	-2.29	0.981
7	-0.18	-0.91	0.995
8	2.43	-3.77	0.989
9	0.87	-3.22	0.999
All final syllables	3.29	-4.93	0.995
Horizontal dimension	0.45	-1.47	0.992

Very good results of the fitting may be recognized as rather unexpected. In verse, where there exist semantic, morphological, and syntactic links and interconnections between both words and text structures (not to speak of the rules of purely poetic restrictions), and where, on the other hand, there are numerous rules of possible deviations from the metric scheme, it was difficult to expect any order in the vertical arrangement of syllabic types, not to speak of the order which is similar to the order of their distribution in the horizontal dimension.

Since all the three poems were written in the iambic tetrameter, the stressed positions (ictuses) predominantly fall on the 2nd, the 4th, the 6th, and the 8th syllables. Comparing the syllabic types distribution in these strong positions with those observed in the unstressed positions (1st, 3rd, 5th, 7th, and 9th), we do not see any difference, except that the first position displays a little lower values of R^2. Contrary to the beginning of the line, the final syllables distribution (all final syllables) is fitted very well.

The long poems chosen for the analysis are different not only in the year of their creation, but also in the circumstances of the author's life.

Ruslan and Ludmila, one of the first literary works which made Pushkin famous, was written at his early age – he began working on it during his studies at a lyceum and finished the poem soon after the completion of his studies. This romantic poem combines the style of poetic ballads possessing heroic, tragic, and satiric themes. It was written in 1818–1820 (with additional parts in 1825 – they were not included in our sample).

Graf Nulin is a poem of a highly humorous nature with a frivolous plot and a large number of colloquial words. It is one of the first works of the author in the realistic style, and was written by Pushkin in exile during two mornings in 1825.

Medniy Vsadnik is one of the most celebrated works by Pushkin, which he wrote during a very creative period of his life in 1833. Working on the poem, Pushkin made tremendous efforts to achieve an ideal form, rewriting all its parts many times (some lines – up to ten times). The chosen sample contains emotionally elevated and lofty lexis, describing the statuesque beauty of the capital of the empire.

Thus, we see that there are very strong divergences among the texts in quite a number of aspects (genre, the age of the author, plot, time spent writing the poems, lexis, style), but the general tendencies of the distribution of syllabic types in the same positions hold good for all of them.

One more important aspect of characterizing the distribution of the elements in text consists in finding out their density. For this purpose, an indicator of repeat-rate was suggested (Altmann, Kohler, 2015; Herfindahl, 1950). It is computed as follows (Andreev, Místecký, Altmann, 2018: 96):

(3) $$R = \sum_{i=1}^{k} p_i^2 \,;$$

for p_i, we get:

(4) $$p_i = \frac{x_i}{N},$$

where k is the number of types, p is the relative frequency of the given type, x_i is the frequency of syllabic type, and N is the total number of all syllables in the same position in the vertical sequence.

It is also recommended to relativize R using the formula (Andreev, Místecký, Altmann, 2018: 97):

(5) $$R_{rel} = \frac{1 - R}{1 - \frac{1}{N}}.$$

The frequencies of the types of syllables in different positions of the lines in *Graf Nulin* were given in Table 2. The calculated values of the repeat-rate indicator for the poems are presented below (Tables 8–10).

Table 8
Relativized values of the repeat-rate indicator for types of syllables
in different metric positions of *Graf Nulin*

Positions	R	R_{rel}	Positions	R	R_{rel}
P1	0.207	0.797	P6	0.334	0.670
P2	0.363	0.640	P7	0.355	0.648
P3	0.379	0.624	P8	0.334	0.669
P4	0.314	0.689	P9	0.475	0.530
P5	0.346	0.657	Final (PF)	0.406	0.597

Table 9

Frequencies of types of syllables in different metric positions in *Ruslan i Ludmila* and relativized values of the repeat-rate indicator

Types	P1	P2	P3	P4	P5	P6	P7	P8	P9	PF
V	34	1	6	2	0	11	9	1	0	1
CV	72	78	99	66	86	90	92	79	56	75
CCV	26	23	14	19	16	11	25	13	1	2
CVC	32	77	63	76	71	62	60	87	43	108
CCVC	13	13	7	24	11	14	10	12	0	9
CVCC	1	2	0	1	4	2	0	2	0	0
CCCV	5	0	2	0	2	2	1	0	0	0
VC	16	2	8	5	10	7	3	3	0	3
CCVCC	0	1	0	0	0	0	0	0	0	0
CCCVC	1	3	1	5	0	1	0	3	0	2
CCCCVC	0	0	0	1	0	0	0	0	0	0
CVCCC	0	0	0	1	0	0	0	0	0	0
N	200	200	200	200	200	200	200	200	100	200
R	0.212	0.318	0.353	0.278	0.323	0.311	0.322	0.354	0.499	0.435
R_{rel}	0.792	0.685	0.650	0.725	0.680	0.692	0.681	0.650	0.507	0.568

Table 10

Frequencies of types of syllables in different metric positions in *Medniy Vsadnik* and relativized values of the repeat-rate indicator

Types	P1	P2	P3	P4	P5	P6	P7	P8	P9	PF
V	40	9	10	3	12	12	9	2	0	0
CV	62	90	91	71	109	79	95	89	64	93
CCV	34	16	25	16	12	17	30	10	3	4
CVC	31	67	62	79	57	70	53	79	25	85
CCVC	15	9	6	19	4	7	7	11	0	8
CVCC	1	0	0	3	0	4	0	4	2	6
CCCV	3	0	3	4	0	1	3	0	0	
VC	12	4	3	4	4	7	3	3	0	3
CCVCC	0	1	0	0	0	1	0	0		
CCCVCC	0	0	0	0	0	1	0	0		
CCCVC	2	3	0	1	2	1	0	2		1
CCCCVC	0	1	0	0	0	0	0	0		
N	200	200	200	200	200	200	200	200	94	200

R	0.199	0.326	0.323	0.299	0.386	0.292	0.322	0.360	0.536	0.400
R_{rel}	0.805	0.678	0.681	0.705	0.617	0.711	0.681	0.643	0.469	0.603

It is possible to study metrical positions themselves from the point of view of the peculiarities which syllabic types distributions display in them. For this purpose, both measures, the Zipf-Alekseev formula and the repeat-rate indicator, are used. In the Zipf-Alekseev formula, parameter "a" is interpreted as the feature of the language at large, whereas parameter "b" shows the changes made by the author of the text (Hřebíček, 2002). This is why for the study of the relationship of different metric positions, out of the two parameters the latter was chosen (Table 11). Graphically, this is represented in three scatterplots, in which the horizontal axis represents the values of the repeat rate indicator and the vertical axis – those of the parameter "b" (Fig. 1–3).

Table 11

Repeat-rate indicator and parameter "b" of the Zipf-Alekseev function

Position in line	*Ruslan i Ludmila*		*Graf Nulin*		*Medniy Vsadnik*	
	R_{rel}	b	R_{rel}	b	R_{rel}	b
1	0.79	-0.24	0.80	-0.22	0.81	-0.49
2	0.69	-2.32	0.64	-4.12	0.68	-2.21
3	0.65	-2.18	0.62	-0.66	0.68	-1.42
4	0.73	-1.46	0.69	-2.10	0.70	-2.06
5	0.68	-2.36	0.66	-3.34	0.62	-1.67
6	0.69	-1.75	0.67	-2.09	0.71	-2.29
7	0.68	-1.26	0.65	-0.27	0.68	-0.91
8	0.65	-3.34	0.67	-2.50	0.64	-3.77
9	0.51	-8.10	0.53	-5.78	0.47	-3.22
LF	0.57	-4.15	0.47	-2.93	0.60	-4.93

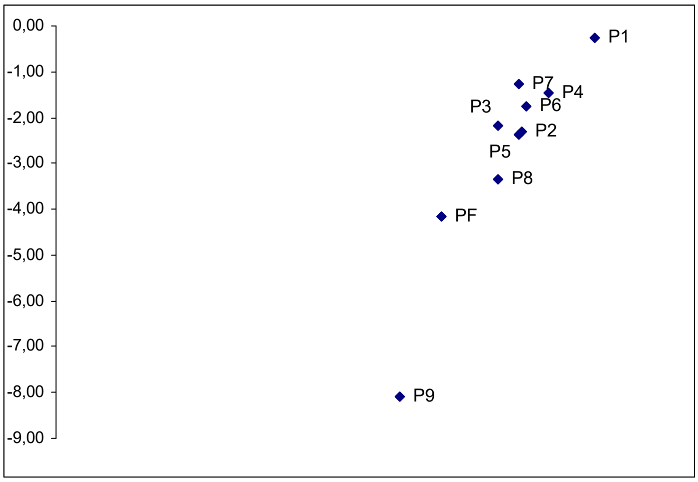

Fig. 1. Scatterplot of the metric positions in *Ruslan i Ludmila*

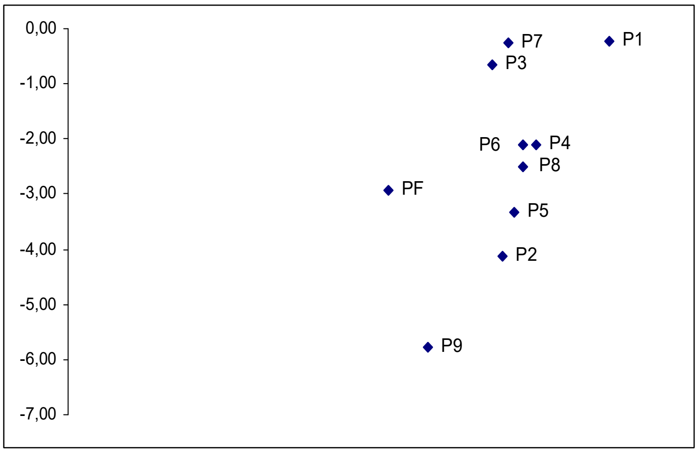

Fig. 2. Scatterplot of the metric positions in *Graf Nulin*

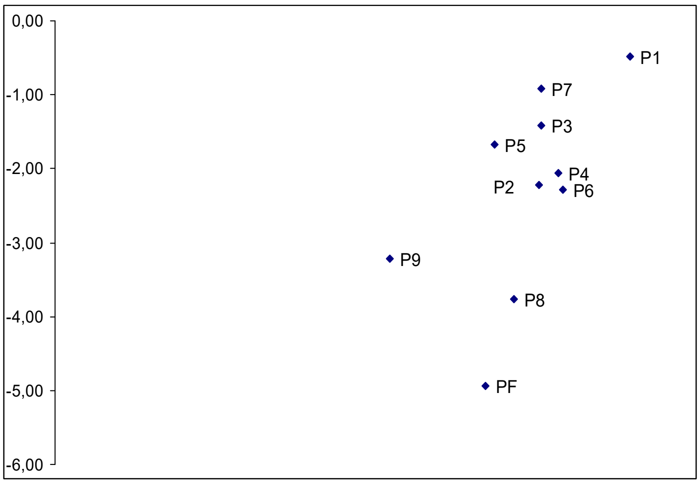

Fig. 3. Scatterplot of the metric positions in *Medniy Vsadnik*

As seen in the scatterplots, the earliest poem (Fig. 1) demonstrates some signs of correlation of two measures, which is less obvious in two other poems.

P1 and P9 (Fig. 1 and Fig. 2) are positioned at a long distance from each other and from the “nucleus”, which consists of strong (stressed) positions P2, P4, P6, and weak (unstressed) positions P3 and P5. On the other hand, positions P8 and PF, which are forming the end of the line, are rather different in their positions in two scatterplots (Fig. 1 and 2) and to some extent in Figure 3, too.

Positions P1 and P7, which precede the first and the last ictuses in the line respectively, demonstrate similar characteristics, as seen in all three diagrams, thus forming a certain “frame” of the poetic line. One more remark refers to the general layout – in the early poem, the scatterplot is more concentrated, in the third scatterplot (the mature creative period), the points are dispersed most of all.

On the whole, the study demonstrated that the vertical distributions of syllables (vertical sequences of syllables) are ordered, and are fitted very well by the Zipf-Alekseev function. The period of creative activity and genre do not influence the distribution of syllabic types in metric positions very much.

The distribution of types of syllables in the first metric position is comparatively less fitted by the above-mentioned function, forming an opposition to most of the other positions, especially ictuses, and together with the 7th metric position creates a sort of a frame in poetic lines. Ictuses (even syllables) form the distribution nucleus of the line, whereas odd syllables are less uniform, especially in their values of repeat-rate.

The presented results are only a first step, and indicate the potential of the utilized approach to uncover syllabic types distribution in verse. Further research may include a broader investigation of long poems by various authors in different languages.

References

Altmann, G., Köhler, R. (2015). *Forms and Degrees of Repetitions in Texts. Detection and Analysis*. Berlin/Munich/Boston: de Gruyter Mouton.

Andreev, S., Místecký, M., Altmann, G. (2018). *Sonnets: Quantitative Inquiries*. *Studies in Quantative Linguistics 29.* Lüdenscheid: RAM-Verlag.

Herfindahl, O. C. (1950). *Concentration in the steel industry*. Diss. New York: Columbia University.

Hřebíček L. (2002). Zipf's Law and Text. *Glottometrics* 3, 27–38.

Zörnig, P. Stachowski, K., Rácová, A., Qu, Y., Místecký, M., Ma, K., Lupea, M., Kelih, E., Gröller, V., Gnatchuk, H., Galieva, A., Andreev, S., Altmann, G. (2019). *Quantitative Insights into Syllabic Structures. Studies in Quantitative Linguistics 30*. Lüdenscheid: RAM-Verlag.

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:
Studies in Quantitative Linguistics 1–30

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.
29. S. Andreev, M. Místecký, G. Altmann, *Sonnets: Quantitative Inquiries.* 2018, 129 pp.
30. P. Zörnig, K. Stachowski, A. Rácová, Y. Qu, M. Místecký, K. Ma, M. Lupea, E. Kelih, V. Gröller, H. Gnatchuk, A. Galieva, S. Andreev, G. Altmann, *Quantitative Insights into Syllabic Structures.* 2019, IV+134 pp.