# Glottometrics 46
# 2019

**RAM-Verlag**

**ISSN 1617-8351**
**e-ISSN 2625-8226**

# Glottometrics

**Indexed in ESCI by Clarivate Analytics and SCOPUS by Elsevier**

**Glottometrics** ist eine unregelmäßig er-scheinende Zeitdchrift (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 **Druck-version** 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



## External Academic Peers for Glottometrics

**Prof. Dr. Haruko Sanada**
Rissho University,Tokyo, Japan (http://www.ris.ac.jp/en/);
Link to Prof. Dr. Sanada:: http://researchmap.jp/read0128740/?lang=english; mailto:hsanada@ris.ac.jp
**Prof. Dr.Thorsten Roelcke**
TU Berlin, Berlin, Germany ( http://www.tu-berlin.de/ )
Link to Prof. Dr.Roelcke: http://www.daf.tu-berlin.de/menue/deutsch_als_fremd_und_fachsprache/mitarbeiter/professoren_und_pds/prof_dr_thorsten_roelcke
mailto:Thosten Roellcke (roelcke@tu-berlin.de)

**Bestellungen** der CD-ROM oder der gedruckten Form sind zu richten an
**Orders** for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de
**Herunterladen/ Downloading:** https://www.ram-verlag.eu/journals-e-journals/glottometrics/


Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 46 (2019), Lüdenscheid: RAM-Verlag, 2019. 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 46 (2019)
**online/ e-version ISSN 2625-8226 (print version ISSN 1617-8351)**

# Contents

*Glottometrics 46, 2019, 1-11*

# Simple Definition of Distances between Texts from Rank–Frequency Distributions. A Case of Ukrainian Long Prose Works by Ivan Franko

*Solomija Buk[1], Andrij Rovenchak (Lviv, Ukraine)*

**Abstract.** We present the analysis of long prose texts using several simple definitions of distance based on rank–frequency distributions. Various types of the Euclidean distance, the Jaccard distance, and the cosine distance are calculated. Our approach is useful for studies of groups of texts, where different definitions of distance show different relations with respect to the shortest and the longest text.

**Keywords***: Ukrainian, Ivan Franko, rank-frequency distribution*

## 1. Introduction

The notion of distance associated primarily with the location of objects in space has left its original geometric cradle long ago and entered many different fields. Modern examples include both related natural sciences, like the geometric measure of distance between states in quantum mechanics (Laba & Tkachuk 2017, Kuzmak 2018), and so called digital humanities, where similar notions are applied in authorship and style studies (Labbé & Labbé 2001; Burrows 2002; Labbé 2007; Cortelazzo et al. 2013; Pavlyshenko 2013a; Pavlyshenko 2013b; Kocher & Savoy 2018), thematic concentration analysis (Chen & Liu 2017), plagiarism identification (Álvarez-Carmona et al. 2018), and some other natural language processing tasks (Suleymanov 2007; Sidorov et al. 2015; Kushnir et al. 2016). In abstract contexts, distance is associated with (dis)similarity, so that larger distances mean less similarity and vice versa.

To be a distance in strict mathematical sense, a function $d$ should satisfy the following conditions (Cortelazzo et al. 2013):

- Non-negativity: $d(A,B) \geq 0$;
- Identity $d(A,B) = 0$ if and only if $A = B$;
- Symmetry: $d(A,B) = d(B,A)$;
- Triangle inequality: $d(A,B) \leq d(A,C) + d(C,B)$.

Three types of distances are calculated in the present paper. The first one is the Euclidean distance given for the vectors **A** and **B** in a $D$-dimensional space by the following expression

$$E(\mathbf{A},\mathbf{B}) = \sqrt{\sum_{i=1}^{D} (A_i - B_i)^2},$$

where $A_i$ and $B_i$ are the vector coordinates.

We will also consider the cosine similarity of vectors **A** and **B**,

[1] Andrij Rovenchak (andrij.rovenchak@gmail.com); Solomija Buk (solomija@gmail.com)

$$C_{\mathrm{sim}}(\mathbf{A},\mathbf{B}) = \frac{\sum_{i=1}^{D} A_i B_i}{\sqrt{\sum_{i=1}^{D} A_i^2}\ \sqrt{\sum_{i=1}^{D} B_i^2}} = \frac{\mathbf{A}\cdot\mathbf{B}}{|\,\mathbf{A}\,||\,\mathbf{B}\,|},$$

where the numerator contains the scalar (dot) product of two vectors and the denominator is the product of vector magnitudes. From the geometrical point of view, this is just cosine of the angle between **A** and **B.** The respective "distance" can be defined as

$$C_{\mathrm{dist}}(\mathbf{A},\mathbf{B}) = 1 - C_{\mathrm{sim}}(\mathbf{A},\mathbf{B}).$$

Note however that this definition contradicts the triangle inequality (Korenius et al. 2007), hence the very term *distance* is given in the quotation marks.

The Jaccard index (Jaccard similarity coefficient), which was originally applied in the comparative studies of floral distribution (Jaccard 1901), is formally defined as the relation between the number of common objects in the set and the total number of objects in the two sets $A$ and $B$, i. e.:

$$J_{\mathrm{sim}}(A,B) = \frac{|\,A\cap B\,|}{|\,A\cup B\,|}.$$

The Jaccard distance is obtained by subtracting the similarity coefficient from unity:

$$J_{\mathrm{dist}}(A,B) = 1 - J_{\mathrm{sim}}(A,B).$$

This is a proper distance measure as it satisfies the triangle inequality (Grygorian & Iacob 2018; Kosub 2019).

Note that other types of distances between texts are also known (cf. Labbé & Labbé 2001; Burrows 2002; Chen & Liu 2017) but they are defined in a more complex way, so we will not use them in our study.

In the present paper, we consider several simple measures to analyze distance between texts of the long prose fiction by Ivan Franko, a famous Ukrainian author, philosopher and public figure from the turn of the 20th century. Quantitative studies of his heritage as a writer include both our recent papers (cf. Kelih et al. 2014; Buk & Rovenchak 2016; Rovenchak & Buk 2018) and other authors (Best & Zinenko 1998; Best & Zinenko 1999; Holovatch and Palchykov 2007; Vasilev & Vasileva 2018). The analyzed texts are as follows, listed in the chronological order and marked by a three-letter abbreviation (according to Buk & Rovenchak 2016):

1. **BC1:** *Boa constrictor* (1st edition: 1878–84);
2. **BSm:** *Boryslav smijetsja* (Boryslav Laughs) (1880–81);
3. **ZBe:** *Zakhar Berkut* (1883);
4. **NSB:** *Ne spytavšy brodu* (Without Asking a Wade) (1885–86);
5. **DDO:** *Dlja domašnjoho ohnyšča* (For the Hearth) (1892);
6. **OSu:** *Osnovy suspil'nosty* (Pillars of Society) (1894–95);
7. **PSt:** *Perekhresni stežky* (The Cross-paths) (1900);
8. **BC2:** *Boa constrictor* (2nd edition: 1905–07);
9. **VSh:** *Velykyj šum* (The Great Noise) (1907);

10. **PD2:** *Petriji j Dovbuščuky* (2nd edition: 1909–12).

The first edition of *Petriji j Dovbuščuky* (1876) was written in a rather specific language with significant Church-Slavonic influences, so it is not included in the analysis.

In Section 2, we illustrate the calculations of distances between two sample texts. Results for ten Franko's texts are presented in Section 3. Brief discussion is given in Section 4.

## 2. Sample illustration

Consider Text1, the stanza of Robert Burn's poem,

> My heart's in the Highlands, my heart is not here,
> My heart's in the Highlands, a-chasing the deer;
> Chasing the wild-deer, and following the roe,
> My heart's in the Highlands, wherever I go.

and Text2, the sentence "I go to Highlands".

The frequency list (ordered by frequency, higher to lower, then alphabetically) of orthographic wordforms corresponding to two texts is shown in Table 1. Note that for simplicity we did not apply any lemmatization in this example.

**Table 1**

Rank–frequency lists of Text1 and Text2.

| Rank | Wordform | Abs.freq. | Text1 | Text2 |
|---|---|---|---|---|
| 1. | the | 6 | 6 | 0 |
| 2. | Highlands | 4 | 3 | 1 |
| 3. | my | 4 | 4 | 0 |
| 4. | heart's | 3 | 3 | 0 |
| 5. | in | 3 | 3 | 0 |
| 6. | go | 2 | 1 | 1 |
| 7. | I | 2 | 1 | 1 |
| 8. | a-chasing | 1 | 1 | 0 |
| 9. | and | 1 | 1 | 0 |
| 10. | chasing | 1 | 1 | 0 |
| 11. | deer | 1 | 1 | 0 |
| 12. | following | 1 | 1 | 0 |
| 13. | heart | 1 | 1 | 0 |
| 14. | here | 1 | 1 | 0 |
| 15. | is | 1 | 1 | 0 |
| 16. | not | 1 | 1 | 0 |
| 17. | roe | 1 | 1 | 0 |
| 18. | to | 1 | 0 | 1 |
| 19. | wherever | 1 | 1 | 0 |
| 20. | wild-deer | 1 | 1 | 0 |

It would yield a 20-dimensional space, where the vector of the first text has the following coordinates:

$$\mathbf{A} = (6; 3; 4; 3; 3; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 0; 1; 1).$$

The second text is represented by the vector

$$\mathbf{B} = (0; 1; 0; 0; 0; 1; 1; 0; 0; 0; 0; 0; 0; 0; 0; 0; 0; 1; 0; 0).$$

The Euclidean distance between two texts is calculated as follows:

$$E(\mathbf{A},\mathbf{B}) = \sqrt{\sum_{j=1}^{20}(A_i - B_i)^2} = \sqrt{(6-0)^2 + (3-1)^2 + \ldots + (0-1)^2 + (1-0)^2 + (1-0)^2} \approx 9.33.$$

The cosine similarity is

$$C_{\mathrm{sim}}(\mathbf{A},\mathbf{B}) = \frac{5}{\sqrt{4}\,\sqrt{93}} \approx 0.259,$$

so that $C_{\mathrm{dist}}(\mathbf{A},\mathbf{B}) \approx 0.741$.

As there are four words in common in Text1 and Text2 and twenty words in total, so the Jaccard similarity and distance are

$$J_{\mathrm{sim}}(\mathbf{A},\mathbf{B}) = \frac{4}{20} = 0.2, \qquad J_{\mathrm{dist}}(\mathbf{A},\mathbf{B}) = 0.8.$$

The value is rather close to unity, which means that the two texts are quite distant.

Note that the Jaccard distance and the cosine distance always remain in the segment [0;1], so to a certain extent they represent a relative measure, while the Euclidean distance might be considered as an absolute measure. This is better illustrated in the next Section.

## 3. Distances between texts of the long prose fiction by Ivan Franko

Before proceeding to the calculation of distances between texts it is worth giving a summary of statistical data for text under study. In Table 2, we list text size (in tokens which are orthographic words). As all the texts were previously lemmatized, the vocabulary size is given as the number of lemma types. The number of hapax legomena is listed for convenience and future references.

**Table 2**

Some statistical data on Franko texts

| | Text size, tokens | Vocabulary size, lemma types | Type-token ratio | Number of hapax legomena | Fraction of hapax legomena in text | Fraction of hapax Legomena in dictionary |
|---|---|---|---|---|---|---|
| BC1 | 25 427 | 5 060 | 0.199 | 2 853 | 0.112 | 0.564 |
| BSm | 77 454 | 8 761 | 0.113 | 4 517 | 0.058 | 0.516 |
| ZBe | 50 220 | 6 688 | 0.133 | 3 411 | 0.068 | 0.510 |
| NSB | 49 170 | 7 342 | 0.149 | 4 014 | 0.082 | 0.547 |
| DDO | 44 841 | 6 564 | 0.146 | 3 417 | 0.076 | 0.521 |
| OSu | 67 172 | 8 561 | 0.127 | 4 492 | 0.067 | 0.525 |
| PSt | 93 890 | 10 080 | 0.107 | 4 960 | 0.053 | 0.492 |
| BC2 | 34 214 | 6 238 | 0.182 | 3 431 | 0.100 | 0.550 |
| VSh | 36 996 | 6 500 | 0.176 | 3 678 | 0.099 | 0.566 |
| PD2 | 52 751 | 7 341 | 0.139 | 3 693 | 0.070 | 0.503 |
| **Total** | **532 135** | **27 807** | **0.052** | **14 224** | **0.027** | **0.512** |

In the present study, we did not unify euphonic variants (*вчора/учора* ‘yesterday’, *сміятися/сміятись* ‘to laugh’, etc.) under one dictionary entry for simplicity. That is why the data from Table 2 might differ from statistical parameters reported for frequency dictionaries of the respective texts.

As one can see from Table 2, each text can be represented as a vector in a 27 807-dimensional space. The coordinates are absolute frequencies of respective lemmas. Obviously, zeros correspond to lemmas not occurring in a specific text. While numbering of coordinates can be arbitrary for all types of distance calculated below (except for the one in Table 4), we suggest adhering to the approach described in Section 2 and numbering the coordinates with the lemma ranks in the total frequency list.

The calculated Euclidean distances are shown in Table 3. As one would expect, the lowest value (marked in boldface and red) corresponds to two editions of *Boa Constrictor* (see Buk 2012 for a detailed quantitative comparison of these editions). The next closest pair is BC2 and VSh. Separation of other pairs is more significant. The leading novel with the most diverse vocabulary is PSt. Not surprisingly, this yields the largest distances to seven other works: 4655 with BC2, 4281 with VSh, 3988 with ZBe, 3908 with DDO, 3866 with PD2, 3569 with NSB. The next largest distance of 3955 is attested between BC1 and BSm. This fact could be explained by the largest text and vocabulary sizes of PSt and BSm.

**Table 3**

The Euclidean distances between Franko texts. The numbers are rounded to integers. The smallest distance is given in boldface and red. The largest distances are in italics and highlighted

| | BC1 | BSm | Zbe | NSB | DDO | Osu | PSt | BC2 | VSh | PD2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1 | 0 | *3955* | 1986 | 2137 | 1800 | 3147 | *5213* | **832** | 1400 | 2220 | BC1 |
| BSm | | 0 | 2703 | 2442 | 2879 | 1989 | 2491 | 3413 | 3166 | 2606 | BSm |
| ZBe | | | 0 | 1486 | 1477 | 2167 | *3988* | 1539 | 1378 | 1543 | Zbe |
| NSB | | | | 0 | 1313 | 1641 | *3569* | 1682 | 1380 | 1528 | NSB |
| DDO | | | | | 0 | 1958 | *3908* | 1400 | 1202 | 1579 | DDO |
| OSu | | | | | | 0 | 2832 | 2613 | 2320 | 2077 | Osu |
| PSt | | | | | | | 0 | *4655* | *4281* | *3866* | PSt |
| BC2 | | | | | | | | 0 | 950 | 1780 | BC2 |
| VSh | | | | | | | | | 0 | 1622 | VSh |
| PD2 | | | | | | | | | | 0 | PD2 |

The number of lemmas found in all ten texts is 956. It would be interesting to calculate the distances between texts based on frequencies of most frequent words, and the number 1000 seems thus to be a good limit. The respective results are shown in Table 4.

Comparing the data in Table 4 with those in Table 3 one can notice that the difference between the values of distances do not differ more than 4% and in many cases remains within 1%. On the other hand, the same effect concerning the largest distance between PSt and other texts could be noticed here as well.

**Table 4**

The Euclidean distances between Franko texts calculated from 1000 most frequent words. The numbers are rounded to integers. The smallest distance is given in boldface and red. The largest distances are in italics and highlighted

| | BC1 | BSm | ZBe | NSB | DDO | OSu | PSt | BC2 | VSh | PD2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1 | 0 | *3941* | 1962 | 2121 | 1781 | 3132 | *5200* | **808** | 1379 | 2198 | BC1 |
| BSm | | 0 | 2674 | 2415 | 2855 | 1950 | 2454 | 3395 | 3144 | 2574 | BSm |
| ZBe | | | 0 | 1447 | 1438 | 2135 | *3966* | 1505 | 1339 | 1500 | Zbe |
| NSB | | | | 0 | 1279 | 1609 | 3548 | 1659 | 1350 | 1488 | NSB |
| DDO | | | | | 0 | 1931 | *3891* | 1371 | 1168 | 1542 | DDO |
| OSu | | | | | | 0 | 2805 | 2593 | 2298 | 2044 | Osu |
| PSt | | | | | | | 0 | *4639* | *4266* | *3844* | PSt |
| BC2 | | | | | | | | 0 | 913 | 1752 | BC2 |
| VSh | | | | | | | | | 0 | 1588 | VSh |
| PD2 | | | | | | | | | | 0 | PD2 |

Next we present results of calculations for the cosine (Table 5) and the Jaccard (Table 6) distances based on data from the full 27 807-dimensional space.

**Table 5**

The cosine distances between Franko texts. The smallest distances are given in boldface. The largest distances are in italics and highlighted.

| | BC1 | BSm | ZBe | NSB | DDO | OSu | PSt | BC2 | VSh | PD2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1 | 0 | 0.080 | *0.120* | *0.113* | *0.125* | *0.118* | *0.123* | **0.046** | *0.129* | *0.120* | BC1 |
| BSm | | 0 | 0.068 | 0.050 | 0.070 | 0.060 | 0.064 | 0.053 | 0.052 | 0.076 | BSm |
| ZBe | | | 0 | 0.089 | 0.099 | 0.099 | 0.102 | 0.082 | 0.075 | 0.093 | Zbe |
| NSB | | | | 0 | 0.068 | **0.047** | 0.057 | 0.080 | 0.055 | 0.089 | NSB |
| DDO | | | | | 0 | 0.056 | 0.058 | 0.088 | 0.070 | 0.097 | DDO |
| OSu | | | | | | 0 | 0.057 | 0.076 | 0.059 | 0.098 | Osu |
| PSt | | | | | | | 0 | 0.080 | **0.047** | *0.109* | PSt |
| BC2 | | | | | | | | 0 | 0.064 | 0.088 | BC2 |
| VSh | | | | | | | | | 0 | 0.085 | VSh |
| PD2 | | | | | | | | | | 0 | PD2 |

The smallest value of the cosine distance is found for the two editions of *Boa Constrictor* as in the case of the Euclidean distance. The difference is however much less pronounced. The next closest value of 0.047 is observed for PSt–VSh (note, in previous Tables 3 and 4 these novels were within the furthest ones) and NSB–OSu pairs. One can see the highest cosine distances between BC1 (the shortest text) and other novels, as well as between PSt and PD2.

In Table 6, the Jaccard distances are shown. Again, the *Boa Constrictor* editions are less distant, $J_{\text{dist}}(\text{BC1},\text{BC2}) < 0.6$, while most values exceed 0.7.

**Table 6**

The Jaccard distances between Franko texts. The smallest distance is given in boldface. The largest distances are in italics and highlighted

| | BC1 | BSm | ZBe | NSB | DDO | OSu | PSt | BC2 | VSh | PD2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1 | 0 | 0.735 | *0.749* | *0.759* | *0.754* | *0.763* | *0.773* | **0.596** | *0.781* | 0.744 | BC1 |
| BSm | | 0 | 0.717 | 0.723 | 0.731 | 0.713 | 0.711 | 0.718 | *0.747* | 0.721 | BSm |
| ZBe | | | 0 | 0.728 | 0.730 | 0.727 | 0.733 | 0.728 | 0.746 | 0.708 | Zbe |
| NSB | | | | 0 | 0.721 | 0.704 | 0.718 | 0.734 | 0.749 | 0.722 | NSB |
| DDO | | | | | 0 | 0.693 | 0.692 | 0.720 | 0.728 | 0.710 | DDO |
| OSu | | | | | | 0 | 0.678 | 0.726 | 0.726 | 0.708 | Osu |
| PSt | | | | | | | 0 | 0.719 | 0.713 | 0.697 | PSt |
| BC2 | | | | | | | | 0 | 0.730 | 0.702 | BC2 |
| VSh | | | | | | | | | 0 | 0.724 | VSh |
| PD2 | | | | | | | | | | 0 | PD2 |

In Table 6, one can see once more higher distances between BC1 and other novels, as well as between VSh and BSm. Remarkable is the fact that the distance between BC1 and BSm is not large enough neither within the Jaccard nor the cosine definitions.

We have also calculated Euclidean distances between texts using relative frequencies as the coordinates instead of absolute frequencies as in Table 3. The results are presented in

Table 7. The behavior of BC1 towards other texts (except BSm) from Tables 5 and 6 is observed here once more.

**Table 7**

The Euclidean distances between Franko texts based on relative frequencies. The smallest distance is given in boldface. The largest distances are in italics and highlighted.

| | BC1 | BSm | ZBe | NSB | DDO | OSu | PSt | BC2 | VSh | PD2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1 | 0 | 0.0289 | *0.0345* | *0.0345* | *0.0358* | *0.0344* | *0.0358* | **0.0216** | *0.0367* | *0.0348* | BC1 |
| BSm | | 0 | 0.0260 | 0.0230 | 0.0268 | 0.0245 | 0.0258 | 0.0232 | 0.0233 | 0.0277 | BSm |
| ZBe | | | 0 | 0.0301 | 0.0309 | 0.0306 | 0.0318 | 0.0278 | 0.0274 | 0.0296 | ZBe |
| NSB | | | | 0 | 0.0266 | 0.0222 | 0.0245 | 0.0287 | 0.0241 | 0.0302 | NSB |
| DDO | | | | | 0 | 0.0235 | 0.0242 | 0.0295 | 0.0269 | 0.0309 | DDO |
| OSu | | | | | | 0 | 0.0240 | 0.0271 | 0.0245 | 0.0307 | OSu |
| PSt | | | | | | | 0 | 0.0283 | 0.0220 | *0.0331* | PSt |
| BC2 | | | | | | | | 0 | 0.0256 | 0.0292 | BC2 |
| VSh | | | | | | | | | 0 | 0.0292 | VSh |
| PD2 | | | | | | | | | | 0 | PD2 |

It might be also interesting to look at the distances calculated using the coordinates defined according to Boolean logic, so that the coordinate is "1" if the respective word occurs in the text at least once and "0" otherwise. The results are shown in Table 8.

**Table 8**

The Euclidean distances between Franko texts based on the Boolean coordinate values. The smallest distance is given in boldface. The largest distances are in italics and highlighted.

| | BC1 | BSm | ZBe | NSB | DDO | OSu | PSt | BC2 | VSh | PD2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1 | 0 | 89.6 | 83.9 | 87.1 | 83.8 | 91.6 | *97.7* | **69.3** | 86.1 | 85.7 | BC1 |
| BSm | | 0 | 93.0 | *95.5* | 94.0 | *97.9* | *102.0* | 91.6 | 95.4 | 95.3 | BSm |
| ZBe | | | 0 | 89.6 | 87.3 | 93.3 | *98.5* | 86.0 | 88.6 | 87.7 | Zbe |
| NSB | | | | 0 | 88.5 | 92.9 | *98.8* | 88.7 | 91.0 | 91.0 | NSB |
| DDO | | | | | 0 | 89.6 | *93.8* | 84.9 | 86.5 | 87.4 | DDO |
| OSu | | | | | | 0 | *97.8* | 91.9 | 92.6 | 93.4 | Osu |
| PSt | | | | | | | 0 | *95.7* | *95.8* | *96.6* | PSt |
| BC2 | | | | | | | | 0 | 85.6 | 85.7 | BC2 |
| VSh | | | | | | | | | 0 | 88.6 | VSh |
| PD2 | | | | | | | | | | 0 | PD2 |

Not surprisingly, the BC1–BC2 pair is closest in this case as well (with distance 69.3). The next closest pair, BC1–DDO, has distance 83.8, which is 21% larger. Comparing to Table 3 one can see that the Euclidean distance of BC2–VSh is just 14% larger than the shortest one for BC1–BC2. The behavior of PSt towards other texts from Tables 3 and 4 is repeating here once more.

So, summarizing the data from Tables 3–8, we conclude that the best discrimination is provided by the simple Euclidean distance (Table 3) and its modification based on Boolean values of coordinates (Table 8). Only a slightly worse estimation is obtained for the Jaccard

distance: the pair OSu–PSt is almost 14% more distant than BC1–BC2. The respective results are visualized in Fig. 1 for the simple Euclidean distance.

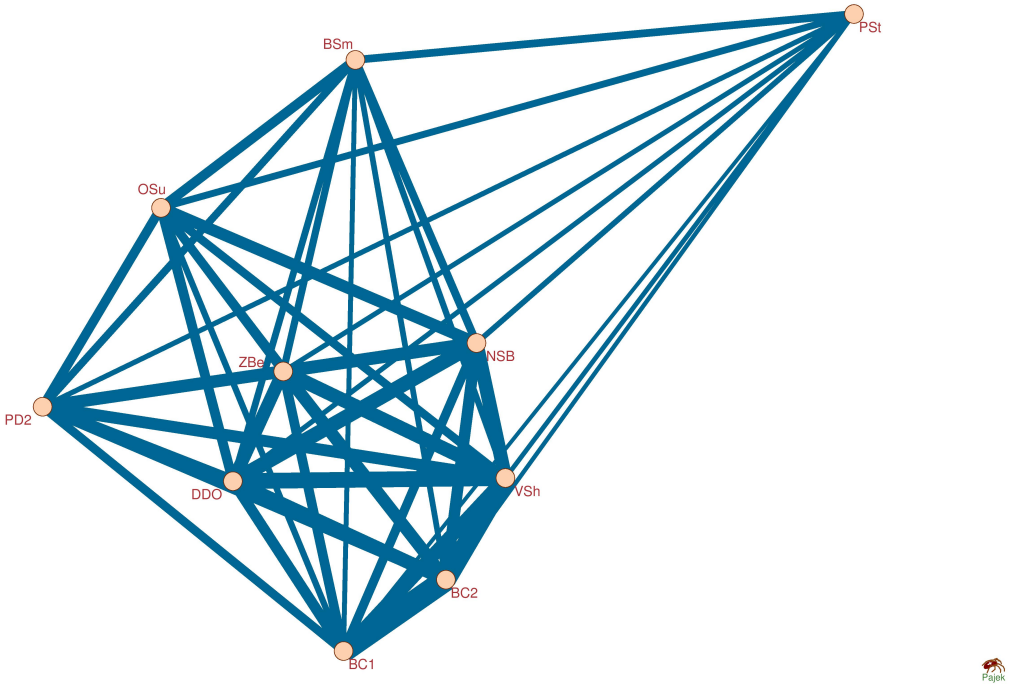


Figure 1. Euclidean distances between long prose texts by Ivan Franko. The data are visualized using the Pajek software (Mrvar & Batagelj 1996–2018). Line width is inverse proportional to the distance between the respective texts.

The analysis of the obtained results shows that patterns of distances with respect to shortest and largest values are of two types. The first one is observed in Tables 3, 4, and – to a certain extent – in Table 8. All those results correspond to various approaches to calculate the Euclidean distance, namely, using absolute frequencies and Boolean values of coordinates. The second pattern is found in Tables 5 (the cosine distance) and 6 (the Euclidean distance based on relative frequencies) as well as slightly different in Table 6 for the Jaccard distance. In the second case, special attention should be paid to the pair BC1–BSm, which falls out from the generally observed behavior of BC1 being mostly distant from other texts. This can indicate, in particular, significant similarities of the vocabulary.

## 4. Discussion

The approaches proposed in this work are mostly suitable for the analysis of groups of texts. The distance between the shortest text and other texts as well as between the longest text and other texts calculated by applying different definitions demonstrate different patterns. The cosine distance and the Euclidean distance based on relative frequencies as well as the Jaccard distance have the largest values in the first case (the shortest text – other texts), while the Euclidean distances based on absolute frequencies and on Boolean values are the largest in the second case (the longest text – other texts).

We also suggest that the distance between texts can be generalized in the following manner. With a list of most frequent words being large enough, preferably several thousand items, one can define the basis of a subspace used to ascribe text coordinates. Alternatively,

the list of the words in the basis can be chosen from other lexicostatistical considerations (cf. Buk 2009; Dellert & Buch 2018; Kwaik et al. 2018).

In prospect, it would be useful to check the discussed approaches using other similar text material, both from one author and different authors. Interpretation of the defined distance values in view of textological and literary studies is another interesting task for future work.

**Acknowledgement.** This work was partially supported (A. R.) by the project FF-83F (registration number 0119U002203) from the Ministry of Education and Science of Ukraine.

## References

**Álvarez-Carmona, M. A., Franco-Salvador, M., Villatoro-Tello, E., Montes-y-Gómez, M., Rosso, P., & Villaseñor-Pineda, L.** (2018). Semantically-informed distance and similarity measures for paraphrase plagiarism identification. *Journal of Intelligent & Fuzzy Systems* 34.5, pp. 2983–2990. DOI: 10.3233/jifs-169483.

**Best K.-H., Zinenko S.** (1998). Wortkomplexität im Ukrainischen und ihre linguistische Bedeutung. In: *Zeitschrift für Slavische Philologie* 58.1, pp. 107–123.

**Best K.-H., Zinenko S.** (1999). Wortlängen in Gedichten des ukrainischen Autors Ivan Franko. In: *Pange lingua. Zborník na počest' Viktora Krupu*. Ed. by J. Genzor, S. Ondrejovič. Bratislava: Veda, pp. 201–213.

**Buk, S.** (2009). Lexical base as a compressed language model of the world (on the material of the Ukrainian language). *Psychology of Language and Communication* 13.2, pp. 35–44.

**Buk, S.** (2012). Quantitative comparison of texts (on the material of the 1884 and 1907 editions of the novel «Boa Constrictor» by Ivan Franko). *Ukrainian Literary Studies* 76, pp. 179–192.

**Buk, S. & Rovenchak, A.** (2016). Probing the 'temperature' approach on Ukrainian texts: Long-prose fiction by Ivan Franko. In: *Studies in Quantitative linguistics 23: Issues in Quantitative Linguistics 4*. Ed. by E. Kelih, R. Knight, J. Mačutek, & A. Wilson. Lüdenscheid: RAM-Verlag, pp. 160–175.

**Burrows, J.** (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. In: *Literary and Linguistic Computing* 17.3, pp. 267–287. DOI: 10.1093/llc/17.3.267.

**Chen, R., & Liu, H.** (2017). Thematic concentration as a discriminating feature of text types: *Journal of Quantitative Linguistics* 25.1, pp. 53–76. DOI: 10.1080/09296174.2017.1339441.

**Cortelazzo, M. A., Nadalutti, P., & Tuzzi, A.** (2013). Improving Labbé's intertextual distance: Testing a revised version on a large corpus of Italian literature *Journal of Quantitative Linguistics* 20.2, pp. 125–152. DOI: 10.1080/09296174.2013.773138.

**Dellert, J. & Buch, A.** (2018). A new approach to concept basicness and stability as a window to the robustness of concept list rankings. *Language Dynamics and Change* 8.2, pp. 157–181. DOI: 10.1163/22105832-00802001.

**Grygorian, A., & Iacob, I. E.** (2018). A concise proof of the triangle inequality for the Jaccard distance. *The College Mathematics Journal* 49.5, pp. 363–365. DOI: 10.1080/07468342.2018.1526020.

**Holovatch, Yu. & Palchykov, V.** (2007). Fox Mykyta and networks of language. *Journal of Physical Studies* 11.1, pp. 22–33.

**Jaccard, P.** (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37, pp. 547–579. DOI:

10.5169/seals-266450

**Kelih, E., Rovenchak, A., & Buk, S.** (2014). Analysing h-point in lemmatised and non-lemmatised texts. In: *Studies in Quantitative Linguistics 17: Empirical Approaches to Text and Language Analysis; dedicated to Luděk Hřebiček on the occasion of his 80th birthday*. Ed. by G. Altmann, R. Čech, J. Mačutek, L. Uhlířová. Lüdenscheid: RAM-Verlag, pp. 81–93.

**Kocher, M., & Savoy, J.** (2018). Evaluation of text representation schemes and distance measures for authorship linking. In: *Digital Scholarship in the Humanities*, in press. DOI: 10.1093/llc/fqy013.

**Korenius, T., Laurikkala, J., & Juhola, M.** (2007). On principal component analysis, cosine and Euclidean measures in information retrieval. *Information Sciences* 177.22, pp. 4893–4905. DOI: 10.1016/j.ins.2007.05.027.

**Kosub, S.** (2019). A note on the triangle inequality for the Jaccard distance. *Pattern Recognition Letters* 120, pp. 36–38. DOI: 10.1016/j.patrec.2018.12.007.

**Kushnir, O., Volosko, A., Ivanitskyi, L., Rykhlyuk, S.** (2016). On the statistics of inter-word distances and the problem of recognition of content words. *Electronics and Information Technologies* 6, pp. 155–164.

**Kuzmak, A. R.** (2018). Geometry of quantum state manifolds generated by the Lie algebra operators. *Journal of Geometry and Physics* 126, pp. 1–6. DOI: 10.1016/j.geomphys.2018.01.007.

**Kwaik. K. A., Saad, M. Chatzikyriakidis, S., & Dobnika, S.** (2018). A lexical distance study of Arabic dialects. *Procedia Computer Science* 142, pp. 2–13. DOI: 10.1016/j.procs.2018.10.456.

**Laba, H. P. & Tkachuk, V. M.** (2017). Geometric characteristics of quantum evolution: curvature and torsion. *Condensed Matter Physics* 20.1, article 13003 (7 pp.). DOI: 10.5488/CMP.20.13003.

**Labbé, D.** (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics* 14.1, pp. 33–80. DOI: 10.1080/09296170600850601.

**Labbé, C., & Labbé, D.** (2001). Inter-textual distance and authorship attribution Corneille and Molière. *Journal of Quantitative Linguistics* 8.3, pp. 213–231. DOI: 10.1076/jqul.8.3.213.4100.

**Mrvar, A. & Batagelj, V.** (1996–2018). *Pajek: analysis and visualization of large networks.* http://mrvar.fdv.uni-lj.si/pajek/.

**Pavlyshenko, B.** (2013a). Classification analysis of authorship fiction texts in the space of semantic fields. *Journal of Quantitative Linguistics* 20.3, pp. 218–226. DOI: 10.1080/09296174.2013.799914

**Pavlyshenko, O.** (2013b). The lexical-semantic fields of verbs in English texts. *Glottometrics* 25, pp. 69–84.

**Rovenchak, A. & Buk, S.** (2018). Part-of-speech sequences in literary text: Evidence from Ukrainian. In: *Journal of Quantitative Linguistics* 25.1, pp. 1–21. DOI: 10.1080/09296174.2017.1324601.

**Sidorov, G., Gomez-Adorno, H., Markov, I., Pinto, D., & Loya, N.** (2015). Computing text similarity using Tree Edit Distance.. In: *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*. DOI: 10.1109/nafips-wconsc.2015.7284129.

**Suleymanov, A. Sh.** (2007). Semantic proximity and semantic interval between texts. In: *Adaptyvni systemy avtomatyčnoho upravlinnja* 10(30), pp. 132–136.

**Vasilev, A. & Vasileva, I.** (2018). Text length and vocabulary size: Case of the Ukrainian writer Ivan Franko. *Glottometrics* 43, pp. 1–10.

*Glottometrics 46, 2019, 12-28*

# The Romantic Clash: Influence of Karel Sabina over Mácha's *Cikáni* from the Perspective of the Numerals Usage Statistics

*Andrei V. Zenkov[1,2], Michal Místecký[3]*

**Abstract.** The paper deals with the long-standing stylometric problem of Czech fiction – the authorship of the novel *Cikáni*. Although the text has been usually attributed to K. H. Mácha, there is a widespread hypothesis that its final shape was substantially influenced by a friend of his, K. Sabina. To solve the problem, we have exposed the works by Mácha (*Cikáni*) and Sabina (*Hrobník* and *Oživené hroby*) to the novel statistical attribution method, which takes into account the usage of numerals in texts. To provide a contrast to the new procedure, we have also employed a more conventional MFW analysis. The results, which are rather contradictory, are accounted for by various interpretations. The goal of the article is to show the soundness of the new method and check its applicability on Czech pieces of literature.

***Keywords:*** *Karel Hynek Mácha; Karel Sabina; Cikáni; Benford's Law; first significant digit; authorship attribution; stylometry; MFW analysis*

## 1. Introduction

Over the recent years, a lot of different authorship-attribution methods have appeared in the domain of quantitative linguistics (cf. Juola 2006; Eder, Rybicki, Kestermont 2016). In the present paper, we are going to make use of the one founded upon the usage of numerals in the text (Zenkov 2018). The employment of them seems to be typical of a given author's style.

The texts which have been selected for analysis are *Cikáni* ("The Gypsies"), the only novel supposedly written by Karel Hynek Mácha (1810–1836), otherwise a prominent Czech poet and legendary figure of the Romantic movement, and *Hrobník* ("A Grave-Digger") as well as *Oživené hroby* ("Graves Brought to Life"), two pieces of fiction by Karel Sabina (1813–1877), his friend and author of prose and opera librettos. Written in 1835, but published as late as 1857 by Sabina, the novel of *Cikáni* was not conserved in the manuscript version, the same being the case for several minor poems by Mácha. This has led some scholars to the conclusion that these works had actually been written by Sabina, who wanted to create the cult of the Romantic rebellious Mácha (cf. Králík 1969; Stich 1976; Charypar 2004; Zlamalová 2010). This discussion resonated in the Czech literary scholarship in the 1960, but seems to have been closed in the mid-1970s, as the proposed hypothesis was discredited by many speculations (e.g., Sabina's interferences in Mácha's works were supported by little evidence, except for the discrete counts of language phenomena in the

[1] Ural Federal University, Ekaterinburg, Russia; e-mail: zenkow@mail.ru.
[2] This work was partially supported by a scholarship from the Slovak Academic Information Agency and by the Russian Foundation for Basic Research, under Grant No. 19-012-00199A.
[3] University of Ostrava, Ostrava, the Czech Republic; e-mail: mmistecky@seznam.cz.

given texts). However, at least in case of one letter, Charypar (2004) considers Sabina's interpolation in the text possible, even plausible, as there is a peculiar trait untypical of Mácha's production – the usage of extensive literary allusions. It is to be noted that this feature is to be found in *Cikáni*, too, where each chapter is introduced with a long quote from Polish literature.

Given the aforementioned, Mácha's and Sabina's texts offer themselves as a good testing material for the numerals method of authorship attribution. The article will be organized as follows: first, the principles and the procedures of the attribution will be explained; second, the novel numerals statistics method as well as the more traditional MFW-based attribution procedure will be applied to the texts; third, interpretations of the obtained results will be provided; and, last but not least, conclusions will be drawn from the investigations. The goal of the paper is both to test the utility of the methods, and provide a new vista for a still-open literary-scholarship issue.

## 2. Methods

First, we have exposed the texts – of Mácha's *Cikáni*, as well as Sabina's *Hrobník* and *Oživené hroby* – to the novel method of text attribution based on the statistics of numerals occurring in them (Zenkov 2018).

The starting point of our study was Benford's Law (Benford 1938), which refers to the probability of occurrence of the first significant (leftmost nonzero) digit in the distributions of various real-life data. The first significant digits of seemingly random numbers often fail to follow a flat distribution – 1's, 2's, 3's, etc., thus do not occur with the same frequencies, as one would expect, but instead, they obey a decreasing distribution, with more 1's than 2's, more 2's than 3's, and so on. According to Benford's Law, in the decimal system, the probability of the occurrence of digit *d* as the first significant is

$$P(d) = \lg\left(1 + \frac{1}{d}\right). \quad (1)$$

All the probabilities are summed up in the upcoming table.

**Table 1**
Probabilities of the digit occurrences in a data file according to Benford's Law

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| 0.301 | 0.176 | 0.125 | 0.097 | 0.079 | 0.067 | 0.058 | 0.051 | 0.046 |

A classic experiment of Benford's, which shows a good agreement with the law, is the analysis of the occurrence of numerals contained in articles of a randomly selected issue of a magazine. Note that Benford himself analyzed the occurrence of numerals expressed only via *figures*.

A conclusive explanation of Benford's Law, covering all cases of its manifestation, is still absent, although some conditions favouring its emergence are stated. The incomplete understanding does not preclude the successful use of Benford's Law in detecting frauds in accounting data and in elections; the suggested applications range from physics and astronomy to steganography and scientometrics.

We have shown the efficacy of counting frequencies of different first significant digits of numerals for text attribution (Zenkov 2018). It has been found out that not only for the

**random** combination of heterogeneous texts, but also for the **coherent** Russian- and English-language texts, the frequency distributions resemble that of Benford's Law, with the exception of the quota of digit 1, which considerably exceeds 0.3 – at least since the word 'one', formally being a numeral, can actually play the role of the indefinite article. The frequent tendency of rounding numbers is also of importance.

In contrast to the traditional methodology of application of Benford's Law, which treats deviations from the law as an indication of a possible existence of 'falsification' (broadly defined), we laid emphasis on the **comparison** of Benford-like distributions for texts by different authors, showing that the **differences** between these distributions are statistically robust style features that make it possible to distinguish between works by different authors (the texts should be sufficiently large, about 200 kB or more).

Alongside with this novel approach, a conventional most-frequent-words (MFW) analysis has been employed in the research, too, so as to provide a counterpart to the core investigation. To this end, the STYLO package of the R software was made use of, with the three texts being divided into chapters (e.g., Hrobník_3, Hroby_5, or Cikáni_7) and a dendrogram produced on the basis of the distances among them. The method compares two texts, working with the lists of their 100 most frequent words (MFW), the ranks of which it compares using various metrical tools. In our research, we use the Classic Delta distance (Burrows 2002; Argamon 2008), which yields good results in comparable cases (cf. Eder 2013). Its formula reads –

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - B_i}{\sigma_i}\right| , \qquad (2)$$

with $n$ standing for the total of the most frequent words, $A_i$ for the frequency of the word $i$ in the text $A$, $B_i$ for the frequency of the word $i$ in the text $B$, and $\sigma_i$ for the standard deviation of the frequencies of the given word $i$.

The idea behind the approach is that an author's style is based on his/her usage of autosemantic words, which are the most frequent in texts and cannot be controlled by a conscious brain activity.

## 3. Results

Following the first idea, we present here the research results concerning the distributions of the first significant digits of numerals contained in Czech-language texts by Mácha and Sabina (see Figure 1).

We have studied the frequencies of occurrences of various first significant digits of numerals, taking into account cardinal as well as ordinal numerals expressed both with figures, and (considerably more often) verbally. In the latter case, the first step was to rewrite every numeral with figures and then to pick out the first significant digit. To identify the author's use of numerals, we previously cleared the text from idiomatic expressions and set phrases containing numerals **accidentally only** (like 'one hand washes the other', 'five-o'clock' in English), as well as itemizations, such as 1), 2), 3), etc.

Some of our results, which were confirmed on the basis of statistical tests, are presented here. As usual, digit 1 is occurring far more often than predicted by Benford's Law. The statistical characteristics concerning the occurrence of first significant digits of numerals contained in texts by Mácha, on the one hand, and Sabina, on the other hand, are different – both visually and according to Pearson's chi-squared test: the frequency distribution of first significant digits for Mácha's *Cikáni* differs significantly from Sabina's texts at the

significance level of $\alpha = 0.05$.[4] Therefore, the editing work done by Sabina (if any) seems to be negligible.

Moreover, Sabina's style is characterized by a very stable usage of the first significant digits, although the texts analysed are different both in size and the time of creation. This means that the method is credible, as there are no doubts as to Sabina's authorship of *Hrobník* and *Oživené hroby*.

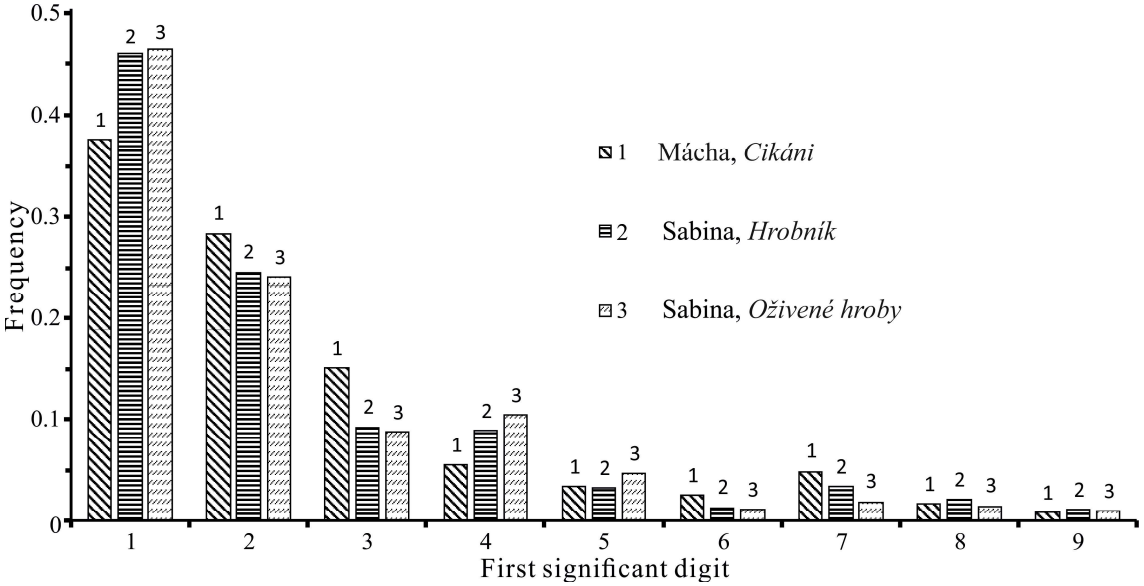


Figure 1. Distribution of the first significant digits of numerals occurring in the texts by Mácha and Sabina

On the contrary, the results of the MFW analysis have shown a different picture. The expected result would have been that the *Hrobník* and *Oživené hroby* chapters would group together and the parts of *Cikáni* would take a branch of their own. However, the three analyses that have been performed – the analysis of the first hundred words, of the two hundred words, and of the first one hundred bigrams[5] – have all grouped *Hrobník* with *Cikáni*, leaving *Oživené hroby* in a separate branch (see Figures 2, 3, and 4).[6] The explanations for the results are multiple; first and foremost, it needs to be said that no morphological analysis of the Czech words has been carried out, as the basic STYLO package does not contain a tool for this operation. This, in itself, substantially lowers the credibility of the MFW analysis research.

Second, what may have played a role, is a change of Sabina's style of writing. *Hrobník* is a 1837 short-story, whereas *Oživené hroby* is a 1870 novel, a piece of a mature and experienced author. Moreover, the fact that Mácha's version of *Cikáni* was written in 1835 means that it belongs to the same period as *Hrobník*. On the other hand, it is interesting that the fact that both *Oživené hroby*, and *Cikáni* are novels does not appear to have influenced the results of the research.

[4] For the counts data and their statistical processing, see Appendix I.

[5] A bigram is a set of two subsequent written characters, spaces and punctuation included. For instance, a sentence "I'm home." contains the following bigrams: /i'/ – /'m/ – /m_/ – /_h/ – /ho/ – /om/ – /me/ – /e./. As bigrams are completely independent of the meanings of words, they are used to eliminate the influence of the treated topics over authorial styles.

[6] As to the distance counts, they are listed in tables in Appendix II.

Unfortunately, it is a usual situation in stylometry that different techniques yield contradictory findings. All the same, the fact that the MFW analysis puts two texts by Sabina to *different* clusters does not testify in its favour.

To conclude, the numerals analysis seems to have provided more trustworthy outcomes at the present moment, though the results obtained via other means should not be left out of scope either.

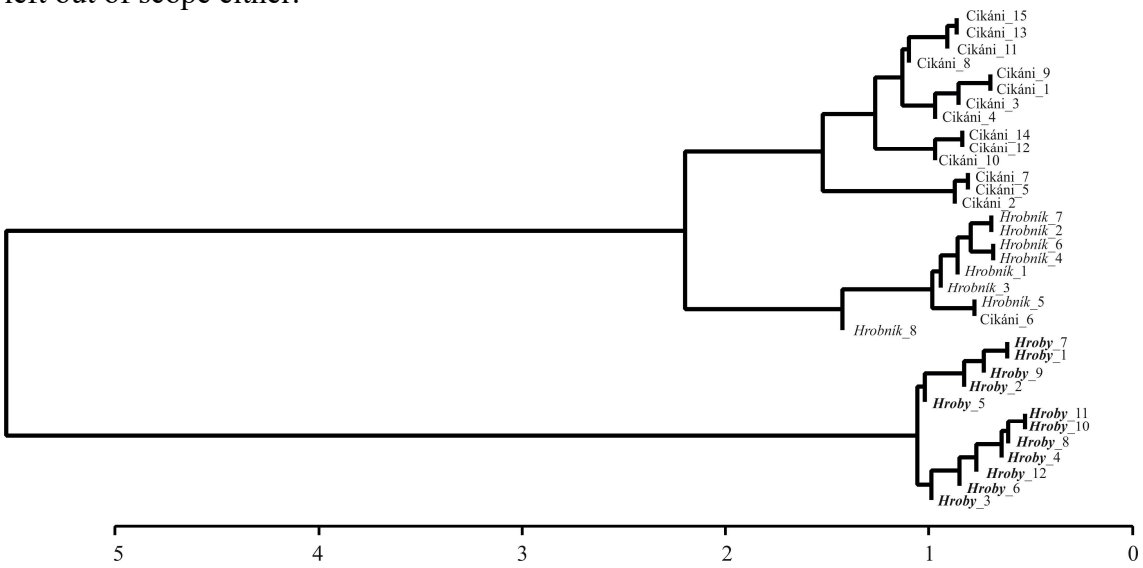


Figure 2. The MFW analysis of the 100 most frequent words of the studied texts

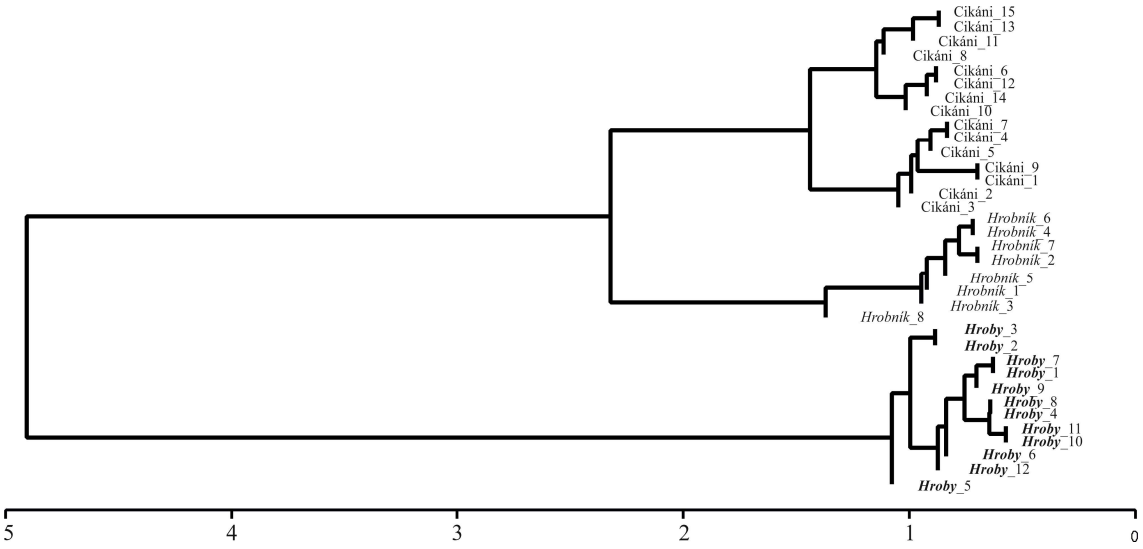


Figure 3. The MFW analysis of the 200 most frequent words of the studied texts.

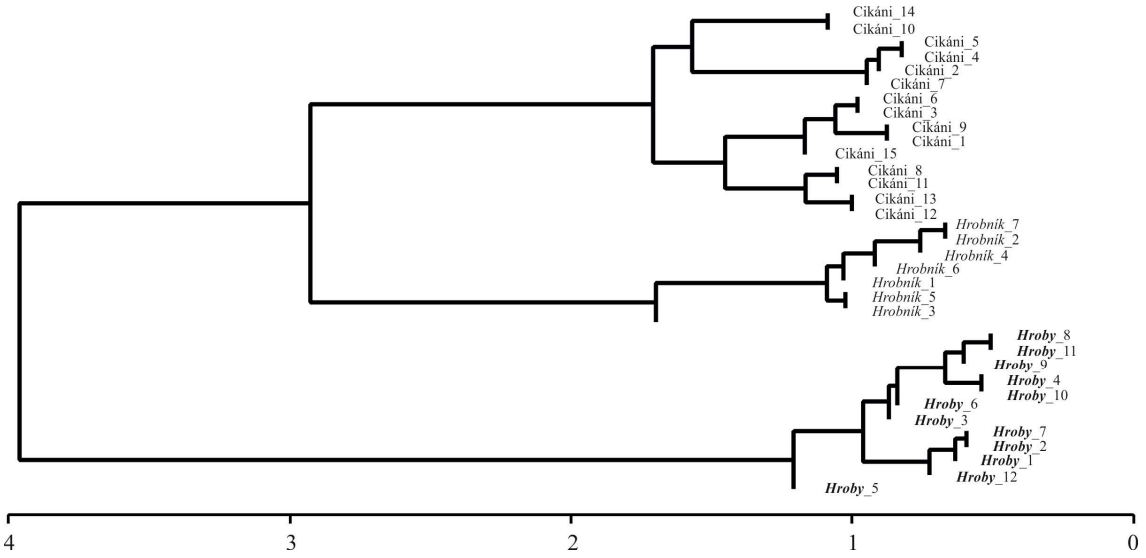


Figure 4. The analysis of the 100 most frequent character bigrams of the studied texts.

## 4. Conclusions

We have shown the applicability of the novel stylometric technique based on the study of the numerals statistics to Czech-language literary texts. We believe that it can be a useful addition to the traditional textual practices of taking into account the length of sentences, the length of words, the frequencies of use of auxiliary words and certain significant parts of speech, etc. – especially as results yielded by them are often controversial. The novel numerals statistics technique seems linguistically more meaningful.

The new approach was confronted against the traditional MFW analysis, which has yielded different results, grouping Sabina's *Hrobník* with Mácha's *Cikáni*. Even though these should not be neglected, it is essential to point out that the conditions of the analysis were not ideal, and the outputs thus cannot be taken as the refutation of the numerals method outcomes. For the time being, we may therefore conclude that the novel of *Cikáni* is part of Mácha's literary heritage.

## Acknowledgement

Andrei Zenkov would like to thank Professor Radko Mesiar, head of the Dept. of Mathematics and Descriptive Geometry, the Slovak Technical University, for his hospitality during the author's stay in Bratislava.

## References

**Argamon, S.** (2008). Interpreting Burrows' delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, 29(2), 147–163.

**Benford, F.** (1938). The law of anomalous numbers. *Proceedings of American Philosophical Society*, 78, 551–572.

**Burrows, J.** (2002). Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), 267–287.

**Charypar, M.** (2004). Dvakrát k problematice autorství u Karla Sabiny. *Slovo a smysl*, 1(1), no pages. Available at: <http://slovoasmysl.ff.cuni.cz/node/15>.

**Eder, M. – Rybicki, J. – Kestermont, M.** (2016). Stylometry with R: a package for computational text analysis. *R Journal*, 8(1), 107–121.

**Eder, M.** (2005). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2), 167–182.

**Juola, P.** (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233–334.

**Králík, O.** (1969). *Demystifikovat Máchu*. Ostrava: Profil.

**Stich, A.** (1976). *Sabina–Němcová–Havlíček (Textologický a stylistický příspěvek k sporům o Sabinových zásazích do cizího díla). Stylistické studie III*. Praha: ÚJČ.

**Zenkov, A. V.** (2018). A Method of Text Attribution Based on the Statistics of Numerals. *Journal of Quantitative Linguistics*, 25(3), 256–270.

**Zlamalová, E.** Mácha jako autorský mýtus.
Available at: <http://www.literarni.cz/rubriky/aktualni/clanky/macha-jako-autorsky-mytus_8025.html#.XFQy8_ZFwkE>.

## APPENDIX I

The comparison of empirical distributions (in our case – the distributions of the first significant digits of numerals occurring in texts by certain authors) is related to the validation of statistical hypotheses about the significance/insignificance of differences between the distributions.

We now formulate hypotheses. The null hypothesis $H_0$ asserts that the tested populations are distributed identically. The alternative hypothesis $H_1$ reads that one population is distributed differently from the other one.

To check the difference between the distributions of the first significant digits, we apply Pearson's chi-squared test, which, among its other applications, is used as a test of homogeneity – it compares the distribution of counts for two groups using the same categorical variable.

Our initial statistical data concerning the occurrence of the different first significant digits in the texts by Mácha and Sabina texts are given in Table 2.

**Table 2**
Empirical results of the texts processing

.

| first significant digit | Counts | | |
|---|---|---|---|
| | *Cikáni* by Mácha | *Hrobník* by Sabina | *Oživené hroby* by Sabina |
| 1 | 73 | 23 | 328 |
| 2 | 55 | 12 | 170 |
| 3 | 29 | 4 | 62 |
| 4 | 10 | 3 | 75 |
| 5 | 7 | 2 | 33 |
| 6 | 5 | 1 | 8 |
| 7 | 9 | 2 | 13 |
| 8 | 3 | 1 | 11 |
| 9 | 2 | 1 | 7 |
| Counts – total | **193** | **49** | **707** |

Next, we show the procedure of hypotheses checking for the texts of *Cikáni* and *Oživené hroby*.

The condition for the applicability of Pearson's chi-squared test is the restriction that the frequency[7] for each cell of the table is not less than 5; we thus merge the cells for digits 6 to 9 (see Table 3).

**Table 3**
Empirical results after the cells merging.

| first significant digit | *Cikáni* | | | *Oživené hroby* | | | total frequency |
|---|---|---|---|---|---|---|---|
| | frequency | relative frequency, % | cell label | frequency | relative frequency, % | cell label | |
| 1 | 73 | 37.6 | I | 328 | 46.4 | II | 73+ 328 = **401** |
| 2 | 55 | 28.4 | III | 170 | 24.1 | IV | 55+ 170 = **225** |
| 3 | 29 | 15.1 | V | 62 | 8.7 | VI | 29 + 62 = **91** |
| 4 | 10 | 5.4 | VII | 75 | 10.6 | VIII | 10 + 75 = **85** |
| 5 | 7 | 3.4 | IX | 33 | 4.7 | X | 7 + 33 = **40** |
| 6 – 9 | 19 | 10.1 | XI | 39 | 5.5 | XII | 19 + 39 = **58** |
| | **Σ = 193** | **Σ = 100%** | | **Σ = 707** | **Σ = 100%** | | **ΣΣ = 900** |

We now juxtapose empirical and theoretical frequencies, the latter obtained by taking into account that the amount of numerals in the texts is different: 193 in *Cikáni*, and 707 in *Oživené hroby*. Thus, of the total quantity is 193 + 707 = 900 numerals in the two texts, the first one accounts for the share 193/900 = 0.21, and the second for 707/900 = 0.79. In all rows, the theoretical frequencies related to the first and second texts should be, respectively, 0.21 and 0.79 of the total frequency of the corresponding row. If the empirical distributions to be compared do not differ from one another, the empirical frequencies should not significantly deviate from the theoretical ones obtained from the proportion.

Now, we recompose the data of Table 3, placing the relative frequencies for both texts in the order indicated by the labels in one column (these will be the *empirical* frequencies $f_{\text{emp}}$), and in the other column, we will place the *theoretical* frequencies $f_{\text{theor}}$, calculated according to the previous as follows:

$$f_{\text{theor}} = \frac{(\Sigma \text{ frequencies over the row}) \cdot (\Sigma \text{ frequencies over the column})}{\Sigma\Sigma}.$$

Here, $\Sigma\Sigma = 900$.

[7] This refers to the theoretical frequency; as can be seen from Table 4, our manipulations with empirical frequencies meet this requirement.

**Table 4**
Calculations for the Pearson's chi-squared test.

| cell | empirical frequency $f_{\text{emp}}$ | theoretical frequency $f_{\text{theor}}$ | $\left(f_{\text{emp}} - f_{\text{theor}}\right)^2 / f_{\text{theor}}$ |
|---|---|---|---|
| I | 73 | $\frac{401 \cdot 193}{900} = 85.99$ | 1.96 |
| II | 328 | $\frac{401 \cdot 707}{900} = 315.01$ | 0.54 |
| III | 55 | $\frac{225 \cdot 193}{900} = 48.25$ | 0.94 |
| IV | 170 | $\frac{225 \cdot 707}{900} = 176.75$ | 0.26 |
| V | 29 | 19.51 | 4.62 |
| VI | 62 | 71.49 | 1.26 |
| VII | 10 | 18.23 | 3.72 |
| VIII | 75 | 66.77 | 1.01 |
| IX | 7 | 8.58 | 0.29 |
| X | 33 | 31.42 | 0.08 |
| XI | 19 | 12.44 | 3.46 |
| XII | 39 | 45.56 | 0.94 |
| | **Σ = 900** | **Σ = 900** | **Σ = 19.08** = $\chi^2_{\text{emp}}$ |

Now, we determine the number of degrees of freedom, *df*. For the test of homogeneity, $df = (r-1)(c-1)$, where *r* corresponds to the number of categories (i.e., rows in the table of empirical frequencies after the cells merging – see Table 3), and *c* corresponds the number of independent groups (i.e., columns in the table). Here, $r = 6$, $c = 2$; so, $df = (6-1)(2-1) = 5$.

With such *df*, the table critical values of the $\chi^2$ distribution for two significance levels α are –

$$\chi^2_{\text{cr}} = \begin{cases} 11.07 & (\alpha = 0.05), \\ 15.09 & (\alpha = 0.01). \end{cases}$$

Since the empirical $\boldsymbol{\chi^2_{\text{emp}}}$= 19.08 exceeds each of these critical values, the hypothesis $H_0$ is rejected even at the significance level of $\alpha = 0.01$; the distributions for *Cikáni* and *Oživené hroby* significantly differ from one another.

The same procedure for the pair *Cikáni–Hrobník* leads to the rejection of the hypothesis $H_0$ as well, and for the pair *Oživené hroby–Hrobník* – to its acceptance (i.e., in the last case, the difference is negligible).

## APPENDIX II

**On typographical gounds, we abbreviated the titles as follows:**
**Ci = Cikáni**
**Hy = Hroby**
**Hk = Hrobník**

**Table 5**
The Delta distances between the samples (MFW analysis of 100 words; part I)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Cikáni_10 | 1.091 | | | | | | | |
| Cikáni_11 | 0.967 | 1.079 | | | | | | |
| Cikáni_12 | 0.938 | 0.990 | 1.073 | | | | | |
| Cikáni_13 | 0.949 | 0.995 | 0.914 | 0.926 | | | | |
| Cikáni_14 | 1.033 | 0.914 | 0.927 | 0.851 | 0.908 | | | |
| Cikáni_15 | 0.915 | 1.109 | 0.914 | 0.985 | 0.878 | 1.062 | | |
| Cikáni_2 | 0.892 | 1.226 | 0.986 | 1.102 | 1.221 | 1.185 | 1.215 | |
| Cikáni_3 | 0.836 | 1.050 | 1.016 | 0.925 | 0.959 | 1.087 | 1.049 | 1.152 |
| Cikáni_4 | 0.879 | 1.048 | 0.987 | 0.921 | 1.035 | 1.091 | 1.011 | 1.006 |
| Cikáni_5 | 0.920 | 1.033 | 1.108 | 1.017 | 1.106 | 1.062 | 1.193 | 0.915 |
| Cikáni_6 | 0.803 | 1.075 | 0.935 | 0.938 | 0.846 | 0.983 | 0.886 | 1.144 |
| Cikáni_7 | 0.957 | 1.089 | 1.158 | 0.975 | 1.108 | 1.167 | 1.145 | 0.828 |
| Cikáni_8 | 1.004 | 1.147 | 1.039 | 1.058 | 1.082 | 1.020 | 1.004 | 1.022 |
| Cikáni_9 | 0.713 | 1.045 | 0.846 | 0.880 | 0.845 | 0.924 | 0.809 | 0.817 |
| Hrobník_1 | 0.889 | 1.027 | 0.925 | 1.084 | 1.096 | 1.264 | 1.078 | 1.108 |
| Hrobník_2 | 0.943 | 0.845 | 1.005 | 0.972 | 1.024 | 1.008 | 1.042 | 1.194 |
| Hrobník_3 | 1.060 | 1.112 | 1.121 | 1.206 | 1.139 | 1.247 | 1.280 | 1.252 |
| Hrobník_4 | 0.858 | 1.016 | 0.919 | 1.082 | 0.883 | 1.016 | 0.932 | 1.110 |
| Hrobník_5 | 1.001 | 1.080 | 1.031 | 1.158 | 1.064 | 1.169 | 0.993 | 1.242 |
| Hrobník_6 | 0.943 | 1.080 | 1.047 | 1.036 | 1.004 | 1.090 | 1.001 | 1.119 |
| Hrobník_7 | 0.893 | 0.968 | 1.053 | 0.985 | 1.041 | 1.071 | 1.137 | 1.129 |
| Hrobník_8 | 1.257 | 1.348 | 1.290 | 1.364 | 1.378 | 1.440 | 1.299 | 1.427 |
| Hroby_1 | 1.014 | 1.155 | 1.143 | 1.186 | 1.217 | 1.419 | 1.320 | 1.194 |
| Hroby_10 | 1.140 | 1.083 | 1.186 | 1.214 | 1.294 | 1.275 | 1.416 | 1.139 |
| Hroby_11 | 1.019 | 1.029 | 1.082 | 1.169 | 1.173 | 1.172 | 1.304 | 1.042 |
| Hroby_12 | 1.252 | 1.258 | 1.263 | 1.343 | 1.314 | 1.370 | 1.430 | 1.327 |
| Hroby_2 | 1.115 | 1.327 | 1.152 | 1.420 | 1.278 | 1.490 | 1.391 | 1.214 |
| Hroby_3 | 1.288 | 1.322 | 1.260 | 1.493 | 1.460 | 1.488 | 1.531 | 1.322 |
| Hroby_4 | 1.022 | 1.063 | 1.070 | 1.126 | 1.200 | 1.215 | 1.302 | 1.088 |
| Hroby_5 | 1.132 | 1.253 | 1.286 | 1.336 | 1.310 | 1.507 | 1.511 | 1.288 |
| Hroby_6 | 1.096 | 1.051 | 1.098 | 1.190 | 1.219 | 1.167 | 1.332 | 1.176 |
| Hroby_7 | 1.092 | 1.233 | 1.206 | 1.320 | 1.302 | 1.389 | 1.358 | 1.190 |
| Hroby_8 | 0.979 | 1.138 | 1.047 | 1.168 | 1.219 | 1.282 | 1.341 | 1.009 |
| Hroby_9 | 1.209 | 1.256 | 1.345 | 1.372 | 1.508 | 1.470 | 1.580 | 1.189 |
| | Ci_1 | Ci_10 | Ci_11 | Ci_12 | Ci_13 | Ci_14 | Ci_15 | Ci_2 |

**Table 6**
The Delta distances between the samples (MFW analysis of 100 words; part II)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Cikáni_4 | 0.997 | | | | | | | |
| Cikáni_5 | 1.010 | 0.950 | | | | | | |
| Cikáni_6 | 0.973 | 0.978 | 1.043 | | | | | |
| Cikáni_7 | 1.100 | 0.878 | 0.823 | 0.962 | | | | |
| Cikáni_8 | 1.126 | 1.060 | 1.172 | 1.036 | 1.057 | | | |
| Cikáni_9 | 0.822 | 0.885 | 0.903 | 0.913 | 0.844 | 0.958 | | |
| Hrobník_1 | 1.031 | 1.083 | 1.104 | 0.958 | 1.076 | 1.094 | 0.992 | |
| Hrobník_2 | 0.998 | 1.010 | 0.873 | 0.798 | 1.063 | 1.055 | 1.074 | 0.849 |
| Hrobník_3 | 1.064 | 1.123 | 0.992 | 0.967 | 1.210 | 1.184 | 1.178 | 0.929 |
| Hrobník_4 | 0.985 | 1.011 | 1.102 | 0.803 | 1.007 | 0.937 | 0.950 | 0.804 |
| Hrobník_5 | 1.156 | 1.061 | 1.112 | 0.789 | 1.050 | 1.137 | 1.040 | 0.948 |
| Hrobník_6 | 0.968 | 0.999 | 1.066 | 0.905 | 1.021 | 1.010 | 0.888 | 0.887 |
| Hrobník_7 | 0.860 | 1.034 | 0.938 | 0.843 | 1.106 | 1.089 | 0.993 | 0.746 |
| Hrobník_8 | 1.282 | 1.305 | 1.378 | 1.187 | 1.215 | 1.158 | 1.307 | 1.149 |
| Hroby_1 | 1.088 | 1.215 | 1.182 | 1.216 | 1.140 | 1.209 | 1.120 | 0.918 |
| Hroby_10 | 1.154 | 1.230 | 1.135 | 1.278 | 1.229 | 1.238 | 1.172 | 1.042 |
| Hroby_11 | 1.059 | 1.154 | 0.993 | 1.153 | 1.050 | 1.167 | 1.014 | 0.983 |
| Hroby_12 | 1.325 | 1.271 | 1.303 | 1.347 | 1.317 | 1.333 | 1.208 | 1.127 |
| Hroby_2 | 1.137 | 1.295 | 1.218 | 1.248 | 1.311 | 1.330 | 1.271 | 1.169 |
| Hroby_3 | 1.227 | 1.346 | 1.402 | 1.474 | 1.398 | 1.317 | 1.281 | 1.341 |
| Hroby_4 | 1.005 | 1.082 | 1.110 | 1.150 | 1.109 | 1.118 | 1.053 | 0.976 |
| Hroby_5 | 1.130 | 1.181 | 1.173 | 1.262 | 1.293 | 1.371 | 1.209 | 1.119 |
| Hroby_6 | 1.077 | 1.132 | 1.136 | 1.157 | 1.164 | 1.194 | 1.024 | 1.014 |
| Hroby_7 | 1.129 | 1.165 | 1.255 | 1.281 | 1.269 | 1.178 | 1.163 | 1.035 |
| Hroby_8 | 1.014 | 1.102 | 1.144 | 1.168 | 1.173 | 1.192 | 0.995 | 0.942 |
| Hroby_9 | 1.289 | 1.307 | 1.208 | 1.468 | 1.269 | 1.336 | 1.341 | 1.244 |
| | Ci_3 | Ci_4 | Ci_5 | Ci_6 | Ci_7 | Ci_8 | Ci_9 | Hk_1 |

**Table 7**
The Delta distances between the samples (MFW analysis of 100 words; part III)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Hrobník_3 | 0.786 | | | | | | | |
| Hrobník_4 | 0.739 | 0.890 | | | | | | |
| Hrobník_5 | 0.756 | 0.973 | 0.744 | | | | | |
| Hrobník_6 | 0.758 | 1.010 | 0.698 | 0.818 | | | | |
| Hrobník_7 | 0.707 | 0.793 | 0.763 | 0.886 | 0.759 | | | |
| Hrobník_8 | 1.138 | 1.366 | 1.141 | 1.136 | 1.125 | 1.157 | | |
| Hroby_1 | 1.057 | 1.036 | 1.016 | 1.140 | 1.139 | 0.913 | 1.441 | |
| Hroby_10 | 1.052 | 1.169 | 1.069 | 1.225 | 1.140 | 0.969 | 1.434 | 0.686 |
| Hroby_11 | 0.995 | 1.118 | 1.006 | 1.109 | 1.072 | 0.835 | 1.335 | 0.657 | 0.542 |
| Hroby_12 | 1.317 | 1.311 | 1.192 | 1.283 | 1.223 | 1.168 | 1.450 | 0.687 | 0.637 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Hroby_2 | 1.085 | 1.232 | 1.117 | 1.181 | 1.178 | 1.112 | 1.418 | 0.726 | 0.896 |
| Hroby_3 | 1.456 | 1.479 | 1.292 | 1.442 | 1.373 | 1.267 | 1.663 | 0.923 | 0.806 |
| Hroby_4 | 1.047 | 1.040 | 1.049 | 1.136 | 1.057 | 0.961 | 1.435 | 0.642 | 0.625 |
| Hroby_5 | 1.204 | 1.210 | 1.229 | 1.254 | 1.149 | 1.094 | 1.517 | 0.885 | 0.899 |
| Hroby_6 | 1.018 | 1.071 | 0.966 | 1.080 | 1.044 | 0.950 | 1.410 | 0.880 | 0.744 |
| Hroby_7 | 1.061 | 1.171 | 1.051 | 1.265 | 1.105 | 0.983 | 1.440 | 0.627 | 0.701 |
| Hroby_8 | 1.091 | 1.044 | 1.081 | 1.173 | 1.069 | 0.912 | 1.370 | 0.663 | 0.649 |
| Hroby_9 | 1.172 | 1.177 | 1.262 | 1.334 | 1.241 | 1.173 | 1.624 | 0.730 | 0.651 |
| | Hk_2 | Hk_3 | Hk_4 | Hk_5 | Hk_6 | Hk_7 | Hk_8 | Hy_1 | Hy_10 |

**Table 8**

The Delta distances between the samples (MFW analysis of 100 words; part IV)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Hroby_12 | 0.685 | | | | | | | | |
| Hroby_2 | 0.777 | 0.931 | | | | | | | |
| Hroby_3 | 0.823 | 0.869 | 0.956 | | | | | | |
| Hroby_4 | 0.642 | 0.787 | 0.879 | 0.937 | | | | | |
| Hroby_5 | 0.914 | 1.005 | 0.945 | 1.072 | 0.831 | | | | |
| Hroby_6 | 0.731 | 0.914 | 1.079 | 0.968 | 0.752 | 1.003 | | | |
| Hroby_7 | 0.627 | 0.776 | 0.741 | 0.888 | 0.769 | 0.928 | 0.812 | | |
| Hroby_8 | 0.556 | 0.758 | 0.858 | 0.843 | 0.626 | 0.814 | 0.758 | 0.675 | |
| Hroby_9 | 0.737 | 0.909 | 0.906 | 0.856 | 0.813 | 0.933 | 0.957 | 0.699 | 0.757 |
| | Hy_11 | Hy_12 | Hy_2 | Hy_3 | Hy_4 | Hy_5 | Hy_6 | Hy_7 | Hy_8 |

**Table 9**

The Delta distances between the samples (MFW analysis of 200 words; part I)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Cikáni_10 | 1.041 | | | | | | | |
| Cikáni_11 | 0.945 | 1.177 | | | | | | |
| Cikáni_12 | 0.957 | 0.951 | 1.106 | | | | | |
| Cikáni_13 | 0.949 | 0.989 | 0.973 | 0.882 | | | | |
| Cikáni_14 | 0.975 | 0.941 | 0.985 | 0.888 | 0.908 | | | |
| Cikáni_15 | 0.899 | 1.117 | 0.939 | 1.006 | 0.868 | 1.055 | | |
| Cikáni_2 | 0.923 | 1.172 | 1.065 | 1.133 | 1.165 | 1.141 | 1.208 | |
| Cikáni_3 | 0.907 | 1.100 | 1.075 | 1.008 | 0.961 | 1.070 | 1.053 | 1.128 |
| Cikáni_4 | 0.838 | 1.073 | 0.976 | 0.949 | 0.997 | 1.078 | 0.972 | 1.033 |
| Cikáni_5 | 0.849 | 0.970 | 1.095 | 0.999 | 1.010 | 1.013 | 1.155 | 0.920 |
| Cikáni_6 | 0.808 | 1.038 | 0.976 | 0.882 | 0.870 | 0.935 | 0.894 | 1.075 |
| Cikáni_7 | 0.916 | 0.956 | 1.137 | 0.976 | 1.008 | 1.064 | 1.032 | 0.874 |
| Cikáni_8 | 0.979 | 1.116 | 1.054 | 1.045 | 1.076 | 1.080 | 1.021 | 1.057 |
| Cikáni_9 | 0.701 | 1.060 | 0.905 | 0.955 | 0.891 | 0.978 | 0.804 | 0.931 |
| Hrobník_1 | 0.844 | 1.017 | 0.973 | 1.083 | 1.037 | 1.076 | 1.055 | 1.073 |
| Hrobník_2 | 0.927 | 0.926 | 1.097 | 0.910 | 1.005 | 1.012 | 1.046 | 1.171 |
| Hrobník_3 | 1.047 | 1.048 | 1.206 | 1.071 | 1.103 | 1.174 | 1.203 | 1.302 |
| Hrobník_4 | 0.905 | 0.994 | 1.016 | 1.019 | 0.897 | 0.970 | 0.992 | 1.154 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Hrobník_5 | 1.011 | 1.128 | 1.097 | 1.104 | 1.104 | 1.150 | 1.004 | 1.328 |
| Hrobník_6 | 0.921 | 1.029 | 1.151 | 0.976 | 1.011 | 1.067 | 0.966 | 1.172 |
| Hrobník_7 | 0.935 | 1.012 | 1.082 | 1.017 | 1.039 | 1.048 | 1.095 | 1.139 |
| Hrobník_8 | 1.179 | 1.337 | 1.300 | 1.360 | 1.322 | 1.411 | 1.177 | 1.476 |
| Hroby_1 | 0.964 | 1.160 | 1.080 | 1.124 | 1.123 | 1.221 | 1.273 | 1.169 |
| Hroby_10 | 1.038 | 1.124 | 1.120 | 1.175 | 1.228 | 1.148 | 1.287 | 1.081 |
| Hroby_11 | 0.921 | 1.071 | 1.063 | 1.119 | 1.132 | 1.095 | 1.224 | 1.027 |
| Hroby_12 | 1.162 | 1.274 | 1.247 | 1.299 | 1.322 | 1.287 | 1.393 | 1.290 |
| Hroby_2 | 1.012 | 1.258 | 1.151 | 1.254 | 1.150 | 1.252 | 1.314 | 1.191 |
| Hroby_3 | 1.196 | 1.277 | 1.225 | 1.378 | 1.373 | 1.333 | 1.380 | 1.239 |
| Hroby_4 | 0.954 | 1.082 | 1.093 | 1.111 | 1.131 | 1.124 | 1.228 | 1.057 |
| Hroby_5 | 1.088 | 1.266 | 1.229 | 1.279 | 1.288 | 1.369 | 1.352 | 1.265 |
| Hroby_6 | 0.968 | 1.117 | 1.103 | 1.129 | 1.193 | 1.125 | 1.248 | 1.162 |
| Hroby_7 | 0.936 | 1.169 | 1.180 | 1.188 | 1.172 | 1.161 | 1.245 | 1.126 |
| Hroby_8 | 0.925 | 1.125 | 1.121 | 1.163 | 1.182 | 1.187 | 1.223 | 1.047 |
| Hroby_9 | 1.081 | 1.203 | 1.239 | 1.237 | 1.309 | 1.266 | 1.375 | 1.113 |
| | Ci_1 | Ci_10 | Ci_11 | Ci_12 | Ci_13 | Ci_14 | Ci_15 | Ci_2 |

**Table 10**

The Delta distances between the samples (MFW analysis of 200 words; part II)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Cikáni_4 | 0.969 | | | | | | | |
| Cikáni_5 | 0.966 | 0.941 | | | | | | |
| Cikáni_6 | 0.955 | 0.933 | 0.903 | | | | | |
| Cikáni_7 | 0.981 | 0.835 | 0.838 | 0.918 | | | | |
| Cikáni_8 | 1.139 | 1.087 | 1.085 | 0.957 | 1.045 | | | |
| Cikáni_9 | 0.913 | 0.839 | 0.907 | 0.930 | 0.858 | 0.966 | | |
| Hrobník_1 | 1.009 | 1.016 | 0.981 | 0.984 | 0.968 | 1.127 | 0.945 | |
| Hrobník_2 | 1.089 | 0.978 | 0.907 | 0.868 | 0.990 | 1.156 | 1.062 | 0.826 |
| Hrobník_3 | 1.071 | 1.076 | 1.015 | 1.050 | 1.119 | 1.220 | 1.125 | 0.944 |
| Hrobník_4 | 1.026 | 1.004 | 1.031 | 0.896 | 0.948 | 1.017 | 0.961 | 0.816 |
| Hrobník_5 | 1.201 | 1.075 | 1.152 | 0.964 | 1.073 | 1.222 | 1.098 | 0.994 |
| Hrobník_6 | 1.047 | 1.004 | 0.985 | 0.944 | 1.010 | 1.088 | 0.970 | 0.886 |
| Hrobník_7 | 0.978 | 1.031 | 0.926 | 0.939 | 1.026 | 1.139 | 0.985 | 0.766 |
| Hrobník_8 | 1.324 | 1.288 | 1.360 | 1.197 | 1.226 | 1.375 | 1.300 | 1.146 |
| Hroby_1 | 1.101 | 1.157 | 1.070 | 1.128 | 1.102 | 1.216 | 1.073 | 0.881 |
| Hroby_10 | 1.130 | 1.167 | 1.057 | 1.162 | 1.126 | 1.219 | 1.134 | 0.925 |
| Hroby_11 | 1.066 | 1.082 | 0.965 | 1.080 | 1.008 | 1.170 | 0.979 | 0.866 |
| Hroby_12 | 1.293 | 1.293 | 1.249 | 1.270 | 1.254 | 1.307 | 1.195 | 1.092 |
| Hroby_2 | 1.136 | 1.246 | 1.103 | 1.181 | 1.181 | 1.295 | 1.142 | 1.001 |
| Hroby_3 | 1.276 | 1.357 | 1.311 | 1.361 | 1.270 | 1.358 | 1.236 | 1.167 |
| Hroby_4 | 1.096 | 1.110 | 1.062 | 1.092 | 1.107 | 1.182 | 1.041 | 0.911 |
| Hroby_5 | 1.198 | 1.203 | 1.119 | 1.225 | 1.247 | 1.331 | 1.184 | 1.093 |
| Hroby_6 | 1.126 | 1.086 | 1.094 | 1.139 | 1.139 | 1.204 | 1.003 | 0.929 |
| Hroby_7 | 1.105 | 1.115 | 1.093 | 1.134 | 1.125 | 1.163 | 1.078 | 0.896 |
| Hroby_8 | 1.078 | 1.130 | 1.037 | 1.132 | 1.125 | 1.211 | 0.985 | 0.890 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Hroby_9 | 1.200 | 1.214 | 1.128 | 1.303 | 1.179 | 1.294 | 1.192 | 1.039 |
| | Ci_3 | Ci_4 | Ci_5 | Ci_6 | Ci_7 | Ci_8 | Ci_9 | Hk_1 |

**Table 11**

The Delta distances between the samples (MFW analysis of 200 words; part III)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Hrobník_3 | 0.806 | | | | | | | | |
| Hrobník_4 | 0.730 | 0.838 | | | | | | | |
| Hrobník_5 | 0.763 | 0.986 | 0.766 | | | | | | |
| Hrobník_6 | 0.742 | 0.883 | 0.720 | 0.807 | | | | | |
| Hrobník_7 | 0.698 | 0.842 | 0.779 | 0.863 | 0.731 | | | | |
| Hrobník_8 | 1.101 | 1.307 | 1.103 | 1.070 | 1.074 | 1.132 | | | |
| Hroby_1 | 0.978 | 1.041 | 0.973 | 1.118 | 1.056 | 0.922 | 1.390 | | |
| Hroby_10 | 1.048 | 1.169 | 1.052 | 1.253 | 1.078 | 0.980 | 1.407 | 0.713 | |
| Hroby_11 | 0.928 | 1.069 | 1.001 | 1.126 | 1.016 | 0.861 | 1.369 | 0.668 | 0.572 |
| Hroby_12 | 1.276 | 1.289 | 1.208 | 1.335 | 1.218 | 1.208 | 1.480 | 0.770 | 0.680 |
| Hroby_2 | 1.066 | 1.162 | 1.097 | 1.226 | 1.156 | 1.070 | 1.421 | 0.713 | 0.833 |
| Hroby_3 | 1.289 | 1.397 | 1.245 | 1.399 | 1.246 | 1.191 | 1.568 | 0.862 | 0.774 |
| Hroby_4 | 1.008 | 1.055 | 1.044 | 1.179 | 0.999 | 0.959 | 1.386 | 0.641 | 0.626 |
| Hroby_5 | 1.207 | 1.248 | 1.260 | 1.303 | 1.109 | 1.130 | 1.447 | 0.926 | 0.899 |
| Hroby_6 | 0.970 | 1.064 | 0.985 | 1.101 | 0.990 | 0.948 | 1.372 | 0.801 | 0.731 |
| Hroby_7 | 1.022 | 1.115 | 1.035 | 1.232 | 1.014 | 0.953 | 1.392 | 0.630 | 0.674 |
| Hroby_8 | 1.079 | 1.076 | 1.105 | 1.232 | 1.049 | 0.971 | 1.342 | 0.679 | 0.658 |
| Hroby_9 | 1.087 | 1.142 | 1.189 | 1.296 | 1.132 | 1.094 | 1.485 | 0.697 | 0.669 |
| | Hk_2 | Hk_3 | Hk_4 | Hk_5 | Hk_6 | Hk_7 | Hk_8 | Hy_1 | Hy_10 |

**Table 12**

The Delta distances between the samples (MFW analysis of 200 words; part IV)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Hroby_12 | 0.751 | | | | | | | | |
| Hroby_2 | 0.792 | 0.929 | | | | | | | |
| Hroby_3 | 0.832 | 0.895 | 0.885 | | | | | | |
| Hroby_4 | 0.625 | 0.805 | 0.794 | 0.874 | | | | | |
| Hroby_5 | 0.911 | 1.009 | 0.949 | 1.057 | 0.842 | | | | |
| Hroby_6 | 0.714 | 0.891 | 0.945 | 0.933 | 0.727 | 0.969 | | | |
| Hroby_7 | 0.611 | 0.782 | 0.728 | 0.884 | 0.660 | 0.874 | 0.749 | | |
| Hroby_8 | 0.597 | 0.787 | 0.811 | 0.817 | 0.642 | 0.838 | 0.727 | 0.637 | |
| Hroby_9 | 0.682 | 0.869 | 0.876 | 0.840 | 0.761 | 0.974 | 0.875 | 0.674 | 0.714 |
| | Hy_11 | Hy_12 | Hy_2 | Hy_3 | Hy_4 | Hy_5 | Hy_6 | Hy_7 | Hy_8 |

**Table 13**
The Delta distances between the samples (MFW analysis of 100 bigrams; part I)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Cikáni_10 | 1.151 | | | | | | | |
| Cikáni_11 | 1.092 | 1.196 | | | | | | |
| Cikáni_12 | 0.985 | 1.102 | 1.108 | | | | | |
| Cikáni_13 | 1.198 | 1.267 | 1.157 | 1.003 | | | | |
| Cikáni_14 | 1.281 | 1.087 | 1.145 | 1.280 | 1.283 | | | |
| Cikáni_15 | 0.938 | 1.358 | 1.308 | 1.293 | 1.206 | 1.462 | | |
| Cikáni_2 | 1.145 | 1.123 | 1.244 | 1.270 | 1.403 | 1.100 | 1.383 | |
| Cikáni_3 | 0.964 | 1.324 | 1.130 | 1.108 | 1.070 | 1.222 | 1.109 | 1.232 |
| Cikáni_4 | 0.997 | 1.187 | 1.103 | 1.147 | 1.185 | 1.156 | 1.151 | 0.923 |
| Cikáni_5 | 0.945 | 1.178 | 1.141 | 1.178 | 1.216 | 1.116 | 1.136 | 0.846 |
| Cikáni_6 | 0.944 | 1.331 | 1.177 | 1.241 | 1.288 | 1.230 | 1.188 | 1.173 |
| Cikáni_7 | 1.092 | 1.396 | 1.245 | 1.144 | 1.311 | 1.307 | 1.318 | 0.941 |
| Cikáni_8 | 1.067 | 1.239 | 1.055 | 1.091 | 1.030 | 1.244 | 1.166 | 1.141 |
| Cikáni_9 | 0.877 | 1.201 | 1.131 | 1.018 | 1.090 | 1.260 | 1.143 | 0.965 |
| Hrobník_1 | 1.127 | 1.080 | 1.097 | 1.222 | 1.368 | 1.244 | 1.377 | 1.128 |
| Hrobník_2 | 0.965 | 1.226 | 1.274 | 1.224 | 1.355 | 1.283 | 1.192 | 1.260 |
| Hrobník_3 | 1.303 | 1.268 | 1.487 | 1.451 | 1.572 | 1.397 | 1.583 | 1.310 |
| Hrobník_4 | 1.108 | 1.328 | 1.285 | 1.238 | 1.343 | 1.309 | 1.261 | 1.282 |
| Hrobník_5 | 1.154 | 1.382 | 1.468 | 1.391 | 1.558 | 1.331 | 1.348 | 1.302 |
| Hrobník_6 | 1.084 | 1.141 | 1.287 | 1.261 | 1.333 | 1.175 | 1.179 | 1.101 |
| Hrobník_7 | 0.940 | 1.144 | 1.128 | 1.212 | 1.220 | 1.223 | 1.153 | 1.149 |
| Hrobník_8 | 1.256 | 1.451 | 1.459 | 1.701 | 1.694 | 1.565 | 1.546 | 1.604 |
| Hroby_1 | 0.953 | 1.038 | 0.960 | 1.170 | 1.188 | 1.128 | 1.237 | 1.085 |
| Hroby_10 | 1.221 | 1.067 | 1.005 | 1.251 | 1.366 | 1.026 | 1.410 | 1.060 |
| Hroby_11 | 0.997 | 0.940 | 0.984 | 1.159 | 1.249 | 1.066 | 1.292 | 0.972 |
| Hroby_12 | 1.268 | 1.279 | 0.972 | 1.262 | 1.417 | 1.280 | 1.497 | 1.267 |
| Hroby_2 | 0.987 | 1.171 | 0.979 | 1.221 | 1.236 | 1.284 | 1.233 | 1.118 |
| Hroby_3 | 1.330 | 1.246 | 1.159 | 1.363 | 1.488 | 1.335 | 1.558 | 1.252 |
| Hroby_4 | 1.110 | 1.035 | 0.957 | 1.198 | 1.320 | 1.009 | 1.328 | 0.937 |
| Hroby_5 | 1.124 | 1.349 | 1.273 | 1.482 | 1.444 | 1.384 | 1.419 | 1.059 |
| Hroby_6 | 1.206 | 1.084 | 1.124 | 1.238 | 1.422 | 1.070 | 1.500 | 1.193 |
| Hroby_7 | 0.958 | 1.077 | 0.937 | 1.129 | 1.273 | 1.107 | 1.328 | 0.981 |
| Hroby_8 | 1.066 | 1.022 | 0.982 | 1.267 | 1.329 | 1.081 | 1.338 | 0.947 |
| Hroby_9 | 1.129 | 1.103 | 1.149 | 1.327 | 1.396 | 1.224 | 1.418 | 1.088 |
| | Ci_1 | Ci_10 | Ci_11 | Ci_12 | Ci_13 | Ci_14 | Ci_15 | Ci_2 |

**Table 14**
The Delta distances between the samples (MFW analysis of 100 bigrams; part II)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Cikáni_4 | 0.992 | | | | | | | |
| Cikáni_5 | 0.941 | 0.825 | | | | | | |
| Cikáni_6 | 0.980 | 1.124 | 0.970 | | | | | |
| Cikáni_7 | 1.113 | 0.991 | 0.831 | 1.064 | | | | |
| Cikáni_8 | 1.074 | 1.057 | 1.002 | 1.158 | 1.074 | | | |
| Cikáni_9 | 1.013 | 0.965 | 0.896 | 1.055 | 1.051 | 0.960 | | |
| Hrobník_1 | 1.285 | 1.129 | 1.027 | 1.236 | 1.308 | 1.318 | 1.097 | |
| Hrobník_2 | 1.122 | 1.118 | 0.892 | 0.959 | 1.094 | 1.193 | 1.049 | 0.963 |
| Hrobník_3 | 1.370 | 1.450 | 1.166 | 1.294 | 1.352 | 1.350 | 1.445 | 1.091 |
| Hrobník_4 | 1.093 | 1.207 | 0.992 | 1.125 | 1.193 | 1.092 | 1.024 | 0.948 |
| Hrobník_5 | 1.375 | 1.379 | 1.215 | 1.045 | 1.433 | 1.348 | 1.292 | 1.120 |
| Hrobník_6 | 1.162 | 1.044 | 0.965 | 1.130 | 1.183 | 1.155 | 1.067 | 0.992 |
| Hrobník_7 | 1.086 | 1.091 | 0.849 | 1.090 | 1.104 | 1.081 | 1.068 | 0.846 |
| Hrobník_8 | 1.472 | 1.598 | 1.461 | 1.545 | 1.593 | 1.571 | 1.505 | 1.277 |
| Hroby_1 | 1.044 | 1.093 | 0.796 | 1.074 | 1.063 | 1.023 | 1.009 | 0.872 |
| Hroby_10 | 1.226 | 1.098 | 0.903 | 1.317 | 1.138 | 1.174 | 1.088 | 0.911 |
| Hroby_11 | 1.098 | 0.999 | 0.794 | 1.190 | 1.144 | 1.079 | 0.993 | 0.878 |
| Hroby_12 | 1.287 | 1.258 | 1.126 | 1.338 | 1.293 | 1.260 | 1.156 | 0.943 |
| Hroby_2 | 1.068 | 1.078 | 0.866 | 1.221 | 1.203 | 1.149 | 0.997 | 0.846 |
| Hroby_3 | 1.257 | 1.244 | 1.135 | 1.505 | 1.315 | 1.303 | 1.229 | 1.134 |
| Hroby_4 | 1.116 | 1.012 | 0.825 | 1.204 | 1.093 | 1.096 | 1.080 | 0.942 |
| Hroby_5 | 1.275 | 1.218 | 0.979 | 1.227 | 1.272 | 1.295 | 1.283 | 1.071 |
| Hroby_6 | 1.280 | 1.196 | 1.064 | 1.361 | 1.338 | 1.247 | 1.136 | 0.984 |
| Hroby_7 | 1.076 | 1.138 | 0.898 | 1.182 | 1.168 | 1.127 | 0.992 | 0.898 |
| Hroby_8 | 1.133 | 1.005 | 0.850 | 1.165 | 1.165 | 1.180 | 1.043 | 0.925 |
| Hroby_9 | 1.201 | 1.151 | 0.945 | 1.305 | 1.212 | 1.260 | 1.067 | 0.984 |
| | Ci_3 | Ci_4 | Ci_5 | Ci_6 | Ci_7 | Ci_8 | Ci_9 | Hk_1 |

**Table 15**
The Delta distances between the samples (MFW analysis of 100 bigrams; part III)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Hrobník_3 | 0.932 | | | | | | | | |
| Hrobník_4 | 0.707 | 1.029 | | | | | | | |
| Hrobník_5 | 0.852 | 1.024 | 0.948 | | | | | | |
| Hrobník_6 | 0.869 | 0.963 | 0.889 | 0.995 | | | | | |
| Hrobník_7 | 0.671 | 0.855 | 0.765 | 0.967 | 0.794 | | | | |
| Hrobník_8 | 1.227 | 1.515 | 1.357 | 1.484 | 1.415 | 1.263 | | | |
| Hroby_1 | 0.834 | 1.005 | 1.022 | 1.098 | 0.953 | 0.736 | 1.284 | | |
| Hroby_10 | 1.138 | 1.257 | 1.195 | 1.308 | 1.026 | 0.989 | 1.412 | 0.651 | |
| Hroby_11 | 0.952 | 1.174 | 1.045 | 1.170 | 1.001 | 0.895 | 1.346 | 0.634 | 0.570 |
| Hroby_12 | 1.150 | 1.308 | 1.229 | 1.417 | 1.136 | 1.059 | 1.419 | 0.679 | 0.622 |
| Hroby_2 | 1.014 | 1.133 | 1.062 | 1.196 | 1.067 | 0.864 | 1.247 | 0.627 | 0.767 |
| Hroby_3 | 1.298 | 1.433 | 1.323 | 1.546 | 1.219 | 1.130 | 1.593 | 0.836 | 0.693 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Hroby_4 | 0.986 | 1.161 | 1.110 | 1.196 | 1.033 | 0.939 | 1.409 | 0.598 | 0.543 |
| Hroby_5 | 1.088 | 1.191 | 1.391 | 1.257 | 1.154 | 1.068 | 1.378 | 0.840 | 1.103 |
| Hroby_6 | 1.135 | 1.241 | 1.199 | 1.223 | 1.145 | 1.068 | 1.419 | 0.815 | 0.723 |
| Hroby_7 | 1.069 | 1.129 | 1.149 | 1.235 | 1.028 | 0.910 | 1.339 | 0.620 | 0.703 |
| Hroby_8 | 1.035 | 1.152 | 1.208 | 1.275 | 1.008 | 0.920 | 1.330 | 0.632 | 0.608 |
| Hroby_9 | 1.157 | 1.255 | 1.255 | 1.338 | 0.994 | 1.050 | 1.462 | 0.719 | 0.625 |
| | Hk_2 | Hk_3 | Hk_4 | Hk_5 | Hk_6 | Hk_7 | Hk_8 | Hy_1 | Hy_10 |

**Table 16**

The Delta distances between the samples (MFW analysis of 100 bigrams; part IV)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Hroby_12 | 0.767 | | | | | | | | |
| Hroby_2 | 0.662 | 0.785 | | | | | | | |
| Hroby_3 | 0.753 | 0.741 | 0.860 | | | | | | |
| Hroby_4 | 0.518 | 0.682 | 0.731 | 0.764 | | | | | |
| Hroby_5 | 0.948 | 1.102 | 0.825 | 1.265 | 0.861 | | | | |
| Hroby_6 | 0.681 | 0.832 | 0.940 | 0.916 | 0.680 | 1.133 | | | |
| Hroby_7 | 0.660 | 0.600 | 0.593 | 0.835 | 0.620 | 0.834 | 0.808 | | |
| Hroby_8 | 0.509 | 0.720 | 0.647 | 0.766 | 0.571 | 0.814 | 0.788 | 0.576 | |
| Hroby_9 | 0.560 | 0.761 | 0.736 | 0.740 | 0.710 | 0.996 | 0.809 | 0.640 | 0.598 |
| | Hy_11 | Hy_12 | Hy_2 | Hy_3 | Hy_4 | Hy_5 | Hy_6 | Hy_7 | Hy_8 |

*Glottometrics 46, 2019, 29-40*

# Types of Syllable Distribution in Russian Long Poems

*Sergey Andreev*[1]

**Abstract.** In the present examination, syllable distribution in 15 long poems written during the period of the end of the 20th and the beginning of the 21st centuries by Russian authors is studied. The distribution of different types of syllables as well as the relationship of the initial and the last syllabic positions in verse lines are explored using one of non-parametric methods – the Kendall's rank correlation coefficient. The syllabic types are formed by vowel–consonant sequences according to the principles of sonorant theory. The results revealed that the rank-frequency distribution of syllabic types in all poems under study is well fitted by the exponential function. Within the poems, the initial and final syllables were found to form a strong opposition, laying the basis for syllabic asymmetry of the verse line.

***Keywords***: *Russian*, *long poems, the exponential function, syllable types, the Kendall rank correlation coefficient, initial and final positions.*

This article continues the study of syllable distribution in Russian poetry. Our previous research (Andreev 2018) was based on the material of the 19th and the beginning of the 20th centuries (the Golden and Silver Ages of Russian poetry), the present study focuses on modern poetry of the second half of the 20th century and the 2000s. The data-base consists of 15 modern long poems by Russian authors, which were selected to represent three periods of important social stages in the life of the country: the 1960s ("Thaw"), the 1990s ("Reforms"), and the 2000s ("Modernity"). The list of poems is given in the Appendix. In all cases, samples were taken from the beginnings of the poems (of 700 to 1000 words). The main principles of syllable division were described in Andreev (2018), and are based on sonorant theory (see Köhler, Altmann 2014, p. 135).

The count of different types of syllables in the poems gave the results which are represented in Table 1 (C – consonant, V – vowel).

**Table 1**
The number of different types of syllables in the samples

| Type | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 |
|---|---|---|---|---|---|---|---|---|
| V | 63 | 119 | 108 | 93 | 81 | 117 | 83 | 114 |
| CV | 855 | 794 | 854 | 752 | 666 | 818 | 1093 | 679 |
| CCV | 162 | 113 | 134 | 160 | 142 | 124 | 213 | 147 |
| CVC | 413 | 466 | 438 | 443 | 427 | 472 | 540 | 472 |
| CCVC | 79 | 89 | 70 | 59 | 67 | 67 | 80 | 66 |
| CVCC | 20 | 18 | 4 | 49 | 11 | 26 | 17 | 3 |
| CCCV | 8 | 20 | 12 | 11 | 17 | 11 | 23 | 1 |
| VC | 63 | 92 | 41 | 40 | 45 | 42 | 81 | 71 |
| CCVCC | 3 | 4 | 0 | 4 | 2 | 2 | 1 | 1 |
| CCCVCC | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| CCCVC | 9 | 4 | 9 | 9 | 6 | 4 | 10 | 1 |

[1] Smolensk State University, Russia. Email: smol.an@mail.ru.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| VCC | 0 | 1 | 0 | 3 | 0 | 0 | 2 | 0 |
| CCCCV | 2 | 1 | 0 | 2 | 2 | 0 | 3 | 0 |
| CVCCCC | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| CCCCVC | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| CVCCC | 0 | 0 | 2 | 2 | 0 | 0 | 1 | 0 |

| Type | T9 | T10 | T11 | T12 | T13 | T14 | T15 |
|---|---|---|---|---|---|---|---|
| V | 91 | 40 | 105 | 66 | 55 | 82 | 87 |
| CV | 702 | 831 | 970 | 567 | 685 | 942 | 717 |
| CCV | 140 | 92 | 146 | 108 | 137 | 130 | 128 |
| CVC | 381 | 366 | 497 | 313 | 458 | 421 | 498 |
| CCVC | 78 | 65 | 72 | 49 | 56 | 59 | 87 |
| CVCC | 13 | 54 | 24 | 18 | 12 | 11 | 27 |
| CCCV | 15 | 14 | 17 | 10 | 6 | 17 | 10 |
| VC | 44 | 40 | 51 | 39 | 41 | 45 | 51 |
| CCVCC | 2 | 12 | 3 | 1 | 0 | 0 | 5 |
| CCCVCC | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| CCCVC | 12 | 13 | 7 | 5 | 11 | 4 | 9 |
| VCC | 0 | 0 | 1 | 0 | 3 | 0 | 0 |
| CCCCV | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| CVCCCC | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| CCCCVC | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CVCCC | 3 | 2 | 0 | 0 | 0 | 1 | 4 |

After rank-ordering of these numbers, for each poem the exponential function was used (Andreev, Místecký, Altmann 2018):

$$y = a * e^{-bx}$$

The results are presented in Table 2. Those types which were not found in a poem were omitted.

**Table 2**
Ranked distribution of syllable types in 15 long poems

| | T1 | | T2 | | T3 | | T4 | |
|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | Observed | Expected | Observed | Expected | Observed | Expected |
| 1 | 855 | 855.17 | 794 | 801.11 | 854 | 858.04 | 752 | 762.43 |
| 2 | 413 | 402.28 | 466 | 413.60 | 438 | 408.29 | 443 | 397.34 |
| 3 | 162 | 189.23 | 119 | 213.53 | 134 | 194.28 | 160 | 207.07 |
| 4 | 79 | 89.02 | 113 | 110.24 | 108 | 92.45 | 93 | 107.91 |
| 5 | 63 | 41.87 | 92 | 56.91 | 70 | 43.99 | 59 | 56.24 |
| 6 | 63 | 19.70 | 89 | 29.38 | 41 | 20.93 | 49 | 29.31 |
| 7 | 20 | 9.27 | 20 | 15.17 | 12 | 9.96 | 40 | 15.27 |
| 8 | 9 | 4.36 | 18 | 7.83 | 9 | 4.74 | 11 | 7.96 |
| 9 | 8 | 2.05 | 4 | 4.04 | 4 | 2.26 | 9 | 4.15 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 10 | 3 | 0.96 | 4 | 2.09 | 2 | 1.07 | 4 | 2.16 |
| 11 | 2 | 0.45 | 1 | 1.08 | – | – | 3 | 1.13 |
| 12 | 1 | 0.21 | 1 | 0.56 | – | – | 2 | 0.59 |
| 13 | – | – | 1 | 0.29 | – | – | 2 | 0.31 |
| 14 | – | – | – | – | – | – | 1 | 0.16 |
| 15 | – | – | – | – | – | – | 1 | 0.08 |
| | a = 1817.927<br>b = 0.754<br>$R^2$ = 0.9952 | | a = 1551.718<br>b = 0.661<br>$R^2$ = 0.9749 | | a = 1803.198<br>b = 0.743<br>$R^2$ = 0.9913 | | a = 1462.979<br>b = 0.652<br>$R^2$ = 0.9909 | |

| | T5 | | T6 | | T7 | | T8 | |
|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | Observed | Expected | Observed | Expected | Observed | Expected |
| 1 | 666 | 682.34 | 818 | 829.31 | 1093 | 1096.41 | 679 | 701.37 |
| 2 | 427 | 364.65 | 472 | 414.53 | 540 | 517.05 | 472 | 391.40 |
| 3 | 142 | 194.88 | 124 | 207.20 | 213 | 243.83 | 147 | 218.42 |
| 4 | 81 | 104.14 | 117 | 103.57 | 83 | 114.98 | 114 | 121.89 |
| 5 | 67 | 55.66 | 67 | 51.77 | 81 | 54.22 | 71 | 68.02 |
| 6 | 45 | 29.74 | 42 | 25.88 | 80 | 25.57 | 66 | 37.96 |
| 7 | 17 | 15.90 | 26 | 12.93 | 23 | 12.06 | 3 | 21.18 |
| 8 | 11 | 8.49 | 11 | 6.47 | 17 | 5.69 | 1 | 11.82 |
| 9 | 6 | 4.54 | 4 | 3.23 | 10 | 2.68 | 1 | 6.60 |
| 10 | 2 | 2.43 | 2 | 1.62 | 3 | 1.27 | 1 | 3.68 |
| 11 | 2 | 1.30 | 1 | 0.81 | 2 | 0.60 | – | – |
| 12 | 1 | 0.69 | – | – | 1 | 0.28 | – | – |
| 13 | – | – | – | – | 1 | 0.13 | – | – |
| | a = 1817.927<br>b = 0.754<br>$R^2$ = 0.9952 | | a = 1551.718<br>b = 0.661<br>$R^2$ = 0.9749 | | a = 1803.198<br>b = 0.743<br>$R^2$ = 0.99132 | | a = 1462.979<br>b = 0.652<br>$R^2$ = 0.9909 | |

| | T9 | | T10 | | T11 | | T12 | |
|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | Observed | Expected | Observed | Expected | Observed | Expected |
| 1 | 702 | 704.23 | 831 | 833.44 | 970 | 976.45 | 567 | 571.68 |
| 2 | 381 | 359.42 | 366 | 344.03 | 497 | 456.98 | 313 | 287.16 |
| 3 | 140 | 183.44 | 92 | 142.01 | 146 | 213.86 | 108 | 144.24 |
| 4 | 91 | 93.62 | 65 | 58.62 | 105 | 100.09 | 66 | 72.45 |
| 5 | 78 | 47.78 | 54 | 24.20 | 72 | 46.84 | 49 | 36.39 |
| 6 | 44 | 24.39 | 40 | 9.99 | 51 | 21.92 | 39 | 18.28 |
| 7 | 15 | 12.45 | 40 | 4.12 | 24 | 10.26 | 18 | 9.18 |
| 8 | 13 | 6.35 | 14 | 1.70 | 17 | 4.80 | 10 | 4.61 |
| 9 | 12 | 3.24 | 13 | 0.70 | 7 | 2.25 | 5 | 2.32 |
| 10 | 3 | 1.65 | 12 | 0.29 | 3 | 1.05 | 1 | 1.16 |
| 11 | 2 | 0.84 | 2 | 0.12 | 1 | 0.49 | – | – |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 12 | 1 | 0.43 | 1 | 0.05 | – | – | – | – |
| | a = 1256.822<br>b = 0.583<br>$R^2$ = 0.9723 | | a = 2019.112<br>b = 0.885<br>$R^2$ = 0.9899 | | a = 2086.430<br>b = 0.759<br>$R^2$ = 0.9910 | | a = 1138.109<br>b = 0.689<br>$R^2$ = 0.9909 | |

| | T13 | | T14 | | T15 | |
|---|---|---|---|---|---|---|
| | Observed | Expected | Observed | Expected | Observed | Expected |
| 1 | 685 | 710.03 | 942 | 943.54 | 717 | 741.87 |
| 2 | 458 | 370.11 | 421 | 403.91 | 498 | 400.92 |
| 3 | 137 | 192.92 | 130 | 172.90 | 128 | 216.67 |
| 4 | 56 | 100.56 | 82 | 74.02 | 87 | 117.09 |
| 5 | 55 | 52.42 | 59 | 31.68 | 87 | 63.28 |
| 6 | 41 | 27.32 | 45 | 13.56 | 51 | 34.20 |
| 7 | 12 | 14.24 | 17 | 5.81 | 27 | 18.48 |
| 8 | 11 | 7.42 | 11 | 2.49 | 10 | 9.99 |
| 9 | 6 | 3.87 | 4 | 1.06 | 9 | 5.40 |
| 10 | 3 | 2.02 | 1 | 0.46 | 5 | 2.92 |
| 11 | – | – | 1 | 0.20 | 4 | 1.58 |
| 12 | – | – | – | – | 1 | 0.85 |
| | a = 1362.168<br>b = 0.652<br>$R^2$ = 0.9722 | | a = 2204.149<br>b = 0.848<br>$R^2$ = 0.9950 | | a = 1372.756<br>b = 0.615<br>$R^2$ = 0.9658 | |

The results demonstrate very good fitting of the formula for all 15 texts. Besides the fact that certain regularity was established, two other things should be emphasized. Firstly, the obtained results show that the exponential function has remained without changes for at least 50 years of the modern period. Secondly, it is possible to state that this kind of distribution is present in the poems written by different authors with highly different styles.

Research in the sphere of linguistics of verse has revealed that in verse lines (at least in Russian poetry), there are two most important positions. They are the first and the last positions (Gasparov 2004). The first position is often considered to determine the organization of the verse line (Beliy 2010; Krasnoperova 2000), whereas the last position usually serves to unite lines together. Thus, the beginning element in the line refers to the horizontal aspect of the poem, and the last element is an important factor to organize its vertical ties. In most cases, such elements were studied at the level of words or syntactic ties (Gasparov, Skulacheva 2004). In this examination, we set the goal to study the first and the last elements of verse lines at the level of syllables.

In Table 3, the frequencies of the types of syllables in the first (initial) and the last (final) positions are represented.

**Table 3**
Syllable types in the initial and final positions of the line

| TYPE | T1 | | T2 | | T3 | | T4 | | T5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Initial | Final | Initial | Final | Initial | Final | Initial | Final | Initial | Final |
| V | 18 | 0 | 33 | 1 | 23 | 2 | 41 | 2 | 42 | 1 |
| CV | 68 | 69 | 45 | 74 | 75 | 102 | 56 | 58 | 46 | 66 |
| CCV | 18 | 3 | 20 | 0 | 31 | 6 | 28 | 4 | 15 | 6 |
| CVC | 43 | 75 | 30 | 72 | 29 | 51 | 25 | 74 | 23 | 57 |
| CCVC | 11 | 12 | 5 | 7 | 9 | 5 | 5 | 8 | 3 | 4 |
| CVCC | 0 | 5 | 2 | 3 | 0 | 3 | 2 | 11 | 1 | 2 |
| CCCV | 1 | 3 | 3 | 3 | 0 | 0 | 2 | 2 | 3 | 0 |
| VC | 8 | 0 | 26 | 3 | 6 | 3 | 3 | 2 | 6 | 2 |
| CCVCC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 |
| CCCVCC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CCCVC | 1 | 1 | 0 | 0 | 2 | 2 | 2 | 0 | 1 | 1 |
| CVCCC | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| VCC | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| CVCCCC | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Total | 168 | 168 | 164 | 164 | 175 | 175 | 165 | 165 | 140 | 140 |

| TYPE | T6 | | T7 | | T8 | | T9 | | T10 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Initial | Final | Initial | Final | Initial | Final | Initial | Final | Initial | Final |
| V | 32 | 0 | 22 | 3 | 50 | 0 | 27 | 0 | 11 | 1 |
| CV | 56 | 45 | 69 | 144 | 50 | 54 | 46 | 58 | 67 | 111 |
| CCV | 16 | 3 | 41 | 14 | 28 | 1 | 24 | 1 | 20 | 4 |
| CVC | 23 | 68 | 35 | 20 | 22 | 108 | 25 | 64 | 28 | 48 |
| CCVC | 4 | 10 | 5 | 5 | 2 | 16 | 4 | 13 | 16 | 4 |
| CVCC | 2 | 7 | 1 | 0 | 1 | 0 | 2 | 0 | 13 | 2 |
| CCCV | 1 | 2 | 3 | 1 | 0 | 0 | 5 | 0 | 3 | 0 |
| VC | 4 | 2 | 10 | 0 | 26 | 0 | 9 | 2 | 12 | 0 |
| CCVCC | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 2 |
| CCCVCC | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CCCVC | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 3 | 1 | 1 |
| CVCCC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| VCC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CVCCCC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 139 | 139 | 188 | 188 | 179 | 179 | 142 | 142 | 173 | 173 |

| TYPE | T11 | | T12 | | T13 | | T14 | | T15 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Initial | Final | Initial | Final | Initial | Final | Initial | Final | Initial | Final |
| V | 20 | 0 | 27 | 1 | 28 | 3 | 24 | 0 | 34 | 1 |
| CV | 31 | 48 | 38 | 83 | 67 | 73 | 56 | 104 | 43 | 60 |
| CCV | 13 | 0 | 21 | 2 | 24 | 7 | 26 | 4 | 24 | 6 |
| CVC | 15 | 35 | 36 | 54 | 24 | 57 | 32 | 47 | 26 | 66 |
| CCVC | 5 | 4 | 8 | 5 | 12 | 7 | 10 | 5 | 7 | 14 |
| CVCC | 6 | 8 | 5 | 5 | 0 | 10 | 0 | 2 | 2 | 4 |
| CCCV | 2 | 0 | 2 | 1 | 0 | 0 | 3 | 1 | 2 | 0 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| VC | 5 | 2 | 14 | 2 | 0 | 2 | 13 | 0 | 14 | 1 |
| CCVCC | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CCCVCC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CCCVC | 0 | 0 | 1 | 0 | 0 | 6 | 0 | 1 | 2 | 0 |
| CVCCC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| VCC | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CVCCCC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 98 | 98 | 153 | 153 | 165 | 165 | 164 | 164 | 154 | 154 |

There are two possible ways to proceed, comparing these two positions: using the frequencies themselves, or their ranks. Keeping in mind that the size of the samples is not the same, we chose the second approach and ranked all the frequencies of syllabic types found in both positions of each poem.

To compare the initial and the final syllables in one and the same poem, the Kendall's non-parametric rank correlation coefficient was used (Sheskin 2004, pp. 1079–1092). The results are presented in Table 4. Statistically significant coefficients at $p < 0.05$ are marked with an asterisk.

**Table 4**
Kendall's coefficient

| Text | Initial – Final |
|---|---|
| T1 | 0.23 |
| T2 | 0.35 |
| T3 | 0.67 * |
| T4 | 0.58 * |
| T5 | 0.50 * |
| T6 | 0.34 |
| T7 | 0.70 * |
| T8 | 0.28 |
| T9 | 0.14 |
| T10 | 0.68 * |
| T11 | 0.39 |
| T12 | 0.59 * |
| T13 | 0.30 |
| T14 | 0.38 |
| T15 | 0.40 |

As seen in Table 4, only in 6 cases the beginning and the end are correlated. Out of these poems, five were written by the poets who became famous during the 1960s and formed the basis of a literary movement. Emelin (T12), to some extent, follows the style of the literary trend of the 60s, ironically imitating it.

Afterwards, different poems were compared with the help of the Kendall coefficient for ranked types of syllables, placed in (a) the initial and (b) the final positions of the lines.

The Kendall coefficients of such comparison are presented in Table 5. The upper part of the table contains correlations of the initial syllable types, the lower part – correlations of

syllable types in final positions. Coefficients which are *not* statistically significant for $p < 0.05$ and $df = 28$ are given in bold type.

**Table 5**

Kendall's correlation coefficients of ranked syllable types in the initial and final positions of 15 poems

| | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 | X | 0.77 | 0.90 | 0.86 | 0.85 | 0.80 | 0.87 | 0.65 | 0.77 | 0.64 |
| T2 | 0.63 | X | 0.63 | 0.81 | 0.92 | 0.90 | 0.76 | 0.78 | 0.91 | 0.58 |
| T3 | 0.58 | 0.57 | X | 0.86 | 0.71 | 0.71 | 0.87 | 0.70 | 0.63 | 0.60 |
| T4 | 0.75 | 0.62 | 0.62 | X | 0.87 | 0.89 | 0.83 | 0.84 | 0.81 | 0.60 |
| T5 | 0.54 | 0.53 | 0.96 | 0.65 | X | 0.91 | 0.84 | 0.79 | 0.97 | 0.57 |
| T6 | 0.88 | 0.66 | 0.66 | 0.83 | 0.67 | X | 0.75 | 0.86 | 0.90 | 0.70 |
| T7 | 0.54 | **0.40** | 0.65 | **0.45** | 0.59 | **0.40** | X | 0.72 | 0.76 | 0.60 |
| T8 | 0.77 | 0.53 | 0.73 | 0.76 | 0.74 | 0.78 | 0.74 | X | 0.78 | 0.54 |
| T9 | **0.46** | **0.38** | 0.53 | **0.35** | 0.56 | 0.50 | **0.46** | 0.71 | X | 0.58 |
| T10 | 0.61 | 0.35 | 0.70 | 0.65 | 0.78 | 0.57 | 0.63 | 0.77 | 0.51 | X |
| T11 | 0.50 | 0.72 | 0.57 | 0.72 | 0.65 | 0.66 | **0.20** | 0.50 | 0.50 | 0.55 |
| T12 | 0.71 | 0.80 | 0.77 | 0.85 | 0.73 | 0.81 | 0.51 | 0.68 | **0.39** | 0.56 |
| T13 | 0.69 | **0.46** | 0.78 | 0.68 | 0.74 | 0.63 | 0.60 | 0.64 | **0.45** | 0.71 |
| T14 | 0.88 | 0.58 | 0.72 | 0.65 | 0.68 | 0.79 | 0.66 | 0.80 | 0.51 | 0.76 |
| T15 | 0.68 | 0.60 | 0.84 | 0.83 | 0.83 | 0.76 | 0.59 | 0.86 | 0.51 | 0.71 |

| | T11 | T12 | T13 | T14 | T15 |
|---|---|---|---|---|---|
| T1 | 0.72 | 0.81 | 0.80 | 0.90 | 0.86 |
| T2 | 0.77 | 0.86 | 0.82 | 0.84 | 0.91 |
| T3 | 0.63 | 0.72 | 0.80 | 0.81 | 0.76 |
| T4 | 0.81 | 0.81 | 0.94 | 0.76 | 0.90 |
| T5 | 0.78 | 0.87 | 0.83 | 0.87 | 0.96 |
| T6 | 0.90 | 0.94 | 0.95 | 0.82 | 0.94 |
| T7 | 0.63 | 0.81 | 0.72 | 0.92 | 0.83 |
| T8 | 0.73 | 0.82 | 0.88 | 0.70 | 0.84 |
| T9 | 0.77 | 0.86 | 0.82 | 0.84 | 0.91 |
| T10 | 0.72 | 0.72 | 0.68 | 0.69 | 0.60 |
| T11 | X | 0.82 | 0.87 | 0.70 | 0.81 |
| T12 | 0.75 | X | 0.87 | 0.88 | 0.91 |
| T13 | 0.61 | 0.74 | X | 0.75 | 0.89 |
| T14 | 0.50 | 0.71 | 0.74 | X | 0.86 |
| T15 | 0.60 | 0.83 | 0.71 | 0.71 | X |

As seen from the table, the initial types are much more closely correlated than the final types, as is shown in the histogram (Fig 1).

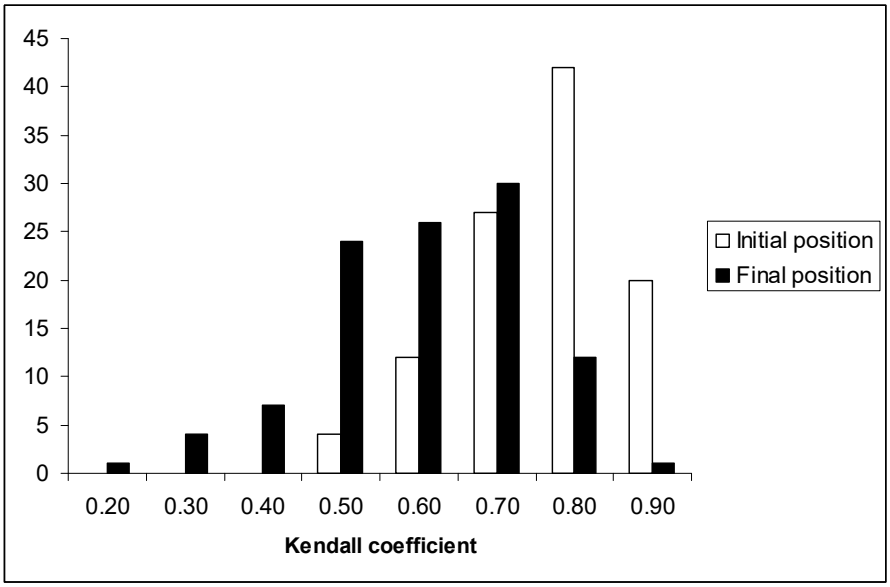


**Fig. 1.** Histogram of Kendall's correlation coefficients of ranked syllable types in the initial and final positions of 15 poems

Among others, T10 has the lowest coefficients, which can be explained by the fact that Voznesensky in his poem *RU* violated long poem's traditions and wrote it in the form of 10 chats with the file extension "ru" (the internet country code top-level domain for Russia) in the postmodernist manner and using a combination of different genres (mass culture, elite fiction, mixture of bookish lexis and obscene words, etc.)

In the final position, the lowest coefficients are observed again in Voznesensky's poem, but this time his *Rossiya voskrese* consists of a big number of psalms, written in literary colloquial style in combination with religious terms. On the contrary, T9, another text with low coefficients of the final types, is written in an unusually light and "easy" manner, with exact rhymes. Thus, both poems stand out from the rest in other aspects, too.

In general, it is possible to conclude that initial types are much more similar in the poems than final types.

Another possible way to examine the structure of a verse line is to find out to what extent it is symmetrical by matching concrete types of syllables at the beginning and the end of the line. In order to carry out this comparison, we used Busemann's coefficient (Zörnig, Altmann 2016; see also Místecký 2018):

$$B = \frac{I}{I + F},$$

where $I$ is the frequency of a given syllable type in the initial positions, and $F$ is the frequency of the same syllable type in the final position.

To estimate the significance, the chi-square test is used (Altmann, Köhler 2015):

$$\chi^2 = \frac{(I - F)^2}{I + F} .$$

The results are interpreted according to the scheme in Zörnig, Altmann (2016). With 1 degree of freedom and $\chi^2 \leq 0.05$ (3.84), the interpretation is as follows:

(1) SIP – initial position is significantly predominant ($B > 0.55$, $\chi^2 > 3.84$);
(2) IP – initial position has a tendency to predominance ($B > 0.55$, $\chi^2 < 3.84$);
(3) BL – the line is balanced (symmetrical) [$0.45 \leq B \leq 0.55$];
(4) FP – final position has a tendency to predominance ($B < 0.45$, $\chi^2 < 3.84$);
(5) SFP – final position is significantly predominant ($B < 0.45$, $\chi^2 > 3.84$).

Following this scheme, we obtained the following results and interpretations for T1 (Table 6).

**Table 6**
Relationship of types in the two positions of the lines in T1

| Syllable type | Frequency I | Frequency F | Busemann coefficient | Chi-square | Line structure |
|---|---|---|---|---|---|
| V | 18 | 0 | 1.00 | 18.00 | SIP |
| CV | 68 | 69 | 0.50 | 0.01 | BL |
| CCV | 18 | 3 | 0.86 | 10.71 | SIP |
| CVC | 43 | 75 | 0.36 | 8.68 | SFP |
| CCVC | 11 | 12 | 0.48 | 0.04 | BL |
| CVCC | 0 | 5 | 0.00 | 5.00 | SFP |
| CCCV | 1 | 3 | 0.25 | 1.00 | FP |
| VC | 8 | 0 | 1.00 | 8.00 | SIP |
| CCCVC | 1 | 1 | 0.50 | 0.00 | BL |

In Table 7, only cases of significant predominance (SIP or SFP) in all 15 poems is shown.

**Table 7**
Busemann coefficient and interpretation

| Texts | Type | Frequency I | Frequency F | Line structure |
|---|---|---|---|---|
| T1 | V | 18 | 0 | SIP |
| | CCV | 18 | 3 | SIP |
| | CVC | 43 | 75 | SFP |
| | CVCC | 0 | 5 | SFP |
| | VC | 8 | 0 | SIP |
| T2 | V | 33 | 1 | SIP |
| | CV | 45 | 74 | SFP |
| | CCV | 20 | 0 | SIP |
| | CVC | 30 | 72 | SFP |
| | VC | 26 | 3 | SIP |

| | | | | |
|---|---|---|---|---|
| T3 | V | 23 | 2 | SIP |
| | CV | 75 | 102 | SFP |
| | CCV | 31 | 6 | SIP |
| | CVC | 29 | 51 | SFP |
| T4 | V | 41 | 2 | SIP |
| | CCV | 28 | 4 | SIP |
| | CVC | 25 | 74 | SFP |
| | CVCC | 2 | 11 | SFP |
| T5 | V | 42 | 1 | SIP |
| | CCV | 15 | 6 | SIP |
| | CVC | 23 | 57 | SFP |
| T6 | V | 32 | 0 | SIP |
| | CCV | 16 | 3 | SIP |
| | CVC | 23 | 68 | SFP |
| T7 | V | 22 | 3 | SIP |
| | CV | 69 | 144 | SFP |
| | CCV | 41 | 14 | SIP |
| | CVC | 35 | 20 | SIP |
| | VC | 10 | 0 | SIP |
| T8 | V | 50 | 0 | SIP |
| | CCV | 28 | 1 | SIP |
| | CVC | 22 | 108 | SFP |
| | CCVC | 2 | 16 | SFP |
| | VC | 26 | 0 | SIP |
| T9 | V | 27 | 0 | SIP |
| | CCV | 24 | 1 | SIP |
| | CVC | 25 | 64 | SFP |
| | CCVC | 4 | 13 | SFP |
| | CCCV | 5 | 0 | SIP |
| | VC | 9 | 2 | SIP |
| T10 | V | 11 | 1 | SIP |
| | CV | 67 | 111 | SFP |
| | CCV | 20 | 4 | SIP |
| | CVC | 28 | 48 | SFP |
| | CCVC | 16 | 4 | SIP |
| | CVCC | 13 | 2 | SIP |
| | VC | 12 | 0 | SIP |
| T11 | V | 20 | 0 | SIP |
| | CCV | 13 | 0 | SIP |
| | CVC | 15 | 35 | SFP |
| T12 | V | 27 | 1 | SIP |
| | CV | 38 | 83 | SFP |
| | CCV | 21 | 2 | SIP |
| | VC | 14 | 2 | SIP |
| T13 | V | 28 | 3 | SIP |
| | CCV | 24 | 7 | SIP |
| | CVC | 24 | 57 | SFP |

| | | | | |
|---|---|---|---|---|
| | CVCC | 2 | 10 | SFP |
| | CCCVC | 0 | 6 | SFP |
| T14 | V | 24 | 0 | SIP |
| | CV | 56 | 104 | SFP |
| | CCV | 26 | 4 | SIP |
| | VC | 13 | 0 | SIP |
| T15 | V | 34 | 1 | SIP |
| | CCV | 24 | 6 | SIP |
| | CVC | 26 | 66 | SFP |
| | VC | 14 | 1 | SIP |
| Total | | 1655 | 1528 | |

To test the significance of the difference of initial and final frequencies, the chi-square test is used (Kelih et al. 2019):

$$\chi^2 = \sum_{i,j} \frac{(n_{iI} - n_{iF})^2}{n_{iI} + n_{iF}}$$

Inserting the data from T1 (Table 7) into the formula, we obtain the following result:

$$\chi^2 = \frac{(18-0)^2}{18+0} + \frac{(18-3)^2}{18+3} + \frac{(43-75)^2}{43+75} + \frac{(0-5)^2}{0+5} + \frac{(8-0)^2}{8+0} = \frac{324}{18} + \frac{225}{21} + \frac{1024}{118} + \frac{25}{5} + \frac{64}{8} = 50.392$$

For $p < 0.05$ and $df = 4$ the chi-square is significant.

Using the chi-square test for all the data in Table 7, we obtain $\chi^2 = 1070.146$ with $df = \infty$, which is highly significant.

The initial position predominates over the final position in 43 cases, and the final position appears only 24 times. As seen from the table, in most cases, V, CCV, VC and CV, CVC form a strong opposition, which leads to asymmetry of the line. The first three types have a tendency to take the initial positions and the last two the final ones.

Both CV and CVC are the most frequent types in the poems in general, including positions in the middle of the line. This may be the reason of their predominance in the final position which, as mentioned earlier, helps to unite different lines. The higher the frequency of the final types is, the stronger the links between the lines are.

CCV, V, VC, despite having high frequency ranks (3, 4, and 6, respectively), are several times less frequent than CV and CVC. Nevertheless, they are present in nearly all poems in the initial position, making, as stated above, a strong opposition to the final types.

On the whole, the results of the analysis, carried out in the study of syllabic arrangement, demonstrate certain regularity in the distribution of types of syllables in poems. This distribution did not change over time and does not depend on the genre, or an author's individual style. This coincides with the results of the study of syllabic distribution in Russian sonnets (Andreev 2018).

Syllabic similarity of different texts, judging by the initial and final positions in them, is rather strong. Syllable types in the initial and final positions have a tendency to form an opposition, which leads to syllabic asymmetry of the poetic line.

It must be emphasized that the obtained results should be regarded as preliminary findings and need to be checked further on a bigger material.

## References

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

**Andreev, S.** (2018). Distribution of Syllables in Russian Sonnets. *Glottometrics*, 41, 13–23.

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

**Beliy, A.** (2010). *Sobranije sochinenij. Simvolism*. [Collection of works. Symbolism]. Moskva: Kulturnaja revolutsija.

**Gasparov, M. L.** (2004). Ritmiko-syntaksicheskije kishe i formuly v epiloge «Ruslan i Ludmila». [Rythmic-syntactic clichés and formulas in the epilogue of "Ruslan and Ludmila"]. In: Gasparov, M. L., Skulacheva, T. V. (eds.) *Slavjanskij stih. VII. Lingvistika i struktura stiha*. [Slavonic Verse. VII. Linguistic Aspects and Structure of Verse]. Moskva: Yaziki slavyanskoy kul'turi, 149–166.

**Gasparov, M. L., Skulacheva, T. V.** (2004). *Statji o lingvistike stiha*. [Articles on the Linguistics of Verse]. Moskva: Yaziki slavyanskoy kul'turi.

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

**Krasnoperova, M. A.** (2000). *Osnovi rekonstuktivnogo modelirovanija stihoslozheniya: Na materiale ritmiki russkogo stiha.* [Basics of Reconstructive Modelling of Versification: On the Material of Russian Verse]. Saint-Petersburg: SPb universitet.

**Místecký, M.** (2018). Counting Stylometric Properties of Sonnets: A Case Study of Machar's Letní sonety. *Glottometrics*, 41, 1–12.

**Sheskin, D.J.** (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*: *Third Edition.* London / New York / Washington, D. C.: CRC Press.

**Zörnig, P., Altmann, G.** (2016). Activity in Italian Presidential Speeches. *Glottometrics*, 35, 38–48.

**Zörnig, P.** et al. (2019). *Quantitative Inquiries into the Syllable Structure.* [In print].

## Appendix

T1. J. Brodsky *Zofia*
T2. J. Brodsky *Felix*
T3. R. Rozhdestvensky *Poema o raznyh tochkah zreniya*
T4. Y. Yevtushenko *Bratskaya GES*
T5. Y. Yevtushenko *Pushkinskij pereval*
T6. R. Dyshalenkova *Begu po cementu*
T7. A. Voznesensky *Rossiya voskrese*
T8. F. Grimberg *Andrej Ivanovich vozvrashchaetsya domoj*
T9. S. Kekova *Po obe storony imeni*
T10. A. Voznesensky *RU*
T11. A. Parschikov *Neft'*
T12. V. Yemelin *Pechen'*
T13. V. Yemelin *Poema truby*
T14. M. Stepanova *Proza Ivana Sidorova*
T15. A. Kalinina *Peterburggo*

*Glottometrics 46, 2019, 41–60*

# Syllable Structure in Romani: A Statistical Investigation

*Anna Rácová[1], Peter Zörnig[2], Gabriel Altmann*

**Abstract.** We present some methods to analyze tendencies that can be discovered in the syllable structure. To this end we study regularities in the Romani language as spoken in Slovakia**.** The results may be useful to classify languages and to support a future theory.

***Keywords:*** *Romani, syllable, syllable types, syllable length, syllable distances*

## 1. Introduction

In any kind of theoretical research, we strive for finding some regularities leading to a description as simple as possible. In linguistics, there are two kinds of regularities: rules and laws. Rules, e.g. grammatical rules, do not explain anything and are not explainable. Hence they can be omitted in theoretical research. The other regularities which can be expressed (characterized, measured, modelled) by some numbers, formulas, etc. are candidates for laws. Laws are statements derived from a background theory and sufficiently tested and confirmed by data. As well known from physics, this way has no end because any regularity is (at least partially) influenced by some other regularities or properties. Nevertheless, quite frequently one can find individual phenomena following a mathematically expressible regularity. At the beginning, one must try to test the given hypothesis using one sole language and in case of positive testing one may extend the research to other languages.

In many languages, the syllable has been "defined" in some way. The authors mostly speak of undecidable cases, of theory, etc. It is well known that in written text, every "rule" is a compromise, not necessarily present in the spoken text. When speaking, we do not make visual boundaries between sentences, not even between words, we need not make pauses. There are many works which try to solve the problem of syllables in different languages (e.g. Best 2013; Bičan 2015; Cutler, Carter 1987; Ivanecký, Majchráková 2007; Jones 1971; Kar 2010; Kelih, Mačutek 2013; Lee 1986; Narayana, Ramakrishnan (s.d.); Ohala (s.d.); Sabol 1994), for example, whether the word "texts" consist of one or two syllables in English, is the Italian "i" in diphthongs a vowel, how to define syllabic segmentation in Slovak, how many diphthongs are in Bangla, problem of syllabification of intervocalic consonant clusters in Hindi, etc. The only possible decision is given by a mathematical model: that segmentation is better whose results follow a law. But the law must be given quantitatively, it must be derived

[1] Anna Rácová ,Institute of Oriental Studies, Slovak Academy of Sciences, Bratislava, e-mail: racova.anna@gmail.com.
[2] Peter Zönig, Dept. of Statistics, University of Brasilia, e-mail: peter@unb.br.

and the hypothesis must be tested in many cases. As is well known, things simply exist but we (humans) say what they are, how they are, and how do they behave.

Here we shall apply some models to the Romani data. It is to be remarked that this language has never been studied quantitatively though there are scientific books on the Romani language in general and on various Romani dialects in particular (e.g. Boretzky 1999, 2003; Elšik, Matras 2006; Halwachs, Mentz 1999; Matras, Bakker, Kyuchukov 1997; Matras 2002; Soravia 1977), as well as numerous articles, grammars, and dictionaries, etc.

We shall study the structure of syllables in Romani texts, namely the frequency order of syllable types, their length in terms of phoneme numbers and the dichotomic classification into open and closed syllables. Open syllables end with a vowel, closed ones with a consonant. Further, we shall compare the texts and study the distances between equal syllables occurring in a text. Needless to say, it will be the Romani as it is spoken in Slovakia, namely the Slovak and Hungarian varieties of Romani which belong to the Central group of Romani dialects.

Like other Romani dialects in Europe, the Romani language in Slovakia descends from a common ancestor, Early Romani, which was spoken from the tenth or eleventh century and up till the late fourteenth century in Byzantium. After the Roma left the Byzantine Empire in the late fourteenth century and dispersed throughout Europe, their language gradually lost contacts with other Romani dialects and developed under the massive influence of different European languages (Elšik, Matras 2006: 5).

The first record about Roma in the Slovak area appeared in 1322 in Spiš, but Roma were only settled in present Slovakia in the 18th century due to the policy of the Enlightenment of Maria Theresa and Joseph II who attempted to regulate Gypsies and strived for their permanent settlement and their joining in economic activities. These are so-called Slovak and Hungarian Roma today, who differ mainly by the degree of influence of Slovak and Hungarian on their language. The third group of Roma living in Slovakia are so-called Vlach Roma, who were only forced to settle down in 1958-1959 and up to the present day preserved the variety of language belonging to the Vlach group of Romani.

There are no exact data about the number of Roma and the speakers of Romani in Slovakia. In the 2011 population and housing census in the Slovak Republic, 105,738 inhabitants of Slovakia identified themselves as Roma, however, qualified estimates of the number of Roma in the Slovak Republic are significantly higher. Mušinka et al. (2014) estimate that there are 402,840 Roma in Slovakia.

Neither the number of Roma who declared themselves as speakers of Romani as their mother tongue in the 2011 Census (122,518 people) matches reality. When estimating the number of Romani-speaking population the degree of integration of the Roma in majority society is significant. The more segregated the Roma population is, the more often they use Romani and the less often they speak Slovak (Hungarian).

Based on the analysis by Rácová and Samko (2017), we assume that the majority of Roma who live in Roma settlements on the outskirts of municipalities, in Roma settlements inside municipalities and in segregated settlements (overall 53.9% of Roma in Slovakia, Mušinka et al. 2014) speak Romani. Adult Roma, who speak Romani, are bilingual or even trilingual. Besides Romani they speak Slovak, which is the language of the majority population and the official language used in the state institutions, and also Hungarian in areas inhabited by the Hungarian minority.

Almost half of the Roma in Slovak Republic (46.5%) live scattered among majority population (Mušinka et al. 2014: 6). Many of them gave up Romani many years ago. We estimate that of the total number of these Roma only 10% speak Romani, the rest of them speak Slovak or Hungarian.

As a result, the total number of Romani-speaking Roma in Slovakia is estimated to be 236,166, which accounts for 58.1% of the total Roma population in Slovakia.

The majority of Roma who speak that language in Slovakia speak Slovak Romani. The East Slovak dialect of this variety is the most thoroughly explored and described and it also presents the basis for the standardization of the Romani language in the Slovak Republic in 2008.

Our field research conducted in 2015 as well as the long term knowledge of the environments in which the Roma in Slovakia live and work (Rácová, Samko 2017) proves the constant decrease of the number of Romani speakers in Slovakia. Romani is mainly spoken by the Roma with a lower social and educational background, especially by those living in concentrated settlements. In some cases the Roma cease to pass their language to their children even when they speak Romani at home because the Romani language does not enable them to assert themselves in the majority population. Frequent contacts with the majority population, higher education which is provided in the state language, and mixed marriages contribute to this situation.

Even the changed status of the Romani language after 1991, when the Roma were officially recognized as a national minority in the Slovak Republic and Romani obtained all the rights pertaining to national minority languages in Slovakia, did not stop the Roma from abandoning Romani. Of real help are neither strives of the cultural Romani elite who pursue the idea of Romani as a fully functioning language, equal to all languages which can be used in all spheres of life and sometime also as a marker of Romani identity.

A UNESCO commission (Brenzinger et al. 2003) classified Romani in Slovakia as ´definitely endangered´ and our research confirms this classification (Rácová, Samko 2017). The Romani language is not transmitted from generation to generation in the entire Roma population, the number of Romani-speaking Roma is insufficient, and the proportion of Romani-speaking Roma to the total number of members of the Roma national minority living in Slovakia is on a constant decrease. Although the changed status of the Roma and their language after 1991 has contributed to an increase in the number of the domains in which Romani is used (literature, theatre, media and it is taught as a subject at several private schools), it has not penetrated into the awareness of most Roma. Neither are the development and preservation of Romani sufficiently supported by the state, which focuses mainly on social problems of the Roma population.

The Romani language as it is spoken in Slovakia shares a base lexical core composed of various historical layers common for other Romani dialects: Indic, Iranian, Armenian, Greek (Elšik, Matras, Kyuchukov 1997). There are also a few borrowings from Serbian, and Romanian, relatively frequent loan-words from Hungarian, and numerous borrowings from Slovak. Borrowings from Slovak often undergo the orthoepic, orthographic, morphematic and morphological adaptation. Neologisms and internationalisms are other sources of enrichment of the Romani lexicon.

The influence of Slovak is also reflected on the grammar of Romani. Some changes caused by interference are systemic (for instance the way of expressing the category of mode and aspect by a system of Slovak verbal prefixes, expressing the modality of necessity, the generalization of use of the reflexive pronoun *pes*, and forming of negative pronouns by the prefix *ňi*), others are more or less sporadic (for instance the concord of adjectives, pronouns and numerals in case with nouns) (see more Rácová 2015a).

Some Romani syntactic constructions are also copying the structure of Slovak.

The Romani orthography is based on the writing system of Slovak. The phoneme inventory consists of 31 consonants (*b, c, č, čh, d, ď, dz, dž, f, g, h, ch, j, k, kh, l, ľ, m, n, ň, p, ph, r, s, š, t, ť, th, v, z, ž*) and 5 vowels (*a, e, i, o, u*). There are no diphtongs in Romani. Slovak diphtongs in borrowings from the Slovak language are adapted to Romani, e.g. Slovak

phoneme *i* in diphtongs *ia, ie* is substituted by *j*: Slovak *žiadateľ* (CV, CV, CVC) becomes Romani *žjadateľis* (CCV, CV, CV, CVC), Slovak *podmieniť* (CVC, CV, CVC) becomes Romani *podmjeňinel* (CVC, CCV, CV, CVC).

## 2. Syllable types

For the study of syllabic structure we used Romani as it is used in published texts. They are translations from the parallel Slovak texts to Romani by Roma.

We included texts of different genres: poetry (*O phuvakero, Valakana*), narratives (*Hanka, Johanka, Holokaust*), a fairy tale (*O Baris*), short stories (*Betmen, Romipen*), explanatory notes to the Census forms 2011 (*Census*), text of the Declaration of Roma of the Slovak Republic to standardization of the Romani language in the Slovak Republic in 2008 (*Deklaracija*), an interview with a Slovak politician (*Interview*) as well as an article published in the Romani newspaper *Romano nevo ľil* (*RNĽ*), and an introduction to the Slovak-Romani dictionary of technical words (*Angluno*) (see Appendix).

As can be seen, the Slovak has more types of syllables than Romani but in the latter we found syllables consisting of 5 phonemes. The results are presented in Table 1. We consider the rank-order of types, i.e. we determine the most frequent syllable type, which has rank 1, the second most frequent type having rank 2, etc. These ordered frequencies have been captured by the Zipf-Alekseev function with added 1, i.e. the frequency of rank x is $y = cx^{a+b\ln(x)} + 1$. This function considers the requirements of the writers/speakers in logarithmic terms, hence the differential equation is

$$\frac{y`}{y-1} = \frac{a+b\ln x}{x} \qquad (1)$$

yielding the just mentioned result

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

Here, one can consider *a* as the constant of the language, *b ln x* is the changing influence of the speaker/writer and the parameter *c* representing the normalization constant is simply an equilibrating factor. In the first case we compare the Slovak text with the Romani one.

**Table 1**
Syllable types of Slovak and Romani
*Rol`ník* : *O phuvakero*

| Rank | Slovak Types | Frequ. | Computed | Romani types | Frequ | Computed |
|---|---|---|---|---|---|---|
| | CV | 154 | 153.86 | CV | 205 | 205.04 |
| | CVC | 41 | 43.18 | CVC | 73 | 72.36 |
| | CCV | 24 | 19.03 | V | 22 | 24.70 |
| | CCVC | 11 | 10.45 | VC | 12 | 9.72 |
| | VC | 4 | 6.58 | CCV | 7 | 4.54 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | CCC | 3 | 4.57 | CVCC | 2 | 2.56 |
| | CCCV | 3 | 3.42 | CCVC | 1 | 1.74 |
| | CC | 2 | 2.72 | CCCVC | 1 | 1.37 |
| | CVCC | 2 | 2.26 | | | |
| | V | 1 | 1.95 | | | |
| | CCCC | 1 | 1.73 | | | |
| | a = -1.7068, b = -0.2174<br>c = 152.8608, $R^2$ = 0.9980 | | | a = -0.7567, b = -1.0949,<br>c = 204.0425, $R^2$ = 0.9994 | | |

The fitting of the Zipf-Alekseev function to both data sets is excellent ($R^2 > 0.99$). The goodness-of-fit is measured by the coefficient of determination $R^2$ which must be larger than 0.8. The parameter *c* of the formula represents the frequency at rank 1 and the parameters *a* and *b* do not differ considerably from those obtained from other languages. In the tables below we present several Romani texts.

**Table 2**
Syllable types in the Romani text *Hanka* and *Betmen*

| | **Hanka** | | | **Betmen** | | |
|---|---|---|---|---|---|---|
| Rank | Types | Frequency | Zipf-Alekssev +1 | Types | Frequency | Zipf-Alekssev +1 |
| 1 | CV | 681 | 681.17 | CV | 686 | 686.31 |
| 2 | CVC | 364 | 362.39 | CVC | 364 | 360.97 |
| 3 | V | 61 | 72.25 | V | 82 | 97.23 |
| 4 | CCV | 34 | 13.84 | CCV | 43 | 25.75 |
| 5 | VC | 32 | 3.47 | VC | 30 | 7.79 |
| 6 | CVCC | 6 | 1.52 | CVCC | 8 | 3.01 |
| 7 | CCC | 1 | 1.12 | CCVC | 2 | 1.65 |
| 8 | CCVC | 1 | 1.03 | CCCV | 1 | 1.22 |
| 9 | CCCVC | 1 | 1.01 | | | |
| | a = 1.0386, b = -2.8147,<br>c = 680.1667, $R^2$ = 0.9969 | | | a = 0.5380 , b = -2.1162,<br>c = 685.3133, $R^2$ = 0.9975 | | |

**Table 3**
Syllable types in the Romani texts *O Roma* and *Romipen*

| | **O Roma** | | | **Romipen** | | |
|---|---|---|---|---|---|---|
| Rank | Types | Frequency | Zipf-Alekssev +1 | Types | Frequency | Zipf-Alekssev +1 |
| 1 | CV | 358 | 358.31 | CV | 458 | 458.08 |
| 2 | CVC | 172 | 168.25 | CVC | 208 | 207.05 |
| 3 | V | 67 | 78.80 | V | 31 | 38.56 |
| 4 | CCV | 46 | 40.14 | CCV | 25 | 7.46 |
| 5 | VC | 34 | 22.16 | VC | 12 | 2.20 |
| 6 | CVCC | 9 | 13.14 | CCVC | 2 | 1.25 |
| 7 | CCVC | 8 | 8.31 | CVCC | 1 | 1.06 |
| 8 | VCC | 2 | 5.58 | | | |
| 9 | CCCV | 1 | 3.97 | | | |
| | a = -0.5953, b = -0.7213,<br>c = 357.3139, $R^2$ = 0.9967 | | | a = 0.7739, b = -2.7748,<br>c = 457.0811 $R^2$ = 0.9974 | | |

**Table 4**
Syllable types in the Romani text *Deklaracija* and *Johanka*

| | **Deklaracija** | | | **Johanka** | | |
|---|---|---|---|---|---|---|
| Rank | Types | Frequency | Zipf-Alekssev +1 | Types | Frequency | Zipf-Alekssev +1 |
| 1 | CV | 656 | 656.15 | CV | 652 | 652.12 |
| 2 | CVC | 231 | 228.80 | CVC | 360 | 358,65 |
| 3 | V | 36 | 49.99 | V | 95 | 100.99 |
| 4 | VC | 32 | 11.91 | VC | 32 | 27.77 |
| 5 | CCV | 26 | 3.69 | CCV | 20 | 8.60 |
| 6 | CCVC | 5 | 1.74 | CCVC | 9 | 3.33 |
| 7 | CVCC | 4 | 1.22 | CVCC | 4 | 1.77 |
| 8 | | | | CCCVC | 2 | 1.27 |
| | a = -0.0943 , b = -2.0627, c = 655.1489, $R^2$ = 0.9968 | | | a = 0.5734, b = -2.9743, c = 651.1511, $R^2$ = 0.9994 | | |

**Table 5**
Syllable types in the Romani text *Valakana* and *Interview*

| | **Valakana** | | | **Interview** | | |
|---|---|---|---|---|---|---|
| Rank | Types | Frequency | Zipf-Alekseev +1 | Types | Frequency | Zipf-Alekseev +1 |
| 1 | CV | 173 | 173.00 | CV | 407 | 407.28 |
| 2 | CVC | 47 | 47.01 | CVC | 198 | 195.32 |
| 3 | V | 14 | 14.26 | V | 58 | 66.76 |
| 4 | VC | 7 | 5.44 | CCV | 25 | 23.775 |
| 5 | CCV | 1 | 2.68 | VC | 21 | 9.47 |
| 6 | CCC | 1 | 1.70 | CCVC | 14 | 4.38 |
| 7 | | | | CVCC | 8 | 2.45 |
| 8 | | | | CCCV | 3 | 1.65 |
| 9 | | | | VCC | 1 | 1.31 |
| 10 | | | | CC | 1 | 1.15 |
| | a = -1.1661, b = -1.0630, c = 171.9976, $R^2$ = 0.9997 | | | a = -0.0495, b = -1.4637, c = 406.2766, $R^2$ = 0.9978 | | |

**Table 6**
Syllable types in the Romani text *Census* and *O Baris*

| | **Census** | | | **O Baris** | | |
|---|---|---|---|---|---|---|
| Rank | Types | Frequency | Zipf-Alekssev +1 | Types | Frequency | Zipf-Alekssev +1 |
| 1 | CV | 599 | 599.26 | CV | 539 | 539.12 |
| 2 | CVC | 256 | 252.90 | CVC | 289 | 287.76 |
| 3 | V | 98 | 106.93 | V | 105 | 109.65 |
| 4 | CCV | 48 | 49.76 | CCV | 45 | 42.68 |
| 5 | VC | 47 | 25.36 | VC | 29 | 17.99 |
| 6 | CVCC | 8 | 14.01 | CVCC | 1 | 8.39 |
| 7 | CCVC | 4 | 8.35 | CCVC | 1 | 4.40 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 8 | CCCVC | 2 | 5.34 | CCCV | 1 | 2.65 |
| | a = -0.6874, b = -0.8087, c = 598.2556, , $R^2$ = 0.9980 | | | a = 0.0291, b = -1.3521, c = 538.1203, $R^2$ = 0.9992 | | |

**Table 7**
Syllable types in the Romani text *Holokaust* and *RNĽ*

| | **Holokaust** | | | **RNĽ** | | |
|---|---|---|---|---|---|---|
| Rank | Types | Frequency | Zipf-Alekssev +1 | Types | Frequency | Zipf-Alekssev +1 |
| 1 | CV | 474 | 474.16 | CV | 497 | 496.80 |
| 2 | CVC | 215 | 213.24 | CVC | 196 | 196.80 |
| 3 | V | 65 | 71.44 | CCV | 90 | 94.16 |
| 4 | CCV | 27 | 25.25 | V | 57 | 51.30 |
| 5 | VC | 24 | 10.02 | VC | 53 | 30.63 |
| 6 | CVCC | 2 | 4.62 | CCVC | 7 | 19.61 |
| 7 | CC | 1 | 2.55 | CCCV | 1 | 13.27 |
| 8 | | | | CCCVC | 1 | 9.40 |
| | a = -0.1701, b = -1.4232, c = 473.1602, $R^2$ = 0.9986 | | | a = -1.0301, b = -0.4476, c = 485.7960, $R^2$ = 0.9954 | | |

**Table 8**
Syllable types in the Romani text *Angluno*

| | **Angluno** | | |
|---|---|---|---|
| Rank | Types | Frequency | Zipf-Alekssev +1 |
| 1 | CV | 570 | 570.51 |
| 2 | CVC | 244 | 237.80 |
| 3 | VC | 52 | 77.36 |
| 4 | CCV | 47 | 26.93 |
| 5 | V | 34 | 10.58 |
| 6 | CCVC | 8 | 4.83 |
| 7 | CCCV | 4 | 2.64 |
| 8 | CCCVC | 3 | 1.76 |
| 9 | CVCC | 1 | 1.36 |
| | a = -0.3037, b = -1.3884, c = 569.5106, $R^2$ = 0.9943 | | |

As can be seen in the tables, the main syllables of Romani are CV, CVC and V. A very broad investigation would be necessary to state whether this relation exists also in other Romani dialects. Here, we merely present the problem. Some of the syllable types are taken from Slovak but they do not disturb the general view.

A distribution can be characterized in many ways. One can take the moments, Ord's criterion, entropy, different text richness indicators, etc. The literature concerning these properties is enormous. Here, we have to do with syllables, that is, with purely non-semantic

entities, hence we can characterize the texts only with properties resulting from formal frequencies.

One can compare the individual texts and look whether there is some difference in the use of syllables. To this end one can compare either pairs of texts or all texts using a chi-square test. Here, we suppose that the general difference will be too large, hence we restrict ourselves to the study of a non-parametric comparison. We let the ranks as they are but if two types have the same frequency, we take the mean rank for both.

**Kendall's rank correlation test**

In the previous section we counted the frequencies of the syllable types appearing in diverse texts. Now we shall test whether the Romani texts in Tables 1 to 8 tend to agree in the *rank distribution of syllable types*. Here the most frequent syllable type has rank 1, the second most type has rank 2 etc. Perfect concordance between two texts in this sense means that all their ranks agree. "High concordance" between two texts T1 and T2 with respect to this criterion means that the "frequent" syllable types of T1 are "frequent" types of T2 and "rare" types of T1 are "rare" types of T2. Finally, discordance means that frequent types of T1 are rare types of T2 and vice versa. We present a statistical test due to Kendall that measures the degree of concordance in the given sense between two or more texts. It is out of the scope of the present article to justify the test procedure. The statistically interested reader finds the pertinent theory in Bortz et al. (1990: 465-470) and the literature cited there. Here we restrict ourselves to describe the test procedure in a way that can be easily transferred to similar linguistic contexts. In Table 9 we present the rank distributions of 12 syllable types in 14 Romani texts.

**Table 9**
Kendall´s test for the Romani texts

| | CV | CVC | V | CCV | VC | CCVC | CVCC | CCCV | VCC | CC | CCCVC | CCC | $\mathbf{V_i}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Declaracija | 1 | 2 | 3 | 5 | 4 | 6 | 7 | 10 | 10 | 10 | 10 | 10 | **120** |
| Romipen | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 10 | 10 | 10 | 10 | 10 | **120** |
| O phuvakero | 1 | 2 | 3 | 5 | 4 | 7.5 | 6 | 10.5 | 10.5 | 10.5 | 7.5 | 10.5 | **66** |
| Hanka | 1 | 2 | 3 | 4 | 5 | 8 | 6 | 11 | 11 | 11 | 8 | 8 | **48** |
| Valanka | 1 | 2 | 3 | 5.5 | 4 | 9.5 | 9.5 | 9.5 | 9.5 | 9.5 | 9.5 | 5.5 | **216** |
| Betmen | 1 | 2 | 3 | 4 | 5 | 7 | 6 | 8 | 10.5 | 10.5 | 10.5 | 10.5 | **60** |
| RNL | 1 | 2 | 4 | 3 | 5 | 6 | 10.5 | 7 | 10.5 | 10.5 | 8 | 10.5 | **66** |
| Angluno | 1 | 2 | 5 | 4 | 3 | 6 | 9 | 7 | 11 | 11 | 8 | 11 | **24** |
| O Roma | 1 | 2 | 3 | 4 | 5 | 7 | 6 | 9 | 8 | 11 | 11 | 11 | **24** |
| Johanka | 1 | 2 | 3 | 5 | 4 | 6 | 7 | 10.5 | 10.5 | 10.5 | 8 | 10.5 | **60** |
| Interview | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9.5 | 9.5 | 11.5 | 11.5 | **12** |
| Census | 1 | 2 | 3 | 4 | 5 | 7 | 6 | 10.5 | 10.5 | 10.5 | 8 | 10.5 | **60** |
| O Baris | 1 | 2 | 3 | 4 | 5 | 7 | 6 | 8 | 10.5 | 10.5 | 10.5 | 10.5 | **60** |
| Holocaust | 1 | 2 | 3 | 4 | 5 | 6 | 10 | 10 | 10 | 7 | 10 | 10 | **120** |
| **Rank sums** $T_j$ | **14** | **28** | **45** | **59.5** | **64** | **95** | **103** | **129** | **142** | **142** | **130.5** | **140** | |

The test statistic is given by

$$X_r^2 = \frac{12\,RSD}{m\,N(N+1) - \frac{1}{N-1}\sum_{i=1}^{m} V_i}, \qquad (3)$$

where N is the number of syllable types, m the number of texts, $T_i$ denotes the rank sum of the i-th column, $\overline{T} = \frac{1}{N}\sum_{i=1}^{N} T_i$ and

$$RSD = \sum_{i=1}^{N} (T_i - \overline{T})^2 \qquad (4)$$

is the rank sum deviation. Finally, the sum in the denominator of (3) can be interpreted as a kind of "tie correction" applied in the case of identical ranks. Under the null hypothesis $H_0$ that there is ***no*** *concordance in the ranks of the texts*, the statistic (3) is chi square distributed with N-1 degrees of freedom.

We shall now explain the calculations in detail. Considering e.g. the text "Valakana" in Table 5, we observe that CV is the most frequent syllable type, having therefore rank 1, CVC is the second-most frequent type, having rank 2, etc. Assigning to every syllable type its rank we get the following ranks for the given syllable types:

| CV | CVC | V | CCV | VC | CCVC | CVCC | CCCV | VCC | CC | CCCVC | CCC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 5 | 4 | 7 | 8 | 9 | 10 | 11 | 12 | 6 |

A look at Table 5 shows that the ranks 5 and 6 correspond to equal frequencies of value 1. So we substitute the ranks 5 and 6 in the above table by their mean value (5+6)/2=5.5. Moreover, the ranks 7 to 12 correspond to the same frequency of value 0 (in fact these syllable types do not occur in the text "Valakana"). In analogy to the previous case we substitute the ranks 7, 8,…,12 by their mean value (7+8+…+12)/6 = 9.5. In this way originates the line in Table 9 corresponding to the text "Valakana". Next, the column sums $T_i$ of Table 9 are needed, which are already presented in the last line. By using formula (4) we obtain $\overline{T}$ = (14+28+…+140)/12 = 91 and $RSD = (14\text{-}91)^2 + \ldots + (140\text{-}91)^2 = 24501.5$. In order to compute (3) we still need the values of the $V_j$, which are given by

$$V_j = \sum_{k=1}^{s_j} (v_k^{\,3} - v_k)\,, \qquad (5)$$

where $s_j$ is the number of ties and the $v_k$ are the lengths of sequences of equal ranks. For example, in the text "Valakana", which is the text number 5 of Table 9, we obtain two ties of lengths 2 a nd 6, thus

$$V_5 = \sum_{k=1}^{2} (v_k^{\,3} - v_k) = (2^3 - 2) + (6^3 - 6) = 216\,. \qquad (6)$$

The numbers $V_j$ are presented in the last column of Table 9. We get $V_1 + \ldots + V_m = 1056$.

The value of the test statistic is therefore

$$X_r^2 = \frac{12 \cdot 24501.5}{14 \cdot 12 \cdot 13 - \frac{1}{11} 1056} = 140.8 . \qquad (7)$$

For N-1=11 degrees of freedom we obtain a probability of exceeding this value under the null hypothesis as $P(X_r^2 > 140.8) = P(\chi_{11}^2 > 140.8) \approx 10^{-24}$. This means that the null hypothesis of lack of concordance in the rank assignments must be rejected, i.e. there is a concordance in the rank distribution. The chance of an error in the test decision is practically zero. The result becomes plausible by looking at Table 5. All considered texts choose CV as the most frequent syllable type and CVC as the second most frequent type, and almost all texts use V as the third most frequent type. Variations occur only for higher ranks

## The parameters

If one considers the parameter *a* as the independent and *b* as the dependent variable in formula (2) one can easily state that they are linearly correlated, i.e. $b = k + ma$ where *k* and *m* are the parameters of the straight line. The relation is shown in Table 14

**Table 10**
The relation between the parameters *a* and *b* in the Zipf-Alekseev formula for Romani syllable types

| Text | a | b | b = -1.6901- 0.9499a |
|---|---|---|---|
| Valakana | -1.1661 | -1.0630 | -0.5824 |
| O phuvakero | -0.7567 | -1.0949 | -0.9713 |
| Census | -0.6874 | -0.8087 | -1.0371 |
| O Roma | -0.5953 | -0.7213 | -1.1246 |
| RNĽ | -0.3037 | -1.3884 | -1.4016 |
| Angluno | -0.3035 | -1.3884 | -1.4018 |
| Holokaust | -0.1701 | -1.4232 | -1.5285 |
| Deklaracija | -0.0943 | -2.0627 | -1.6005 |
| Interview | -0.0495 | -1.4637 | -1.6431 |
| O Baris | 0.0291 | -1.3521 | -1.7177 |
| Betmen | 0.5380 | -2.1162 | -2.2011 |
| Johanka | 0.5734 | -2.0743 | -2.2348 |
| Romipen | 0.7739 | -2.7748 | -2.4252 |
| Hanka | 1.0386 | -2.8147 | -2.6767 |
| | $R^2 = 0.8200$ | | |

The Piotrowski function yields $R^2 = 0.8587$ but we prefer the simpler function with two parameters.

We accept a determination coefficient which is greater than 0.8, however, a number of other texts must be examined in order to obtain a stronger relation and the result must be compared with that obtained from the analysis of other languages.

**Syllable length**

For the modeling of syllable length we use the usual Menzerathian function, defined as $y = ax^b exp(-cx)$. Here, we can see that the writer`s/speaker`s striving for change is given in simple form and the formula can be derived from the differential equation

$$\frac{y'}{y} = c + \frac{b}{x} \qquad (8)$$

following from the general theory (cf. Wimmer, Altmann 2005) just as the Zipf-Alekseev formula. Here *b* and *c* are the contributions of the writer and reader and *a* is the equilibrating constant. The results for Slovak and Romani concerning the text *Rol`ník : O phuvakero* are presented in Table 11

**Table 11**
Length of syllables in the Slovak and Romani texts:
*Rol'ník* : *O phuvakero*

| Length | Slovak | Computed | Romani | Computed |
|---|---|---|---|---|
| 1 | 1 | 5.65 | 22 | 21.40 |
| 2 | 160 | 159.27 | 217 | 217.26 |
| 3 | 68 | 70.20 | 80 | 79.03 |
| 4 | 17 | 5.64 | 3 | 7.36 |
| 5 | | | 1 | 0.32 |
| | a = 4500.4116, b = 14.4571 c = 6.6811, $R^2$ = 0.9899 | | a = 6416.2017, b = 11.5771, c = 5.7033, $R^2$ = 0.9994 | |

In the other texts we found the "length-regularity" as presented in Tables 12 to 17.

**Table 12**
Syllable length in the Romani text *Betmen* and *O Roma*

| | **Betmen** | | **O Roma** | |
|---|---|---|---|---|
| Length | Frequency | Computed | Frequency | Computed |
| 1 | 82 | 63.09 | 67 | 61.13 |
| 2 | 716 | 723.33 | 392 | 396.02 |
| 3 | 407 | 388.95 | 220 | 209.94 |
| 4 | 11 | 59.76 | 18 | 39.94 |
| | a = 8755.0885, b = 10.6357, c = 4.9328, $R^2$ = 0.9901 | | a = 3026.7713, b = 8.7009, c = 4.1626, $R^2$ = 0.9901 | |

**Table 13**
Syllable length in the Romani text *Hanka* and *Deklaracija*

| | **Hanka** | | **Deklaracija** | |
|---|---|---|---|---|
| Length | Frequency | Computed | Frequency | Computed |
| 1 | 61 | 39.45 | 36 | 33.91 |
| 2 | 713 | 718.25 | 688 | 688.46 |
| 3 | 399 | 386.07 | 257 | 255.37 |

| | | | | |
|---|---|---|---|---|
| 4<br>5 | 7<br>1 | 49.06<br>2.83 | 9 | 18.40 |
| | a = 10515.9844, b = 12.2447 ,<br>c = 5.5886, $R^2$ = 0.9938 | | a =25761.0315 , b = 13.9127,<br>c = 6.6328, $R^2$ = 0.9997 | |

**Table 14**
Syllable length in the Romani text *Johanka* and *Romipen*

| | **Johanka** | | **Romipen** | |
|---|---|---|---|---|
| Length | Frequency | Computed | Frequency | Computed |
| 1 | 95 | 80.48 | 31 | 22.20 |
| 2 | 684 | 691.49 | 470 | 471.84 |
| 3 | 380 | 361.06 | 233 | 227.962 |
| 4 | 13 | 59.90 | 3 | 23.39 |
| 5 | 2 | 5.30 | | |
| | a =7983.1570 , b = 9.7358,<br>c = 4.597 , $R^2$ = 0.9918 | | a = 9516.7664, b = 13.1531,<br>c = 6.0606, $R^2$ = 0.9963 | |

**Table 15**
Syllable length in the Romani text *Interview* and *Census*

| | **Interview** | | **Census** | |
|---|---|---|---|---|
| Length | Frequency | Computed | Frequency | Computed |
| 1 | 58 | 54.88 | 98 | 92.06 |
| 2 | 429 | 430.77 | 646 | 649.73 |
| 3 | 224 | 219.41 | 304 | 293.06 |
| 4 | 25 | 36.47 | 12 | 42.87 |
| | | | 2 | 3.38 |
| | a = 5088.0059, b = 9.5607,<br>c = 4.5295, $R^2$ = 0.9984 | | a = 9847.6848, b = 9.5602,<br>c = 4.6725, $R^2$ = 0.9962 | |

**Table 16**
Syllable length in the Romani text *O Baris* and *Holokaust*

| | **O Baris** | | **Holokaust** | |
|---|---|---|---|---|
| Length | Frequency | Computed | Frequency | Computed |
| 1 | 105 | 89.88 | 65 | 58.14 |
| 2 | 568 | 578.44 | 499 | 502.53 |
| 3 | 334 | 308.03 | 242 | 231.87 |
| 4 | 3 | 59.13 | 2 | 32.23 |
| | a = 5658.0825 b = 8.6623,<br>c = 4.1424, $R^2$ = 0.9781 | | a =,7834.5047, b = 10.1859,<br>c = 4.9035, $R^2$ = 0.9928 | |

**Table 17**
Syllable length in *RNĽ, Angluno and Valakana*

| | **RNL`** | | **Angluno** | | **Valakana** | |
|---|---|---|---|---|---|---|
| Length | Frequency | Computed | Frequency | Computed | Frequency | Computed |
| 1 | 56 | 46.09 | 34 | 28.38 | 14 | 14.00 |
| 2 | 550 | 553.66 | 622 | 623.14 | 180 | 180.00 |
| 3 | 286 | 276.30 | 291 | 287.73 | 48 | 48.00 |
| 4 | 8 | 37.49 | 13 | 27.34 | | |
| 5 | 1 | 2.49 | 3 | 1.09 | | |
| | a, 8176.1555, b = 11.0570, c = 5.1783, $R^2$ = 0.9952 | | a = 14204.3835, b = 13.4235, c = 6.2155, $R^2$ = 0.9991 | | a = 12372.6368, b = 13.4720, c = 6.7841, $R^2$ = 1.0000 | |

Again, if one orders the values according to increasing *b*, one obtains the results presented in Table 18.

**Table 18**
The values of b and c in the length function

| | O Baris | O Roma | Interview | Census | Johanka | Holokaust | Betmen |
|---|---|---|---|---|---|---|---|
| b | 8.66 | 8.70 | 9.56 | 9.57 | 9.74 | 10.18 | 10.64 |
| c | 4.14 | 4.16 | 4.53 | 4.67 | 4.60 | 4.90 | 4.93 |

| | RNĽ | O phuvakero | Hanka | Romipen | Angluno | Valakana | Deklaracija |
|---|---|---|---|---|---|---|---|
| b | 11.06 | 11.58 | 12.24 | 13.15 | 13.42 | 13.47 | 13.91 |
| c | 5.18 | 5.70 | 5.59 | 6.06 | 6.22 | 6.78 | 6.63 |

Again, the dependence is a simple increasing straight line c = 0.7490 + 0.4010b, $R^2$ = 0.8572 .

**Open and closed syllables**

A third property of syllables that will be treated here is their end. If a syllable ends with a vowel, one says that it is open (O), if it ends with a consonant, it is closed (C). The results depend also from the way of treating diphthongs, week vowels etc. Since we adhere to the above segmentation, we obtain for the first text simply

| | | | | |
|---|---|---|---|---|
| Slovak | O = 182, | C = 65 | p = 0.7368, | u = 7.44 |
| Romani | O = 234, | C = 89 | p = 0.7245, | u = 8.07 |

| | | | | |
|---|---|---|---|---|
| Betmen: | O = 812, | C = 404, | p = 0.6678 , | u = 11.70 |
| O Roma: | O = 472, | C = 225, | p = 0.6772, | u = 9.36 |
| Hanka: | O = 776, | C = 405, | p = ,0.6571, | u = 10.80 |
| Deklaracija: | O = 710, | C = 275, | p = 0.7208, | u = 13.86 |
| Johanka: | O = 767, | C = 407, | p = 0.6533, | u = 10.51 |
| Romipen | O = 514, | C = 223, | p = 0.6974, | u = 10.72 |
| Interview: | O = 413, | C = 243, | p = 0.6296, | u = 6.64 |
| Census: | O = 745, | C = 317, | p = 0.7015, | u = 13.13 |
| O Baris: | O = 690, | C = 320, | p = 0.6832, | u = 11.64 |

| | | | | |
|---|---|---|---|---|
| Valakana: | O = 188, | C = 55., | p = 0.7737, | u = 8.53 |
| Holokaust | O = 566, | C = 242, | p = 0.7005, | u = 11.40 |
| RNĽ | O = 644, | C = 257, | p = 0.7148, | u = 12.89 |
| Angluno | O = 655, | C = 308, | p = 0.6801, | u = 11.18 |

Here, *p* is the proportion of the open syllables and *u* is the value of the normal test comparing the proportion of O with 0.5 and divided by the variance of p.

The result says that in Romani one prefers open syllables, a trend that is quite preferred in many languages. The sample sizes are very great hence a normal test is appropriate. The preference for open syllables is a general trend in language evolution: their pronunciation does not require as much effort as that of consonants. There are also languages in which no closed syllables exist (Polynesian).

## 3. Distances

The present section is devoted to the question, how syllable types are ordered in a formalized text. We express the order in terms of the *distances* between equal elements of the sequence of types. To illustrate this, we consider a small hypothetical text containing the syllable types V, VC, CV. Assume that these types form the sequence

$$\text{CV, V, VC, VC, V, CV, V, VC, CV, CV} \qquad (9)$$

At first, we concentrate on the distances between the types V:

-, V, -, -, V, -, V, -, -, -

where the horizontal line "-" indicates any syllable type **different from** V. Between the first two elements V the distance is 2, because there are two other elements "-" between them. Between the second and the third V the distance is 1 because there is exactly one element different from V between them. Now we concentrate on the type CV, yielding

CV, -, -, -, -, CV, -, -, CV, CV

where "-" now expresses any type **different from CV**. The distances between the first and the second appearance of CV is 4. Between the second and the third CV, the distance is 2 and between the third and fourth appearance of CV the distance is 0. In the same way, concentrating on the syllable type VC we obtain the sequence

-, -, VC, VC, -, -, -, VC, -, -

yielding the distances 0 and 3. Altogether, we obtained the distances 2,1,4,2,0,0,3, or – in ordered form - 0,0,1,2,2,3,4. The observed frequencies of the distance 0, 1, 2, 3, 4 are therefore $d_0 = 2$, $d_1 = 1$, $d_2 = 2$, $d_3 = 1$, $d_4 = 1$. The concept of distances can be applied to any sequence of linguistic entities, theoretical results, modifications and applications have been extensively studied in Zörnig (1984a,b, 1987, 2013).

In the following two tables we compute the distance frequencies $d_0$, $d_1$,…,$d_{20}$ for Romani texts, where we restrict our attention to distances smaller than or equal to 20. All 10 Romani texts can be well fitted by a simple exponential curve of the form

$$y = 1 + a\, exp(-\frac{x}{b}) \qquad (10)$$

the use of which will be justified below. For example, for the text *Deklaracija* (Table 19) we found the observed values $d_0 = 474$, $d_1 = 198$, $d_2 = 96$… and the corresponding theoretical values obtained by the above exponential model 470.32, 209.72, 93.82,…

There are languages with an excessive proportion of short distances, however the intrusion of foreign words may change this situation. There are nevertheless languages changing the foreign words in the "usual domestic" forms, e.g. the English word "December" has the form "kekemapa" in Hawaiian.

The formula (3.2) can be justified as follows. We assume that the relative rate of change of frequencies is depends linearly on y-1. Hence we write the relation as

$$\frac{y`}{y-1} = A. \qquad (11)$$

Solving this differential equation we obtain y = 1+ exp(Ax+B) or ln(y – 1) = Ax + B, respectively and setting A = -1/b, and exp(B) = a, we obtain the two parameter exponential function

$$y = 1 + a\, exp(-\frac{x}{b}). \qquad (13)$$

The distances may be different for every linguistic entity. In a certain sense syllables are mechanical entities whose succession is not conscious because one cares for the meaning and for the form of smaller entities (correct pronunciation). Not even in poetry, except for the rhyme, do they play a conscious role. In poetry one cares rather for rhythm, not for syllable types. Nevertheless, one can find regularities even here. In order to find them, we compute the distance between the syllables of the same type for many texts and try to find a regularity holding true at least for one language.

**Table 19**
Distances between equal syllable types in Romani texts

| | **Deklaracija** | | **Johanka** | | **Holokaust** | | **Romipen** | | **Interview** | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dist. | Frequ | Exp | Frequ | Exp. | Frequ | Exp | Frequ | Exp | Frequ | Exp |
| 0 | 474 | 470.32 | 453 | 451.71 | 347 | 341.60 | 353 | 347.06 | 276 | 276.91 |
| 1 | 198 | 209.72 | 262 | 254.43 | 156 | 170.12 | 131 | 153.00 | 159 | 153.88 |
| 2 | 96 | 93.82 | 119 | 143.50 | 83 | 84.98 | 82 | 67.76 | 78 | 85.71 |
| 3 | 49 | 42.28 | 76 | 81.13 | 53 | 42.70 | 32 | 30.32 | 51 | 47.93 |
| 4 | 24 | 19.36 | 70 | 46.06 | 27 | 21.70 | 28 | 13.88 | 23 | 27.01 |
| 5 | 22 | 9.16 | 32 | 26.34 | 22 | 11.28 | 21 | 6.66 | 19 | 15.41 |
| 6 | 14 | 4.63 | 22 | 15.25 | 13 | 6.10 | 9 | 3.48 | 9 | 8.98 |
| 7 | 9 | 2.61 | 14 | 9.01 | 6 | 3.53 | 8 | 2.09 | 9 | 5.42 |
| 8 | 9 | 1.72 | 12 | 5.50 | 16 | 2.26 | 9 | 1.48 | 2 | 3.45 |
| 9 | 8 | 1.32 | 6 | 3.53 | 7 | 1.62 | 2 | 1.22 | 9 | 2.36 |
| 10 | 3 | 1.14 | 4 | 2.42 | 3 | 1.31 | 4 | 1.09 | 8 | 1.75 |
| 11 | 8 | 1.06 | 7 | 1.80 | 4 | 1.15 | 3 | 1.04 | 4 | 1.42 |
| 12 | 3 | 1.03 | 6 | 1.45 | 8 | 1.08 | 4 | 1.02 | 8 | 1.23 |
| 13 | 4 | 1.01 | 5 | 1.25 | 2 | 1.04 | 0 | 1.01 | 2 | 1.13 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 1 | 1.01 | 5 | 1.14 | 4 | 1.02 | 3 | 1.00 | 4 | 1.07 |
| 15 | 3 | 1.00 | 4 | 1.08 | 4 | 1.01 | 3 | 1.00 | 3 | 1.04 |
| 16 | 1 | 1.00 | 3 | 1.05 | 0 | 1.00 | 2 | 1.00 | 4 | 1.02 |
| 17 | 2 | 1.00 | 3 | 1.03 | 3 | 1.00 | 1 | 1.00 | 2 | 1.01 |
| 18 | 2 | 1.00 | 3 | 1.01 | 3 | 1.00 | 4 | 1.00 | 1 | 1.01 |
| 19 | 3 | 1.00 | 1 | 1.01 | 1 | 1.00 | 2 | 1.00 | 2 | 1.00 |
| 20 | 1 | 1.00 | 4 | 1.00 | 4 | 1.00 | 1 | 1.00 | 1 | 1.00 |
| | a = 469.3247<br>b = 1.2341<br>$R^2$ =0.9971 | | a = 450-7108<br>b = 1.7369<br>$R^2$ = 0.9937 | | a = 340.5950<br>b = 1.4284<br>$R^2$ = 0.9934 | | a= 346.0569<br>b =1.2154<br>$R^2$ = 0.9899 | | a = 275.9101<br>b = 1.6936<br>$R^2$ = 0.9967 | |

| | **Rolnik (Slk)** | | **Phuvakero** | | **Hanka** | | **Betmen** | | **Census** | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dist | Frequ | Exp | Frequ | Exp. | Frequ | Exp | Frequ | Exp | Frequ | Exp |
| 0 | 97 | 96.46 | 153 | 150.96 | 509 | 500.72 | 750 | 729.90 | 360 | 356.62 |
| 1 | 47 | 47.84 | 56 | 64.58 | 234 | 254.65 | 346 | 397.20 | 204 | 213.62 |
| 2 | 22 | 23.98 | 36 | 27.95 | 125 | 129.75 | 222 | 216.36 | 138 | 128.13 |
| 3 | 13 | 12.28 | 13 | 12.43 | 80 | 66.36 | 139 | 118.06 | 73 | 77.01 |
| 4 | 11 | 6.53 | 5 | 5.84 | 51 | 34.17 | 79 | 64.63 | 37 | 46.45 |
| 5 | 3 | 3.71 | 12 | 3.05 | 24 | 17.84 | 44 | 35.59 | 31 | 28.17 |
| 6 | 2 | 2.33 | 2 | 1.87 | 23 | 9.55 | 37 | 19.80 | 24 | 17.25 |
| 7 | 4 | 1.65 | 4 | 1.37 | 11 | 5.34 | 33 | 11.22 | 18 | 10.71 |
| 8 | 3 | 1.32 | 7 | 1.16 | 7 | 3.20 | 26 | 6.55 | 9 | 6.81 |
| 9 | 2 | 1.16 | 3 | 1.07 | 7 | 2.12 | 28 | 4.02 | 12 | 4.47 |
| 10 | 2 | 1.08 | 2 | 1.03 | 7 | 1.57 | 14 | 2.64 | 8 | 3.08 |
| 11 | 6 | 1.04 | 0 | 1.01 | 7 | 1.29 | 11 | 1.89 | 11 | 2.24 |
| 12 | 2 | 1.02 | 3 | 1.01 | 4 | 1.15 | 8 | 1.48 | 4 | 1.74 |
| 13 | 2 | 1.01 | 0 | 1.00 | 4 | 1.07 | 9 | 1.26 | 2 | 1.44 |
| 14 | 1 | 1.00 | 3 | 1.00 | 2 | 1.04 | 8 | 1.14 | 9 | 1.27 |
| 15 | 1 | 1.00 | 1 | 1.00 | 5 | 1.02 | 11 | 1.08 | 7 | 1.16 |
| 16 | 0 | 1.00 | 0 | 1.00 | 1 | 1.01 | 9 | 1.04 | 3 | 1.09 |
| 17 | 2 | 1.00 | 0 | 1.00 | 5 | 1.00 | 2 | 1.02 | 3 | 1.06 |
| 18 | 1 | 1.00 | 0 | 1.00 | 1 | 1.00 | 4 | 1.01 | 2 | 1.03 |
| 19 | 0 | 1.00 | 2 | 1.00 | 4 | 1.00 | 5 | 1.01 | 1 | 1.02 |
| 20 | 0 | 1.00 | 1 | 1.00 | 1 | 1.00 | 3 | 1.00 | 5 | 1.01 |
| | a = 95.4573<br>b = 1.4945<br>$R^2$ = 0.9935 | | a = 149.964<br>b = 1.1653<br>$R^2$ = 0.9882 | | a = 499.7178<br>b = 1.4748<br>$R^2$ = 0.9950 | | a = 728.8988<br>b = 1.6404<br>$R^2$ = 0.9901 | | a = 355.6161<br>b = 1.9443<br>$R^2$ = 0.9955 | |

As can be seen, the distances in the Slovak version of a text (Rolnik) follow the same regularity as the Romani text.

## 4. Discussion

Romani has been influenced by many languages. In Slovakia it developed mainly under the influence of Slovak, also of Hungarian. Hence, if some results that are valid for other languages can be found in Romani it is a strong confirmation of the given regularity. The fact

that the rank-order of types follows the Zipf-Alekseev function is no surprise; the fact that the syllable length follows the Menzerathian function has been expected and one could expect also the inhomogeneity of syllable types.

We are aware of the fact that 14 texts are not enough but our aim was rather to show that even syllables behave lawlike. Their examination in other languages would be topical.

The syllable is an entity having both components (phonemes) and belonging to a hierarchy of material entities. One can search for super-syllable and hyper-syllable in two directions. First, the combination of syllables according to the accent yields poetic feet. Hence a foot is a kind of super-syllable. The next rhythmic level is the line of a poem which is a sequence of feet, e.g. in the hexameter there are 16 line types composed of dactyls and spondees, e.g. DDSS. The line of a poem is thus a hyper-syllable. In some poetries, the line can be divided in two parts, thus making the hierarchy still richer. Second, the syllables in a text can be collected into Köhlerian motifs. A qualitative motif is a sequence of entities (here syllables) which are all different; the next motif begins with a syllable that already occurred in the previous motif; The motif must not contain two syllables occurring in the previous motif. The motifs have some properties, e.g. length measured in terms of syllable numbers occurring in it. Thus the motif is a super-syllable. The next level are motifs of motifs, i.e. sequences containing a series of motifs none of which is repeated. One can call them syllabic hyper-motifs. This level has not been examined up to now and we do not know how is it structured. It contains types, their lengths, the distances between them, etc.

In every language there are morphemes which may be in some way related to syllables. The relation is stronger or weaker, e.g. in monosyllabic languages it is one to one. In analytic languages, the relation may slightly change, and in synthetic languages where a morph can be represented by a phoneme or a syllable or by several syllables, the relation is more complex. This fact could be used in typology.

## Analyzed texts

*Roľník*: Banga, Dezider: *Roľník*. In Banga, D. Le Khamoreskere čhavora. Slniečkove deti. Bratislava 2012, p. 200.

O *phuvakero*: Banga, Dezider: *O phuvakero*. In Banga, D. Le Khamoreskere čhavora. Slniečkove deti. Bratislava 2012, p. 201.

*Hanka*: In: Kumanová, Zuzana (ed.). Príbehy rómskych žien. Vakerben pal o romnija. Stories of Roma Women. Vinodol 2016, p. 38. Translated to Romani by Stanislav Cina.

*Betmen*: Berko, Milan. Betmen. In 6. zbierka literárnych prác: Píšeme a čítame spolu. Irinas taj genas jekhetane. Rómsky literárny klub. Available at: www.rolik.eu/zbierka.html.

O Roma: In: Kumanová, Zuzana. Rómovia vo fotografii Jozefa Kolarčíka-Fintického. Bratislava, Občianske združeneie IN MINORITA 2008, Translated to Romani by Erika Godlová.

*Romipen*: Fočár, Martin. Romipen khatar sal. In: Zbierka literánych´prác: O Nelkáčikos. Rómsky literárny klub. Available at www.rolik.eu/zbierka.html.

*Deklaracija*: Romengeri Deklaracija andal Slovakijakri republika pedal romaňi čhibakeri štandardizacija andre Slovakijakeri republika.

*Johanka*:. In: Kumanová, Zuzana (ed.). Príbehy rómskych žien. Vakerben pal o romnija. Stories of Roma Women. Vinodol 2016, p. 46. Translated to Romani by Stanislav Cina.

*Valakana*: Banga, Dezider: *Valakana*. In Banga, D. Le Khamoreskere čhavora. Slniečkove deti. Bratislava 2012, p. 243.

*Interview*: O Alojz Hlina: Hin amen but bare goďaver manuša pro hokej the fudbalos, no the pre romaňi problema. *Interview* by Roman Čonka. Translated to Romani by Inga Lukáčová. In Romano nevo ľil, 6/2012, p. 5.

*Cenzus*: Manušengero kherengero the sobakengero rachitšagos andro berš 2011. Štatisticko urados Slovačiko republikate. E obalka le informacijenca pal o manuš.

*O Baris*: Lacková, Elena: *O Baris* baro primašis. In Banga, Dezider (ed.) Genibarica. Doplnkové čítanie pre žiakov ZŠ. Bratislava, Goldpress 1993, p. 51–52.

*Holokaust*: Na džanav te bisteren pre miro čha. In Rusová, Zlatica: *Holokaust,* utrpenie slovenských Rómov. Holokaust, pharipen serviko romengero. Holokauszt, a szlovákiai romák szenvedései. Bratislava, Úrad vlády Slovenskej republiky 2017, p. 71.

*Angluno*: Cina, Stanislav – Kailáš, Štefan – Samko, Milan – Rusnáková, Jurina – Adam, Matej: Slovensko-rómsky terminologický slovník. Bratislava: Úrad splnomocnenca vlády pre národnostné menšiny 2012. 135 s. Available at: http://www.narodnostnemensiny.gov.sk/data/files/5957_slovensko-romsky-terminologicky-slovnik.pdf.

*RNĽ*: O Romane čhave andal o špecijalna školi hin Anglijate andro ajse školi sar aver čhave. Rómske deti zo špeciálnych škôl sú v Anglicku v bežných školách. Translated to Romani by Inga Lukáčová. In: Romano nevo ľil, No. 10/2011, p. 2.

## References

**Best, K.-H.** (2013). Silbenlängen im Deutschen*. Glottotheory,* 4, 36–44.

**Bičan, A.** (2015). Kvantitativní analýza slabiky v českém lexikonu [Quantitative analysis of the syllable in the Czech lexicon]. *Linguistica Brunensia,* 63(2), 87–107.

**Boretzky, N.** (1999). *Die Verwandschaftsbeziehungen zwischen den Südbalkanischen Romani-Dialekten. Mit einem Kartenanhang.* Frankfurt am Main: Peter Lang.

**Boretzky, N.** (2003). *Die Vlach-Dialekte des Romani Strukturen – Sprachgeschchte – Verwandtschaftsverhältnisse – Dialektkarten.* Wiesbaden: Harrasowitz.

**Brenzinger, M., Dwyer, A. M., Graaf, T. de, Grinevalsd, C., Krauss, M., Miyaoka, O., Ostles, N., Sakiyama, O., Villalón, M. E., Yammamotto, A. Y., Yepeda, O.** (2003). - Language Vitality and Endangerment. Document submitted to the International Expert Meeting on UNESCO. Retrieved from http://www.unesco.org/culture/doc/src/00120-EN.pdf.

**Cutler, A., Carter, D. M.** (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language,* 2, 133–142.

**Halwachs, D. W., F. Mentz** (eds.) (1999). *Die Sprache der Roma. Perspektiven der Romani-Forchung in Österreich im interdisziplinären und internazionalen Kontext.* Klagenfurt: Drava.

**Elšik, V., Matras, Y.** (2006). *Markedness and Language Change. The Romani Sample.* Berlin, Walter de Gruyter GmbH & Co. KG.

**Ivanecký, J., Majchráková, D.** (2007). Precision of Statistical Syllable Segmentation as a Function of Training Data Quality. Slovko. Bratislava.

**Jones, W. E.** (1971). Syllables and word-stress in Hindi. *Journal of the International Phonetic Association*. 1(2), 74–78.

**Kar, S.** (2010). *Syllable Structure of Bangla: An Optimality-Theoretic Approach.* Cambridge Scholars Publishing.

**Kelih, E., Mačcutek J.** (2013). Number of canonical syllable types: A continuous bivariate model. *Journal of Quantitative Linguistics,* 20, 241–251.

**Lee, Sang-Oak** (1986). An explanation of syllable structure change. *Korean Language Research,* 22, 195–213.

**Matras, Y., Bakker, P., Kyuchukov, H.** (1997): *The Typology and Dialectology of Romani.* Amsterdam Studies in ther Theory and History of Linguistic Science. Series IV – Current Inssues in Linguistic Theory. Vol 156. Amsterdam/Phladelphia, John Benjamins Publishing Company.

**Matras, Y.** (2002). *Romani: A Linguistic Introduction.* Cambridge: Cambridge University Press.

**Matras, Y.** (2005). The status of Romani in Europe. Report submitted to the Council of Europe´s Language Policy Division, October. Retrieved from https://romani.humanities.manchester.ac.uk/.../statusofromani.pdf

**Mušinka, A., Škobla, D., Hurrle, J., Matlovičová, K., Kling, J.** (2014). *Atlas rómskych komunít na Slovensku 2013* [The Atlas of Roma Communities in Slovakia 2013]. Bratislava: UNDP.

**Narayana, L. M., Ramakrishnan, A. G.** (s.d.). Defining Syllables and Their Stress Labels in Mile Tamil TTS Corpus. Retrieved from https://www.semanticscholar.org/paper/DEFININGSYLLABLES-AND-THEIR-STRESS-LABELS-IN-MILE-Narayana-amakrishnan/b8ceefd05822170530021962553028 0b3e900968

**Ohala, M.** (s.d.). Hindi syllabification. Retrieved from https://www.internationalphoneticassociation.org./…/p14_0727.pdf

**Rácová, A.** (1995). The lexicon of "Slovak" Romany language. *Asian and African Studies,* 4(1), 8–14.

**Rácová, A.** (2015a). Slovak language and Slovak Romani. In: Kyuchukov, H., Kwadrans, L., Fizik, L. (eds.), *Romani Studies: Contemporary Trends*. "Roma" series 2 (pp. 79–95). LINCOM GmbH.

**Rácová, A.** (2015b). Uplatňovanie rómčiny ako jazyka národnostnej menšiny na Slovensku [Introducing Romani as a national minority language in Slovakia]. In: Wachtarcyzková, J., Satinská, L., Ondrejovič, S. (eds.), *Jazyk v politických, ideologických a interkultúrnych vzťahoch* [Language in political, ideological, and inter-cultural context]: proceedings from the international conference Jazyk v politických, ideologických a interkultúrnych vzťahoch (pp. 302–319). Bratislava: Veda, Vydavateľstvo SAV.

**Rácová, A., Samko, M.** (2017). On the vitality and endangerment of the Romani language in Slovakia. *Asian and African Studies*, 26(2), 185–208.

**Rácová, A.** (2018). The impact of language ideologies on the language practices of Roma in Slovakia. *Asian and African Studies,* 27(2), 192–215.

**Rottmann, O.** (2002). Syllable length in Russian, Bulgarian, Old Church Slavic and Slovene. *Glottometrics,* 2, 87–94.

**Sabol J.** (1994). Slovenská slabika (Náčrt problematiky). In: J. Malcek (ed.), *Studia Academica Slovaca 23, Prednášky XXX. letného seminára slovenského jazyka a kultúry* (pp. 214-224). Bratislava: Stimul – Centrum informatiky a vzdelávania FF UK.

**Soravia, G.** (1977). *Profilo dei dialetti italiani. Dialetti degli Zingari italiani*. Pisa: Consiglio Nazionale delle Ricerche.

**Vennemann, T.** (1982) (ed.), *Zur Silbenstruktur der deutschen Standardsprache. Silben, Segmente, Akzente:* 261–305. Tübingen: Niemeyer.

**Wimmer, G., Altmann, G**. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook:* 791–807. Berlin/New York: de Gruyter.

**Zörnig, P.** (1984). The distribution of the distance between like elements in a sequence I. In: Boy, J., Köhler, R. (eds.), *Glottometrika* 6, 1–15. Bochum: Brockmeyer.

**Zörnig, P.** (1984). The distribution of the distance between like elements in a sequence II. In: U. Rothe (ed*.), Glottometrika* 7, 1–14. Bochum: Brockmeyer.

**Zörnig, P.** (1987). A theory of distances between like elements in a sequence. In: Fickermann, I. (ed.), *Glottometrika 8,* 1–22. Bochum: Brockmeyer.

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

*Glottometrics 46, 2019, 61–82*

# A Corpus-Based Study of the Semantic Prosody of Chinese Light Verb Pattern Across Registers: Taking *jinxing* and *shoudao* as Examples[1]

*Huiying Cai[2], Yunhua Qu[3], and Zhiwei Feng[4]*

**Abstract.** Light verbs constitute a fundamental element of the Chinese language. The construction "V+VN" was identified as a major light verb pattern (LVP) due to its high frequency of occurrence. Although the syntactic aspect of Chinese light verbs has been discussed in detail, the pragmatic aspect remains undetermined. Thus, this study examined *jinxing* 'be in progress' and *shoudao* 'come in for' as representatives of the two types of Chinese light verbs, agentive-action and accusative-action, respectively. Their patterns were regarded as complete entities of form and meaning. Register and semantic prosody, possessing an evaluative and attitudinal function, were introduced in order to explore the pragmatic aspect of Chinese light verbs. Based on a self-built corpus, the preference of semantic prosody created by the two typical LVPs and the influence of registers (spoken vs. written) on both, as well as on the semantic prosody created, were discussed in a quantitative manner. Furthermore, through the two general corpus approaches, variationist and text-linguistic, similarities and differences between the two LVPs themselves and their semantic prosody influenced by the registers were identified. The reasons for the corresponding results were discussed based on a situational analysis within the register analysis framework. The results broaden our understanding of Chinese light verbs from the syntactic aspect to the pragmatic aspect. Our findings would be substantially useful for Chinese language learning and teaching about the usage of Chinese LVPs in second language acquisition.

**Keywords:** ***Chinese, l****ight verb; light verb pattern; semantic prosody; register; corpus-based*

## 1. Introduction

The term "light verb", first put forward by Jespersen (1942), refers to verbs combined with verbal nouns in expressions, such as *take a walk, give a demonstration of the technique, make an offer, have a bite,* and *do research*. The form of these expressions, i.e., the collocation, could be summarized as the "V+VN" construction. The verb here is delexicalized, having no actual meaning, "The main semantic content of the predicate is provided not by it, but by the action nominal complement" (the verbal noun) (Kearns, 1988:595). In German linguistics, such constructions are termed "Funktionsverbgefüge", i.e., function verb constructions. Since the "V+VN" combination occurs relatively frequently and is dependent on the choice of word category, in this case light verbs, the construction can be identified as a light verb pattern (LVP) according to Hunston's definition of pattern (1999). "Delexicalized" light verbs also exist in the Chinese language, possessing similar features to those in English

[1] This project is supported by the National Social Science Foundation of China (17BYY002).

[2] School of International Studies, Zhejiang University, Hangzhou 310058, China. E-mail: zncc1123@163.com. ORCID No.: https://orcid.org/0000-0001-5110-4553.

[3] Corresponding author. School of International Studies, Zhejiang University, Hangzhou 310058, China. E-mail: qu163hua@163.com.

[4] School of Foreign Languages, Hangzhou Normal University, Hangzhou, China.

(Lv, 1980), which is also the case for LVPs. For example, in the patterns *jinxing caifang* 'be interviewing' and *shoudao shanghai* 'be hurt', *jinxing* 'be in progress' and *shoudao* 'come in for', respectively, are light verbs which do not express semantic meaning.

Hitherto, the syntactic aspect of Chinese light verbs has been deeply investigated in order to make more appropriate use of light verbs in sentences. In addition to their syntax and grammatical functions (Liu, 2007), their associated syntactical phenomenon has also been explored. Most of the extant literature largely followed the grammar of English light verbs (Liu et al., 2001; Shen, 2004; Zhu, 2005; Song, 2011; Huang and Lin, 2013). According to Shen (2004), for example, Chinese light verbs must exist in accordance with tense and aspect. However, current studies have had limited success, as the pragmatic aspect of light verbs has not been sufficiently examined.

This lack of a thorough understanding of Chinese light verbs has presented certain challenges. For example, research has found that foreign learners were often confused about how to select appropriate light verbs to express their actual attitudes or feelings when communicating in different situational contexts (Park, 2014). Consequently, situational context must be considered in association with the pragmatic dimension (Zhu, 1985). In pattern grammar, which "focuses on the formal components of a pattern, rather than on a structural interpretation of those components", it is believed that patterns and meaning are connected (Hunston, 1999). Therefore, the purpose of regarding the "V+VN" combination as the LVP here is to treat the formal components as a whole of form and meaning. Then, we can explore the particular meanings possessed by associated lexical items, which would contribute substantially to our understanding of the pragmatic aspect of light verbs.

The important communicative function of Chinese light verbs, which are served by their linguistic features in situational contexts (Biber & Conrad, 2009), was investigated in this study. Specifically, although Chinese light verbs are capable of exchanging without altering the original meaning (Zhu, 1985), it remains necessary to determine whether the semantic meaning of the attitude or evaluation expressed in context would change after the light verb has changed. In this case, semantic prosody, regarded as linguistic features that have an attitudinal and evaluative function (Sinclair, 1996), was included to examine the pragmatic functions of light verbs. Moreover, registers, which connect situational contexts and language (linguistic features) (Biber, 2000), are hypothesized to be the factor causing the difference in the preference of semantic prosody created by LVPs during communication.

Due to a lack of extant pragmatic research concerning light verbs, a corpus-based study was performed to analyze relationships between register, semantic prosody, and LVPs. The following measures have been taken for operationalization. First, according to the two classifications of Chinese light verbs suggested by Zhu (1982), two typical Chinese light verbs were selected as representatives: *shoudao* 'come in for', the only accusative-action light verb; and *jinxing* 'be in progress', the most frequently-used one among the six agentive-action light verbs. Their patterns were regarded as the research objects here. Second, semantic prosody, "a specific halo or profile" created by a pattern (i.e., "a habitual co-occurrence of two or more words") (Bublitz, 1996; Stubbs, 1996), was classified into three categories: positive; negative; and neutral prosody, in accordance with that proposed by Stubbs (1996), who emphasized its pragmatic function. Based on the attitude or evaluation expressed, with the occurrence of a pattern, if a strong favorable or pleasant aura is created, the category of semantic prosody is identified as positive (e.g., provide support); if a strong negative or unpleasant aura is created, the category is identified as negative (e.g., cause death); and if

neutrality is sensed or the context provides no evidence of semantic prosody, the category is identified as neutral (e.g., as a result). In this study, the categories of semantic prosodies created by each LVP were identified according to whether speakers'/writers' attitude or evaluation in the sentence is positive, negative or neutral, based on the intuition of the native speaker (the researcher) within the contexts considered. Third, the spoken register and the written register were compared to explain possible differences in the preference of semantic prosody created by LVPs across registers, due to their distinct difference of interactivity (Biber & Conrad, 2009). These register differences were analyzed by two important corpus approaches: variationist and text-linguistic (Biber, 2011) which complement each other. The discussion focuses on situational analysis within the register analysis framework (Biber & Conrad, 2009). The research outlined above comprises the following research questions and hypotheses:

(1) Is there any characteristic or preference of semantic prosody (positive, negative, or neutral) when created by the different light verb patterns (LVPs) (*jinxing* 'be in progress' and *shoudao* 'come in for')?

(2) How does the register (spoken or written Chinese) influence the employment of the Chinese LVPs?

Hypothesis: Chinese LVPs would be applied more in the written register than in the spoken register.

(3) How does the register influence semantic prosody of LVPs?

Hypothesis: Neutral prosody would be applied more in the written register than in the spoken register.

(4) Is there any difference between the influence of the register on the preference of semantic prosody for the *jinxing* 'be in progress' pattern and for the *shoudao* 'come in for' pattern?

The significance of the present study comprises two aspects. In the linguistic field, research about Chinese light verbs is broadened from syntactic aspects to pragmatic aspects, and a novel perspective is offered by introducing semantic prosody and analyzing the influence of registers. In the practical field of education, the study may provide valuable guidance for Chinese learners and instructors concerning communicative function. Specifically, misleadings and misunderstandings would be substantially avoided from the output, speaking/writing, and the input, listening/reading.

The remainder of this paper is organized as follows. The extant literature about semantic prosody, register, and the syntactic aspect of Chinese light verbs is reviewed in Section 2. The research design, the corpus adopted, and the entire procedure are introduced in Section 3. The distributional relationships between LVPs, semantic prosody and register are presented, and the data are analyzed, in Section 4. A discussion about the findings is given in Section 5. Finally, in Section 6, conclusions are drawn, and future research directions are given.

## 2. Theoretical background

### 2.1 Syntactic features of Chinese light verbs

The present study refers to Jespersenian phonetically light verbs, whose features are as follows: (1) they have a semantic bleaching; (2) they cannot be independently predicated; (3)

they support verbal nouns in collocation with semantic and grammatical functions; and (4) they are affixal verbs (Liu, 2007). Chinese light verbs possess similar features to the ones above. However, unlike in the English language, since there is not as much morphologic change in the Chinese language, the verbal noun after a Chinese light verb always accords with its verb form (Liu, 2007:11). It is widely accepted that the seven Chinese light verbs could be classified into two categories: six agentive-action verbs, including *jinxing* 'be in progress', *you* 'have', *zuo* 'do', *jiayi* 'in addition', *geiyi* 'give' and *yuyi* 'give'; and one accusative-action verb, *shoudao* 'come in for' (Zhu, 1982). The agent of the former category serves as the subject, while the object of the latter category serves as the subject. Liu (2007:151) noted that the combination of verbs and light verbs in the Chinese language exists on the level of syntax, and thus the property of Chinese light verbs is syntax. He also observed three features of the grammatical function of Chinese light verbs: (1) they are all transitive verbs with other verbs or the verbal nouns that follow; (2) they are affixal, and not the actual main predicate; and (3) the object must be disyllabic, in which the property of both verb and noun exists (Liu, 2007).

Current studies of Chinese light verbs have mainly focused on properties and functions, as well as the syntactical phenomenon of the Chinese language. Most Chinese researchers largely followed the grammar of English light verbs. For example, Liu, Pan and Hu (2001) reported that Chinese light verbs also carry the mark of the tense and aspect in forms, containing the meaning of "handle" and "deal with" in themselves. Zhu (2005) analyzed the phenomenon, in which objects are positioned after intransitive verbs by employing Chinese light verbs. Song (2011) explored the standard of event coercion using light verbs. However, Huang and Lin (2013:728) argued that the order of light verbs is closely associated with two kinds of information that the verbs specify: "aspectual eventive information"; and "argument information", when observing that two or more light verbs sometimes co-occur in Mandarin Chinese.

### 2.2 Light verbs in pattern grammar

In pattern grammar, a word is regularly associated with certain words and structures, which are regarded as patterns that contribute its meaning. A pattern possesses three main features: (1) the combination of words occurs relatively frequently; (2) it is dependent on a particular choice of words; and (3) it is associated with a clear meaning (Hunston, 1999:37). For light verbs, the very frequent combination "V+VN" is closely related to this particular word category, and the semantic meaning is normally given by the verbal noun (VN) that follows. Thus, the construction can be identified as the light verb pattern (LVP). The pattern can also be associated with a variety of different words, i.e., different light verbs, and will tend to be associated with VNs that convey particular meanings (Hunston, 1999).

### 2.3 Communicative function of LVPs from the pragmatic aspect

Although the syntactic aspect of light verbs has been elucidated, the pragmatic aspect has not yet been sufficiently determined. The study aimed to identify the communicative function of light verbs through semantic prosody (linguistic features) in spoken and written registers (situational contexts). The patterns of the two typical light verbs, *jinxing* 'be in progress' and *shoudao* 'come in for', are regarded as the research objects.

### 2.3.1 Semantic prosody and its categorization

Semantic prosody has been considered due to its attitudinal and evaluative function, and its pragmatic property (Sinclair, 1996). It was first termed by Louw (1993:157), who defined it as a "consistent aura of meaning with which a form is imbued by its collocates". For example, the subjects of the phrasal verb, *set in*, described as having a negative semantic profile, always referred to some unpleasant states of affairs, such as *rot*, *decay*, *despair*, or *bitterness* (Sinclair, 1987:155-6). According to Cheng (2013), various methods have been applied to describe the characteristics of semantic prosody, since attitudinal and pragmatic meanings comprise multiple aspects (Sinclair, 1996; Bublitz, 1996; Zhang, 2010; Stewart, 2010; Hunston, 2007). Since "prosodies are usually attributed to semantically more neutral items" (Stewart, 2010:2), it can be inferred that patterns of light verbs which accord with semantic neutrality could create corresponding semantic prosody to express speakers'/writers' attitudes or evaluations. Furthermore, a variety of verbs associated with patterns frequently possess common meanings (Hunston, 1999). Consequently, although light verbs have no actual meaning, semantic prosodies created by patterns of different light verbs (combinations of V+VN) might exhibit similarities. Since semantic prosody, having a communicative purpose, is "on the pragmatic side of the semantics-pragmatics continuum" (Sinclair, 1996), it was considered here to specify the communication function (attitudinal or evaluative connotation) of light verbs from the pragmatic aspect by analyzing the relationship between semantic prosody and LVPs.

Bulbiltz (1996) observed that "words can have a specific halo or profile, which may be positive, pleasant and good, or else negative, unpleasant and bad". In order to make the present study more viable, semantic prosody was classified into three categorizations: positive prosody; negative prosody; and neutral prosody, in accordance with Stubbs' proposal (1996:176). The main purpose of the present study was to elucidate the pragmatic aspect of light verbs through semantic prosody, and not necessarily to observe the features of semantic prosody itself. Therefore, Stubbs' classification, the ternary distinction, could constitute the clearest and most reliable one to support exploration of semantic prosody created by LVPs. Since the patterns are usually strikingly different across registers (Biber, 2011) and semantic prosodies are context-dependent and register-dependent (Cheng, 2013), the register, a functional variety of language is introduced due to its influence on light verb patterns and semantic prosody.

### 2.3.2 Spoken register vs. written register

Varieties of registers possess corresponding situational characteristics, "including distinctive aspects of the context and communicative purpose" (Biber, 2000). The linguistic features, caused by the situational context of the register, are functional, (Biber & Conrad, 2009), which makes it possible for the characteristics or preferences of semantic prosody created by LVPs to differ across registers. Moreover, Biber (2011) highlighted the omission of register analyses in pattern grammar, in which "the authors chose to increase their coverage of words rather than investigating the possibility of different patterns occurring in different registers" (2011: 28). Thus, the present study focused on the functional interpretation of the relationship between register (situation) and semantic prosody (linguistic features) created by LVPs. Based on the register analysis framework developed by Biber and Conrad (2009:50), the situational analysis was referred to as an approach, "describing the situational characteristics of the registers, including distinctive aspects of the context and communicative purpose".

Situational characteristics relevant for describing and comparing registers are as follows: participants; channel; production circumstances; setting; communicative purposes; and topic (2009:40).

Based on the two ways of expression, language can be divided into spoken and written. The differences between spoken registers and written registers, for example, the patterns of linguistic variation and use (Biber, 2011), have been documented. Moreover, the importance of one or the other have been thoroughly explored. Differences of situational characteristics were also pointed out by Biber and Conrad (2009). Among these, regarding the aspect of communicative function, the primary purpose in the spoken register is interpersonal interaction, conveying personal feelings and attitudes, rather than describing or explaining factual information. On the other hand, the written register concerns communicating objective information, instead of developing a personal relationship (Biber & Conrad, 2009). In other words, spoken registers are interactive, while written registers are non-interactive. Since the obviously different features between spoken and written language should be acknowledged, especially when learning Chinese (Zhu, 1985), spoken registers and written registers are considered in order to explain the preference of semantic prosody created by LVPs.

Spoken and written registers were compared using two general corpus approaches: variationist; and text-linguistic. Variationist designs, related to the linguistic variation literature, consider each linguistic token (light verb) as an observation and investigate proportional preferences. Text-linguistic designs, on the other hand, related to studies of text-linguistic variation, consider each text as an observation, and investigate the rates of occurrences in texts (Biber, 2011). Although the importance of registers has been more apparent in text-linguistic research (2011:12), both approaches are important, as the results that they produce are "to a large extent complementary" (2011:27). Consequently, they were both employed to investigate the use of LVPs, features of semantic prosody, and the two registers.

## 3. Methods

### 3.1 Research design

The study aimed to elucidate the characteristics or preferences of semantic prosody created by Chinese LVPs. The influence of registers was also a topic of focus. A corpus built by Qu in 2016, the Zhejiang University Spoken and Written Corpus of Mandarin Chinese, was applied as a tool to analyze Chinese LVPs in different registers. Two typical Chinese light verbs, *jinxing* 'be in progress' and *shoudao* 'come in for' were, respectively, selected from the seven light verbs listed by Zhu (1982) as representatives of the two types of Chinese light verbs, agentive-action and accusative-action. The two Chinese light verbs, which are opposite and typically-used, were qualified to represent both agentive-action and accusative-action light verbs.

The patterns of two light verbs, *jinxing* and *shoudao*, were extracted from the corpus. Sentences, including the LVPs, were also examined to identify their semantic prosody within contexts. Based on the three categorizations of semantic prosody classified by Stubbs (1996), positive, negative and neutral prosody, LVPs were classified by the researcher, who is a native speaker of the Chinese language. Then, the relationship between semantic prosody and the LVPs was analyzed. Moreover, as a linguistic feature, semantic prosody in different registers (spoken vs. written) differs as well, as does the LVP. Thus, the influence of the two registers

on the use of the LVPs on semantic prosodies of LVPs (text-linguistic design) and on the preference of semantic prosody for each LVP (variationist design) was also discussed by using the two general corpus approaches, which are complementary and both important. Specifically, the text-linguistic design intended to analyze the direct influence of register on semantic prosody of LVPs by observing the rates of occurrence for semantic prosody of LVPs across texts. Each text was considered as an observation to emphasize the importance of registers. On the other hand, the variationist design intended to analyze the influence of register by observing the proportional preference of semantic prosody for each LVP. Each linguistic token (light verb) was considered as an observation to emphasize the characteristic of the token itself.

### 3.2 Corpus

The study was based on a corpus built by Qu in 2016, the Zhejiang University Spoken and Written Corpus of Mandarin Chinese, comprising 1,100,000 words of spoken Mandarin Chinese and 1,100,000 words of written Mandarin Chinese. In total, this constitutes 1,100 spoken texts and 1,066 written texts. Each text has approximately 1,000 words. This corpus has been shown to be reliable, and determined by the Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS) that it has achieved the accuracy up to 98%. The Zhejiang University Spoken and Written Corpus of Mandarin Chinese has been classified into different registers, as shown in Table 1 and Table 2.

**Table 1**

Composition of the Zhejiang University Corpus of Spoken and Written Mandarin Chinese (spoken)

| Category | Register | Number of texts | Percentage |
|---|---|---|---|
| Conversation | Face-to-face conversation | 55 | 5% |
| | Telephone calls | 275 | 25% |
| | Internet speech | 60 | 5.5% |
| Spoken with written characteristics | Television talk show | 125 | 11.4% |
| | Debate | 78 | 7.1% |
| | Play and movie | 95 | 8.6% |
| | Chinese folk arts (cross talk, storytelling, and storytelling in Beijing dialect with percussion accompaniment) | 40 | 3.6% |
| Narration | Oral narratives | 102 | 9.3% |
| | Edited oral narratives | 270 | 24.5% |
| **Total** | **9** | **1100** | **100%** |

**Table 2**

Composition of Zhejiang University Corpus of Spoken and Written Mandarin Chinese (written)

| Category | Register | Number of texts | Percentage |
|---|---|---|---|
| A | News reports | 93 | 8.7% |
| B | News editorials | 53 | 5% |
| C | News reviews | 154 | 14.4% |
| D | Religion | 31 | 2.9% |

| | | | |
|---|---|---|---|
| E | Skills, trades, and hobbies | 73 original texts + 25 texts on martial arts, Chinese painting= 98 | 9.2% |
| F | Popular lore | 86 | 8.1% |
| G | Biographies and essays | 23 | 2.2% |
| H | Miscellaneous: Reports and official documents | 63 | 5.9% |
| J | Science: Academic prose | 165 original texts + 25 texts on mental calculation and Chinese medicine = 190 | 17.8% |
| K | General fiction | 50 | 4.7% |
| L | Mystery and detective fiction | 42 | 4% |
| M | Science fiction | 11 | 1% |
| N | Adventure and martial arts fiction | 49 | 4.6% |
| P | Romantic fiction | 57 | 5.3% |
| R | Humor | 16 | 1.5% |
| S | Lyrical | 25 texts on modern poems + 25 texts on modern lyrics = 50 | 4.7% |
| **Total** | **16** | **1066** | **100%** |

### 3.3 Procedure

The first step was to apply Antconc, a corpus analysis toolkit for concordancing and text analysis, to extract all of the sentences, including at least one of the two Chinese light verbs, i.e., *jinxing* 'be in progress' and *shoudao* 'come in for', from the corpus. The raw data were checked to determine whether the verb functioned exactly as a light verb, i.e., whether or not it possessed the real meaning of its pattern, which should be expressed only by its following verbal noun. If not, the corresponding sentences were deleted to ensure that the results were appropriate for the study. The occurrence of each light verb in each text was recorded through a self-designed Python program. Then, the sentences with each light verb were classified by register. Thus, the results were divided into four parts: *jinxing* in the spoken register; *shoudao* in the spoken register; *jinxing* in the written register; and *shoudao* in the written register (see Figure 1).

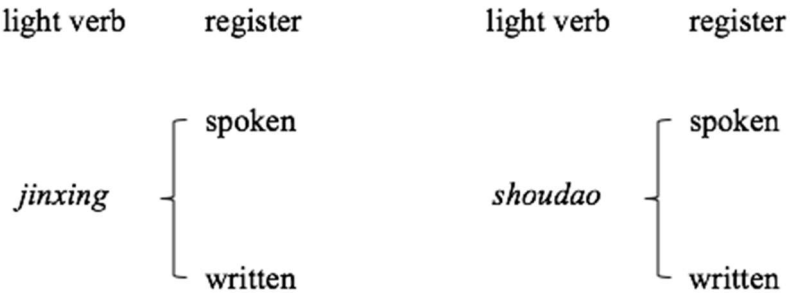


**Figure 1.** Four parts extracted from the corpus.

The second step was to identify the category of semantic prosody created by each LVP from the sentences extracted, with consideration of context and in accordance with the three categories defined by Stubbs (1996): positive; negative; and neutral prosody. The occurrence of each category of semantic prosody of *jinxing* patterns and that of *shoudao* patterns in each text were recorded. Then, the distribution of each category of semantic prosody of each light

verb in the whole corpus was listed, respectively. Moreover, the distribution of each category of semantic prosody of *jinxing* and of *shoudao* in each register, spoken and written, was listed, respectively, as well (see Figure 2). Then, the results were analyzed generally and specifically.

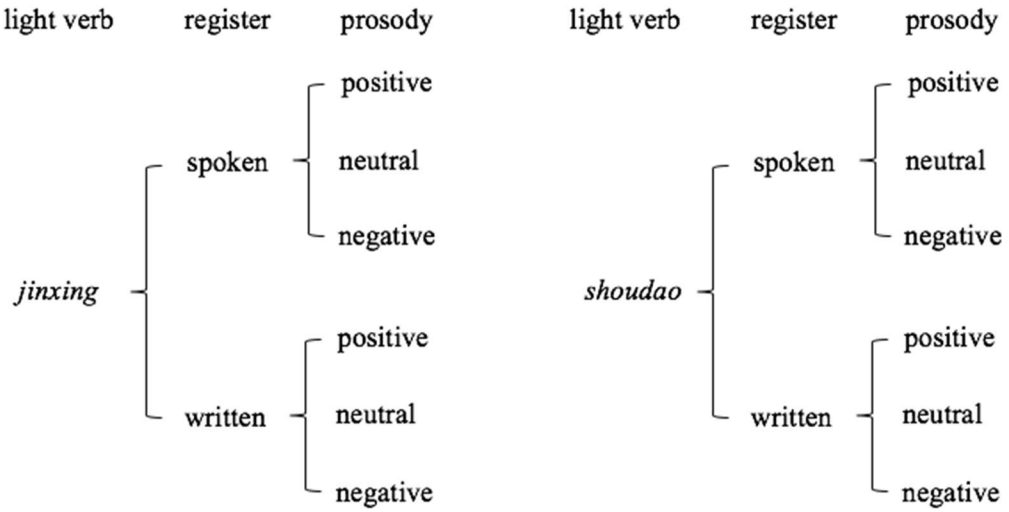


**Figure 2.** Identification of the three categories of semantic prosody.

The third step was to confirm the correlation between semantic prosody and the LVPs by a chi-squared test to compare the preference of semantic prosody of the two LVPs with the data containing the occurrence of each category created by each LVP. Then, the general characteristic or preference of semantic prosody was observed for each LVP by comparing the distribution of each category of semantic prosody of *jinxing* and *shoudao* in the entire corpus, respectively. In addition, the category of semantic prosody of each LVP that comprised the largest portion was identified. The formula is as follows:

$$\frac{prosody\ (\text{positive, negative, or neutral})}{\text{LVP}\ (jingxing\ \text{or}\ shoudao)}$$

The fourth step was to observe the employment of LVPs in different registers. The relationship between the LVPs and the registers was shown through a one-way analysis of variance (ANOVA) and a chi-squared test, which were performed by SPSS. The null hypothesis for the ANOVA was that *the use of each LVP does not change significantly across the registers*. If the result revealed that the difference was significant, it can be inferred that the use of each LVP did change significantly across the registers, and the register did have some influence on the use of LVP. The chi-squared test was conducted to compare the use of the two LVPs considering the register with the data containing the occurrence of each LVP in each register. Furthermore, the distribution of LVPs in the spoken register and that in the written register were compared according to the following formula:

$$\frac{\text{LVP } (jinxing \text{ or } shoudao)}{\text{register (spoken or written)}}$$

The fifth step was to observe semantic prosodies created by the two LVPs across registers from the perspective of text-linguistic variation. Similarly, for each LVP, the relationship between semantic prosody and the registers was identified through a one-way ANOVA and a chi-squared test, which were also performed by SPSS. The null hypothesis for the ANOVA was that *for each LVP, each category of semantic prosody created does not change significantly across the registers.* If the result revealed that the difference was significant, it can be concluded that semantic prosodies did change significantly across the registers, and registers did have some influence on semantic prosodies. The chi-squared test was conducted to compare the semantic prosodies across the registers for each LVP with the data containing the occurrence of each category of semantic prosody for each LVP in each register. In the corpus employed, since the total number of words in spoken texts and that in written texts were identical, i.e., 1,100,000 words, comparing the numbers of each category of semantic prosody created by each LVP in the spoken register with that in the written register was equivalent to comparing the rates of occurrence. Thus, the numbers were analyzed directly.

The sixth step was to observe the proportional preference of semantic prosody for each LVP in each register from the perspective of linguistic variation. The distribution of each category of semantic prosody of *jinxing* in the spoken register and that in the written register were compared, as was that of *shoudao*. In that case, the researcher observed which category of semantic prosody was the most proportionally preferred when using *jinxing* in the spoken register or in the written register, and when using *shoudao* in the spoken register or in the written register. The formulas are listed as follows:

In spoken registers:

$$\frac{prosody \text{ (positive, negative, or neutral)}}{\text{LVP } (jinxing \text{ or } shoudao)}$$

In written registers:

$$\frac{prosody \text{ (positive, negative, or neutral)}}{\text{LVP } (jinxing \text{ or } shoudao)}$$

Through the last three steps, the influence of registers on semantic prosody created by the two LVPs could be made explicit. The reason for this phenomenon will be discussed, as well.

The whole procedure, including the six steps, is briefly illustrated in Figure 3.

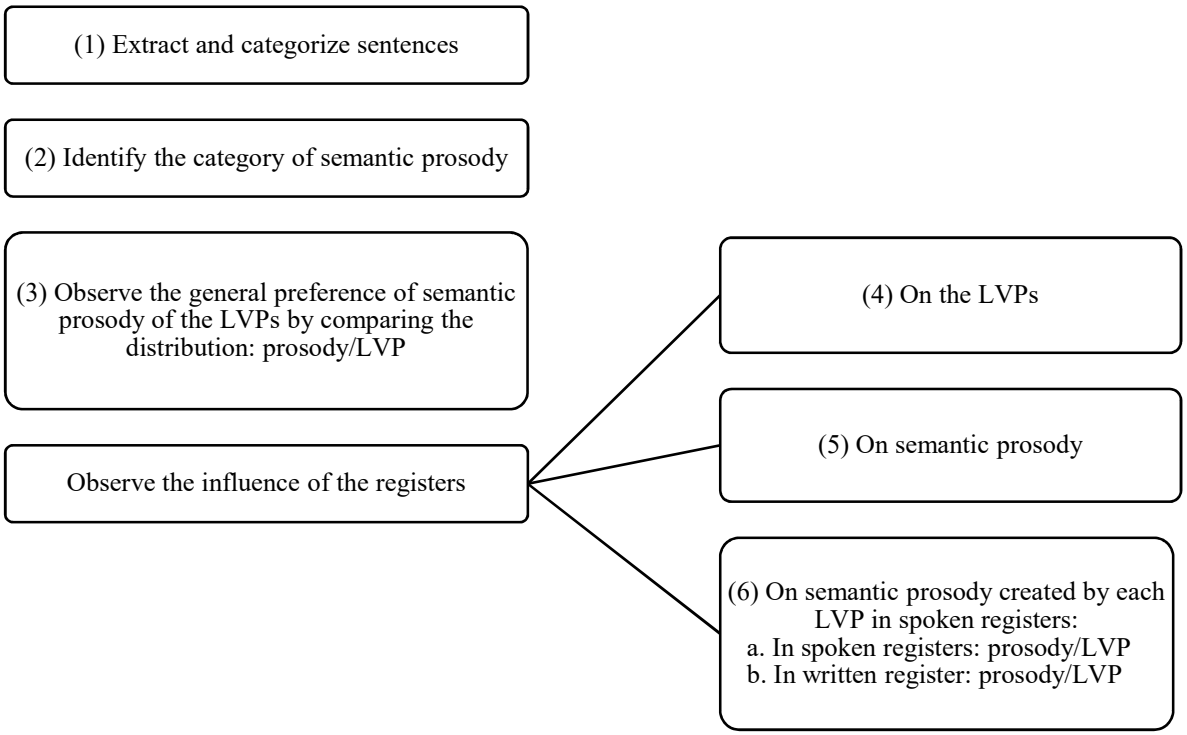


**Figure 3.** Procedure

## 4. Results

1,100 sentences were eventually extracted, including the word *jinxing* 'be in progress', and 241 sentences, including the word *shoudao* 'come in for' in the entire Zhejiang University Spoken and Written Corpus of Mandarin Chinese. According to the definition of the light verb, i.e., the delexicalized verb in "V+VN" patterns which has no actual semantic content (Kearns, 1988:595), those containing the two words functioning as light verbs were extracted, consisting of 972 sentences including *jinxing* and 240 sentences including *shoudao*. The category of semantic prosody of each LVP was examined manually with consideration of context according to Stubbs' standards (1996). The number of each category was collected. These data were applied to identify the semantic prosody created by the LVPs in the spoken register and the written register. The results are presented as follows.

### 4.1 *Jinxing* pattern

*Jinxing* 'be in progress' patterns occurred 972 times in the corpus. All of them were divided into three categories: 243 with positive prosody; 151 with negative prosody; and 578 with neutral prosody, comprising 25.00%, 15.53%, and 59.47%, respectively. Considering registers, there were 271 *jinxing*s in the spoken register and 701 ones in the written register. In the spoken register, there were 103 *jinxing*s with positive prosody, 79 *jinxing*s with negative prosody and 89 *jinxing*s with neutral prosody, constituting 38.01%, 29.15%, and 32.84%, respectively. In the written register, there were 140 *jinxing*s with positive prosody, 72 *jinxing*s

with negative prosody and 489 *jinxing*s with neutral prosody, constituting 19.97%, 10.27%, and 69.76%, respectively. Examples of these three categories of *jinxing* pattern are listed in Table 3. All of the items were extracted from the corpus. Two examples were from the spoken register, and other two were from the spoken register in each category.

**Table 3**
Examples of the light verb *jinxing*

| | Register | Category of semantic prosody | Example |
|---|---|---|---|
| *Jinxing* | Spoken | Positive | 进行重金奖励<br>*Jinxing* zhongjin jiangli<br>To reward with a huge sum of money |
| | | | 对一些设施进行完善<br>Dui yixie sheshi *jinxing* wanshan<br>To improve some facilities |
| | | Negative | 进行疯狂的报复<br>*Jinxing* fengkuang de baofu<br>Taking a furious revenge |
| | | | 请您进行举报<br>Qing nin *jinxing* jubao<br>Please report |
| | | Neutral | 根据提示进行操作<br>Genju tishi *jinxing* caozuo<br>Operating according to prompts |
| | | | 对某个市场要进行研究<br>Dui mouge shichang yao *jinxing* yanjiu<br>To study a market |
| | Written | Positive | 择优进行表彰<br>Zeyou *jinxing* biaozhang<br>To commend outstanding ones |
| | | | 进行了积极的探索<br>*Jinxing* le jiji de tansuo<br>Have had a positive exploration |
| | | Negative | 对我国进行制裁<br>Dui woguo *jinxing* zhicai<br>Carrying out sanctions against our country |
| | | | 对中泰进行起诉<br>Dui zhongtai *jinxing* qisu<br>To prosecute Zhongtai Group |
| | | Neutral | 用遗传学的方法进行筛选<br>Yong yichuanxue de fangfa *jinxing* shuaixuan<br>Screening by genetic methods |
| | | | 从词汇学的角度进行分析<br>Cong cihuixue de jiaodu *jinxing* fenxi<br>Analyzing from the perspective of lexicology |

### 4.2 *Shoudao* pattern

*Shoudao* ‘come in for’ patterns appeared 240 times in the corpus. They were all divided into three categories, as well: 60 with positive prosody; 82 with negative prosody, and 98 with neutral prosody, comprising 25.00%, 34.17%, and 40.83%, respectively. Considering registers, there were 56 *shoudao*s in the spoken register, and 184 in the written register. In the spoken register, there were 17 *shoudao*s with positive prosody, 28 *shoudao*s with negative prosody,

and 11 *shoudao*s with neutral prosody, constituting 30.36%, 50.00%, and 19.64%, respectively. In the written register, there were 43 *shoudao*s with positive prosody, 54 *shoudao*s with negative prosody and 87 *shoudao*s with neutral prosody, constituting 23.37%, 29.35%, and 47.28%, respectively. Examples of these three categories of *shoudao* are presented in Table 4. The items were taken from the corpus. In each category, there were two examples in the spoken register, and the other two in the written register.

**Table 4**

Examples of the light verb *shoudao*

| | Register | Category of semantic prosody | Example |
|---|---|---|---|
| *Shoudao* | Spoken | Positive | 受到热烈欢迎<br>Shoudao Relie Huanying<br>Have been warmly welcomed |
| | | | 而且受到很多鼓励<br>Erqie Shoudao Henduo Guli<br>And have been encouraged |
| | | Negative | 使敌人受到多大损失<br>Shi Diren Shoudao Duoda Sunshi<br>Made enemies lose a lot |
| | | | 受到种种精神折磨<br>Shoudao Zhongzhong Jingshen Zhemo<br>Was tortured mentally in different ways |
| | | Neutral | 这个收入就受到影响了<br>Zhege Shouru Jiu Shoudao Yingxiang Le<br>Income has been affected |
| | | | 比如你受到很多关注<br>Biru Ni Shoudao Henduo Guanzhu<br>As if you received much attention |
| | Written | Positive | 使全体人员受到很大鼓舞<br>Shi Quanti Renyuan Shoudao Henda Guwu<br>Make all of the staff quite inspired |
| | | | 这样的人总是容易受到敬仰<br>Zheyang De Ren Zongshi Rongyi Shoudao Jingyang<br>That kind of people are always easily respected |
| | | Negative | 她受到了不能容忍的侮辱<br>Ta Shoudao Le Buneng Rongren De Wuru<br>She has been insulted unbearably |
| | | | 受到严厉的整治<br>Shoudao Yanli De Zhengzhi<br>Be rectified severely |
| | | Neutral | 为了让观众受到教育<br>Weile Rang Guanzhong Shoudao Jiaoyu<br>In order to educate the audience |
| | | | 新闻受到重视的程度<br>Xinwen Shoudao Zhongshi De Chengdu<br>The extent of attention received by news |

### 4.3 Comparison of the two LVPs

The statistics above are all displayed in Table 5. Comparing the two LVPs, *jinxing* ‘be in progress’ and *shoudao* ‘come in for’, they themselves, and the characteristic or preference of semantic prosody in the two registers, exhibit certain similarities and differences. Concerning

the relationships between semantic prosodies, LVPs and registers, reference is made to the results of one-way ANOVAs, which are displayed in Table 7 and Table 9, and the results of chi-squared tests.

**Table 5**
Semantic prosodies of the LVPs across registers

| | Register (1,100,000 words each) | Positive prosody | Neutral prosody | Negative prosody | Total |
|---|---|---|---|---|---|
| ***Jinxing*** | Spoken | 103 (38.01%) | 89 (32.84%) | 79 (29.15%) | 271 (100%) |
| | Written | 140 (19.97%) | 489 (69.76%) | 72 (10.27%) | 701 (100%) |
| | Total | 243 (25.00%) | 578 (59.47%) | 151 (15.53%) | 972 (100%) |
| ***Shoudao*** | Spoken | 17 (30.36%) | 11 (19.64%) | 28 (50.00%) | 56 (100%) |
| | Written | 43 (23.37%) | 87 (47.28%) | 54 (29.35%) | 184 (100%) |
| | Total | 60 (25.00%) | 98 (40.83%) | 82 (34.17%) | 240 (100%) |

### 4.3.1 Semantic prosodies created by the LVPs

In general, as shown in Table 6, the distribution of neutral prosody of the *jinxing* pattern (59.47%) and that of the *shoudao* pattern (40.83%) are the largest among the three categories in the whole corpus. In addition, unlike the latter, the former is much larger than the other two categories. However, the difference between semantic prosodies of the *shoudao* pattern and those of the *jinxing* pattern is still statistically significant ($p < 0.05$): for the *shoudao* pattern, the distribution of negative prosody (34.17%) is greater than that of positive prosody (25.00%); whereas, for the *jinxing* pattern, the distribution of positive prosody (25.00%) is greater than that of negative prosody (15.53%) (see Table 6).

**Table 6**
LVPs with different categories of semantic prosody

| LVP | Positive prosody | Neutral prosody | Negative prosody | Total |
|---|---|---|---|---|
| *Jinxing* | 243 (25.00%) | 578 (59.47%) | 151 (15.53%) | 972 (100%) |
| *Shoudao* | 60 (25.00%) | 98 (40.83%) | 82 (34.17%) | 240 (100%) |

### 4.3.2 LVPs across the registers

By comparing the mean scores of each LVP in the two different registers, it can be seen in Table 7 that, for both the *jinxing* pattern and the *shoudao* pattern, the difference in use between the spoken texts and the written texts is significant ($p < 0.05$).

**Table 7**
Features of the LVPs across registers

| Light verb | Spoken register mean score | Written register mean score | F value | Significance |
|---|---|---|---|---|
| *Jinxing* | 0.2464 | 0.6576 | 83.375 | < 0.05 |
| *Shoudao* | 0.0509 | 0.1726 | 55.721 | < 0.05 |

Although the total number of *jinxing* patterns is much greater than that of *shoudao* patterns, no significant differences are found between the use of the two LVPs across registers ($p > 0.05$). It can be inferred that the influence of registers on the use of both the LVPs is similar. Through both the mean score comparison in Table 7 and the distribution comparison in Table 8, it can be concluded that the use of the two LVPs (*jinxing*; *shoudao*) is much more frequent in the written register than in the spoken register.

**Table 8** LVPs across registers

| LVP | Spoken register | Written register | Total |
|---|---|---|---|
| *Jinxing* | 271 (27.88%) | 701 (72.12%) | 972 (100%) |
| *Shoudao* | 56 (23.33%) | 184 (76.67%) | 240 (100%) |

### 4.3.3 Semantic prosodies across registers (text-linguistic perspective)

From the text-linguistic perspective, by comparing the mean scores of the semantic prosodies of the LVPs in the two different registers, it can be seen in Table 9 that all of the categories of semantic prosodies created by both of the LVPs in the spoken texts are significantly different from those in the written texts ($p < 0.05$). This reflects the influence of register, except for the negative prosody created by *jinxing* patterns ($p > 0.05$). The register does not seem to have the capacity to influence negative prosody when created by *jinxing* patterns, which is probably because not-preferring negative prosody is already the norm of the *jinxing* pattern.

**Table 9**

Features of semantic prosodies of the two LVPs(*jinxing* and *shoudao*) across registers

| Linguistic feature | Spoken register mean score | Written register mean score | F value | Significance |
|---|---|---|---|---|
| ***Jinxing* patterns** | | | | |
| Positive prosody | 0.0936 | 0.1313 | 4.637 | < 0.05 |
| Neutral prosody | 0.0809 | 0.4587 | 134.375 | < 0.05 |
| Negative prosody | 0.0718 | 0.0675 | 0.077 | > 0.05 |
| ***Shoudao* patterns** | | | | |
| Positive prosody | 0.0155 | 0.0403 | 10.990 | < 0.05 |
| Neutral prosody | 0.0100 | 0.0816 | 47.387 | < 0.05 |
| Negative prosody | 0.0255 | 0.0507 | 7.870 | < 0.05 |

In addition, the influence of the two registers can also be reflected from the rates of occurrence of each category of semantic prosody created by each LVP across the two registers. Since the total number of words in spoken texts and that in written texts are identical, the raw numbers of occurrence, listed in Table 5, are directly compared, representing the rates of occurrence (per 1,100,000 words).

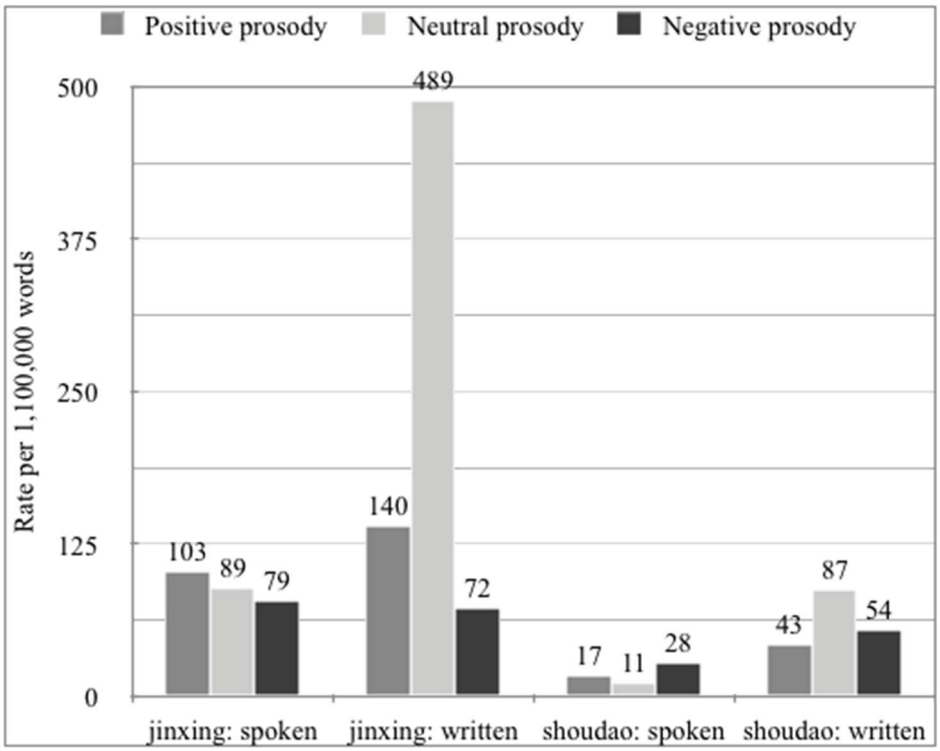


**Figure 4.** Rates of occurrence (per 1,100,000 words) for semantic prosodies of the two LVPs (*jinxing* and *shoudao*) across registers.

For both the *jinxing* pattern ($p < 0.05$) and the *shoudao* pattern ($p < 0.05$), the difference between the occurrences of semantic prosodies created in the spoken register and that in the written register is statistically significant. This means that the registers exert some influence on semantic prosodies created by LVPs, regardless of *jinxing* or *shoudao*. As shown in Figure 4, in the written register: (1) neutral prosodies of both the LVPs (*jinxing*: 489 times; *shoudao*: 87 times) are significantly more frequent than those in the spoken register (103 times and 11 times, respectively); (2) positive prosodies of both (*jinxing*: 140 times; *shoudao*: 43 times) are more frequent than those in the spoken register (103 times and 17 times, respectively) as well, which may be due to the great difference in the total number across the registers; and (3) similarly, negative prosody of *shoudao* (54 times) is also more frequent than in that in the spoken register (28 times), while only negative prosody of *jinxing* (72 times) is less frequent than that in the spoken register (79 times).

#### 4.3.4 Preference of semantic prosody across registers (variationist perspective)

From the variationist perspective, the influence of the two registers is reflected by the proportional preference of each category of semantic prosody for each LVP across the two registers, which is listed in Table 5.

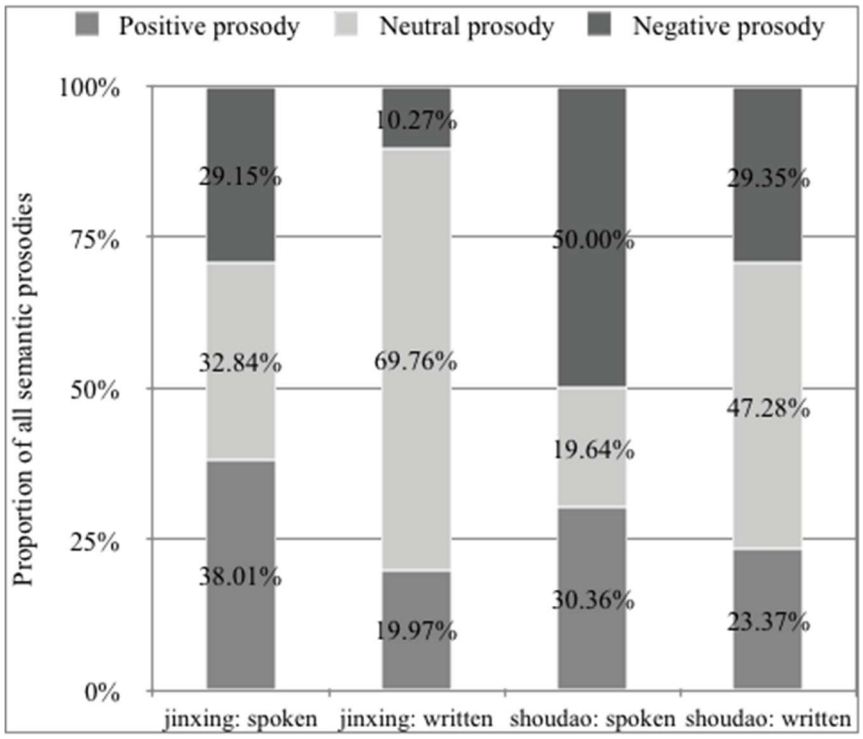


**Figure 5.** Proportional preference of semantic prosodies for each of the LVPs (*jinxing* and *shoudao*) across registers (spoken vs. written).

From Figure 5, it can be inferred that the influence of registers on the preference of semantic prosody for the *jinxing* pattern and for the *shoudao* pattern is quite similar, as well. Specifically, in the spoken register, the proportion of semantic prosody that expressed clear attitudes is the largest: for *jinxing*, positive prosody comprises the largest proportion (38.01%); whereas, for *shoudao*, negative prosody comprises the largest proportion (50.00%). In the written register, in contrast, the proportion of neutral prosody (69.76% and 47.28%, respectively) is the largest. In other words, for both of the LVPs, in the spoken register and the written register, the proportion of neutral prosody increased significantly, while the proportions of the other two prosodies that expressed clear attitudes decreased.

The different features of the LVPs are also worth noting: irrespective of the spoken register or the written register, for the *jinxing* pattern, positive prosody (38.01%; 19.97%) is somewhat proportionally preferred over negative prosody (29.15%; 10.27%); whereas, for the *shoudao* pattern, the distribution of negative prosody (50.00%; 29.35%) is proportionally preferred over positive prosody (30.36%; 23.37%). Especially in the spoken register, for *jinxing* patterns, the proportion of positive prosody (38.01%) is the largest; whereas, for *shoudao* patterns, the proportion of negative prosody (50.00%) is the largest, and the difference (19.64%) between the two prosodies is much more significant.

## 5. Discussion

The categorization of semantic prosody proposed by Stubbs (1996) was applied to analyze the relationship between semantic prosody and the patterns of the two typical Chinese light verbs, i.e., the agentive-action verb *jinxing* 'be in progress' and accusative-action verb *shoudao* 'come in for' (Zhu, 1982), in a quantitative manner. The influence of registers was

also considered from both the variationist perspective and the text-linguistic perspective. From the results shown above, it was discovered that the relationship between semantic prosody, register, and each LVP exhibited certain similarities and differences.

### 5.1 Reflection of communicative function of LVPs

In response to the first research question, generally, the category of neutral prosody was more preferred than the other two when created by both of the Chinese LVPs, the *jinxing* 'be in progress' pattern and the *shoudao* 'come in for' pattern, regardless of the register. This similarity could be regarded as the reflection of the communicative function of light verbs. In addition, unlike *shoudao*, neutral prosody of *jinxing* was much more preferred than the other two categories.

Since semantic prosody exhibits an attitudinal and evaluative function when communicated in situational contexts, the communicative function of the Chinese light verbs could be reflected by the preference of semantic prosody created by LVPs. According to the results above, the communicative function of both the LVPs preferring neutral prosody, especially *jinxing*, conveyed objective information, and the communicative function of expressing clear attitudes or evaluation appeared to be relatively more obvious when employing *shoudao*. Consequently, it could be concluded that light verbs, containing no actual semantic meaning, are more likely to be associated with neutral prosody.

### 5.2 Influence of the register

According to the results, the influence of the registers on the two Chinese LVPs and on the semantic prosody created by them are elucidated and interpreted in the following two subsections, respectively.

#### 5.2.1 The relationship between registers and LVPs

In response to the second research question, different registers did have different preferences of the Chinese LVP. Specifically, both the *jinxing* 'be in progress' pattern and the *shoudao* 'come in for' pattern were utilized more frequently in the written register than in the spoken register.

The differences in production circumstances of spoken and written registers may lead to the above phenomenon. As real-time spoken registers provide no time for planning prior to deciding what to say and no possibility of revision or edition (Biber & Conrad, 2009), speakers would say verbal words (VN) that can instantaneously convey their thoughts directly without employing redundant light verbs, which possess no actual semantic content. In contrast, written registers provide sufficient time for planning, revising, and editing to create more formal works to convey information (Biber & Conrad, 2009). The LVP, which is longer than the single verb, is regarded as more formal and information-condensed, and exhibits more consideration, and thus it is more frequently employed in written registers. Especially, Chinese writers tend to use four-character patterns in their written works. Thus, it is reasonable that the LVP (V+VN), which is more likely to comprise a four-character pattern, is more preferred in the written register.

#### 5.2.2 The relationship between registers and semantic prosody

In response to the third research question, the registers did exert some influence on semantic prosody created by both of the LVPs. Specifically, concerning the *jinxing* pattern

and the *shoudao* pattern, from the text-linguistic perspective, neutral prosody is much more used in the written register. However, from the variationist perspective, semantic prosody that expressed clear evaluations or attitudes, positive or negative, was preferred in the spoken register, while neutral prosody was preferred in the written register. This constitutes the similarity of the two LVPs.

Due to the different primary purpose of spoken and written registers, which concerns the aspect of communicative function, the preference of semantic prosody for each of the LVPs was different in the two registers. Since reaching the primary goal of spoken registers, i.e., interaction, requires an interpersonal communicative function, the expression of a personal stance, such as personal feelings, attitudes, desires, likes and dislikes, is more necessary than objective information (Biber & Conrad, 2009). Thus, in spoken registers, semantic prosodies of both the LVPs prefer to be positive or negative, expressing clear evaluations or attitudes to create interactive situations and to develop interpersonal relationships. Nevertheless, since the primary focus of written registers is communicating information rather than developing a personal relationship (Biber & Conrad, 2009), the language used is more objective and non-interactive, without referring to personal details, such as evaluations or attitudes. Consequently, in written registers, semantic prosody of the two LVPs prefers to be neutral, which shows objective attitudes to convey information effectively without involving personal relationships and sentiment.

Overall, created both by the *jinxing* pattern and by the *shoudao* pattern, in the spoken register, semantic prosody that expressed clear evaluations or attitudes, positive or negative, was more preferred; whereas, in the written register, neutral prosody was much more preferred than the other two categories. These results are consistent with the widely-accepted theory that semantic prosody possesses an attitudinal and evaluative function (Sinclair, 1996). Moreover, positive or negative prosody is applied more in spoken registers to express certain relevant attitudes or evaluations in order to achieve effective communication, especially when face-to-face. However, it is also worth noting that the characteristics of semantic prosody of *jinxing* patterns are not exactly the same as those of *shoudao* patterns. The detailed differences are explained in the following section.

### 5.3 Differences between the jinxing pattern and the shoudao pattern

Providing a confirmative answer to the fourth research question, the differences between the characteristics of semantic prosody of *jinxing* ‘be in progress’ and of *shoudao* ‘come in for’ should be noted. Generally, positive prosody of *jinxing* was more preferred than negative prosody, while negative prosody of *shoudao* was more preferred than positive prosody. From the text-linguistic perspective, comparing the occurrences of all semantic prosodies created by the two across the registers, only the occurrences of negative prosody created by *jinxing* patterns in the written register were used less than in the spoken register. In addition, through the comparison of the mean scores, it was speculated that the incapacity of registers to influence negative prosody when created by *jinxing* patterns is probably because not-preferring negative prosody is already the norm of the *jinxing* pattern. From the variationist perspective, irrespective of the spoken register or the written register, *jinxing* patterns preferred positive prosody more than negative prosody; whereas, *shoudao* patterns preferred negative prosody more than positive prosody. Especially in the spoken register, *jinxing* patterns preferred positive prosody the most. In contrast, *shoudao* patterns preferred negative prosody the most, and the difference of preference between positive prosody and negative

prosody was much more significant.

It can be inferred that the agentive-action light verb *jinxing* is more likely to be related to positive prosody, and the possibility of creating negative prosody is less. Concerning the preference of negative prosody for the *shoudao* pattern rather than positive prosody, it is supposed that the reason might be that the accusative-action light verb *shoudao*, which functions in a similar manner to the passive voice in the English language, is frequently associated with negative semantic profiles in the Chinese language.

### 5.4 Practical significance

The results of the present study for Chinese language learning and teaching offer practical significance. The pragmatic aspect of the two Chinese light verbs, *jinxing* 'be in progress' and *shoudao* 'come in for', has been discussed with consideration of semantic prosody and the registers. It can be concluded that misleadings and misunderstandings will be reduced on both sides of the output, speaking and writing, and the input, listening and reading. Considering the frequency of using Chinese light verbs, *jinxing* and *shoudao*, and also in different registers, our determination of the preferred positive or negative prosodies created by the combination in the spoken register and the preferred neutral prosody in the written register is quite useful for instructors and Chinese language learners, especially foreign learners, who are learning Chinese light verbs. Meanwhile, because of the distinction between the agentive-action verb *jinxing* and the accusative-action verb *shoudao*, in which *jinxing* patterns prefer positive prosody while *shoudao* patterns prefer negative prosody, it must be noted that Chinese light verbs cannot be generalized. Indeed, the selection of Chinese light verbs must still rely on register and context.

## 6. Conclusion

In the present study, semantic prosody, having an attitudinal and evaluative function, was introduced to specify the pragmatic aspect of light verbs. The influence of registers was involved as well, with situational contexts considered. The relationships between semantic prosodies, light verbs, and registers were illustrated by both of the general corpus approaches, variationist and text-linguistic. Two main similarities were revealed. First, there was a similar preference of semantic prosody, i.e., neutral prosody, when created by the two LVPs, the *jinxing* 'be in progress' pattern and the *shoudao* 'come in for' pattern, which reflects the neutrality of light verbs. Second, the registers exerted a similar influence on both of them and on their semantic prosodies. Specifically, the LVPs were applied more in the written register than in the spoken register. This may be because the four-character Chinese LVP is preferred in written works and is regarded as more formal and information-condensed in planned, revised, or edited written works. Semantic prosody that expressed clear attitudes was preferred more in the spoken register and neutral prosody much more in the written register. This is ascribed to the difference in interactivity between the two registers. The difference between the patterns of the two light verbs has also been elucidated: when expressing clear evaluations or attitudes, the agentive-action light verb *jinxing* is more likely to be related to positive prosody and the possibility of creating negative prosody is less. On the other hand, the accusative-action verb *shoudao* tended to be related to negative prosody rather than positive prosody, which is probably due to its association with negative semantic profiles in the Chinese language.

Through the above findings, our understanding of light verbs in the Chinese language has been broadened from syntactic aspects to pragmatic aspects by determining their preference of semantic prosody, and a novel perspective was proposed by relating LVPs with semantic prosody and registers. The results could be substantially useful for Chinese language learning and instructors in second language acquisition concerning the usage of Chinese light verbs by avoiding misleadings and misunderstandings, as much as possible, when making output or input.

The present study is not without limitations. Depending on subjective judgments, the standard applied to identify the category of semantic prosody of LVPs was not sufficiently accurate to infer general regularity. In order to improve reliability, future investigations could be performed to confirm or deny our results. Further advancements could also be accomplished concerning the pragmatic aspect of Chinese light verbs. For example, more representatives of Chinese light verbs could be selected, coverage of the corpus could be broader, and categories of registers could be more subdivided. Through those improvements, pragmatic features could be more reliable and generalizable.

## Acknowledgement

The authors express gratitude to Dan Zhang and Shu Liu for their assistance to extract sentences containing the two light verbs.

## References

Biber, D. (2000). *Corpus Linguistics.* Cambridge University Press.

Biber, D. (2011). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37.

Biber, D., & Conrad S. (2009). *Register, Genre, and Style*. Cambridge University Press.

Bublitz, W. (l996). Semantic prosody and cohesive company: Somewhat predictable. *Leuvense Bijdragon: Tijddschrift voor Germaanse Filologie*, 85(l–2), l–32.

Cheng, W. (2013). Semantic Prosody. In C. A. Chapelle (Eds.), *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. London, England: Oxford University Press.

Halliday, M.A.K., McIntosh, A. & Strevens, P. (1964). *The Linguistic Sciences and Language Teaching*. London: Longman.

Huang, C. R., & Lin, J. (2013). The ordering of mandarin Chinese light verbs. In D. H. Ji, & G. Z. Xiao (Eds.), *Chinese Lexical Semantic Workshop 2012*, LNAI 7717 (pp. 528–735). Berlin Heidelberg: Springer-Verlag.

Hunston, S. (2007). Semantic prosody revisited. *International Journal of Corpus Linguistics*, 12(2), 249–268.

Hunston, S., & Francis, G. (1999). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English.* Philadelphia, America: John Benjamins.

Jespersen, O. (1942). *A Modern English Grammar on Historical Principles, Part VI, Morphology*. London: George Allen and Unwin Ltd.

Kearns, K. (1988). Light verbs in English. Master's thesis, MIT, Cambridge, MA. Reprinted in Linguistics and Philosophy, 595–635.

Liu, C. Q. (2007). Light verb theory and the corresponding study in Chinese Language. Ph.D.

dissertation, Wuhan University, Wuhan, China.

Liu, Y. H., Pan, W. Y., & Hu, W. (2001). *Shiyong Xiandai Hanyu Yufa* 'Chinese Grammar'. Beijing, China: The Commercial Press.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157–175). Amsterdam, Netherlands: John Benjamins.

Lv, S. X. (1980). *Xiandai Hanyu Ba Bai* Ci 'Eight Hundred Words in Modern Chinese'. Beijing, China: The Commercial Press.

Park, Z. X. (2014). Waiguo ren hanyu qing dongci xide qingkuang kaocha 'A study of the acquisition of Chinese light verbs'. Master dissertation, Fudan University, Shanghai, China.

Shen L. (2004). Aspect Agreement and Light verb in Chinese: A comparison with Japanese. *Journal of East Asian Linguistics*, 13, 309–336.

Shi, J., & Wang, H. (2008). Qing dongci jiashe jiqi yingyong jiazhi 'Light-verb hypothesis and the uses in syntax analysis'. *Journal of Shanxi University (Philosophy and Social Sciences)*, 3, 37–39.

Sinclair, J. (1987). *Looking up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London, England: Collins.

Sinclair, J. (1996). The Search for Units of Meaning. *Textus*, 9(1), 75–106.

Song, Z. Y. (2011). Qing dongci, shijian yu hanyu zhong de binyu qiangpo 'Light verbs, events and object forcing in Chinese language'. *Studies of the Chinese Language*, 3, 205–217.

Stewart, D. (2010). *Semantic prosody: A critical evaluation*. London, England: Routledge.

Stubbs, M. (1996). *Text and Corpus Linguistics*. Oxford: Blackwell.

Wang, H. L., & Zhang, K. L. (2014). Jiyu yuliaoku de qingdongci jiegou hanying fanyi yanjiu: yi *jinxing* lei jiegou weili 'A corpus-based study of Chinese-English translation of light verb construction: taking the *jinxing* construction as example'. *Journal of PLA University of Foreign Languages*, 37(2), 62–68.

Yu, S. W., Zhu, X. F., & Duan, H. M. (2005). Xiandai hanyu zhong de xingshi dongci 'Dummy verbs in contemporary Chinese'. *Computational Linguistics and Chinese Language Processing*, 10(4), 509–518.

Zhang, C. H. (2010). An overview of corpus-based studies of semantic prosody. *Asian Social Science*, 6(6), 190–194.

Zhu, D. X. (1982). *Yufa Jiangyi* 'The Handouts of Grammar'. The Commercial Press.

Zhu, D. X. (1985). Xiandai shumian hanyu li de xuhua dongci he ming dongci 'delexical verbs and nominal verbs in modern written Chinese'. *Journal of Peking University (Philosophy and Social Sciences)*, 5, 10–16.

Zhu, X. F. (2005). Qing dongci he hanyu bu jiwu dongci dai binyu xianxiang 'The phenomenon of light verbs and Chinese intransitive verbs with objects'. *Modern Foreign Languages*, 3, 221–231.

*Glottometrics 46, 2019, 83–98*

# Quantifying the Quantitative Meter: On Rhythmic Types in the Dactylic Hexameter

*Karl-Heinz Best*
*Michal Místecký*[1]
*Peter Zörnig*[2]
*Gabriel Altmann*

**Abstract**. In the present article, we show that the rank-order of the hexameter types used in the poetry productions in four languages (Greek, Latin, German, and Czech) abides by the exponential function. Up to now, more complex probability distributions have been used (cf. Best 2008, 2009), namely the negative hypergeometric and the Pólya ones, having one parameter more. We also establish the relation between the parameters of a given model.

***Keywords****: Hexameter, poetry, Greek, Latin, German, Czech, exponential function*

In the present article, we are going to investigate the distributions of types of dactylic hexameter lines in poetic productions of various languages. The line of the Antiquity-originated dactylic hexameter consists of six feet, out of which the last two are usually normed, forming the so-called "heroic ending" (–UU / –U, or –UU / – –). Hence, we have merely four different feet in each line. Since the foot is either a spondee (S), or a dactyl (D), one obtains 16 different line patterns, namely –

DDDD, DDDS, DDSD, DSDD, SDDD, DDSS, DSDS, DSSD,
SDDS, SDSD, SSDD, DSSS, SDSS, SSDS, SSSD, SSSS.

Many examinations were made already in the eighteenth century (cf. Drobisch 1866, 1868a, b), and the research has continued up to now (cf. Grotjahn 1981; Best 2008, 2009 and Internet (cf. Hexameter). Various problems have been analyzed; here, we are interested only in the state/character of the line type in the hierarchy of lower entities. K.-H. Best (2008, 2009) fitted the negative hypergeometric and the Polya distributions with excellent results; here, we shall restrict ourselves to the simpler exponential function. This is the usual development in science – one begins with an appropriate model and, in the sequel, one tries to make it simpler.

The individual foot consists of syllables, representing the basic entities. The foot depends on the quantity of individual syllables and represents a super-unit. The same statement holds true for each kind of regular poetry, but the number of types would be different. The line of the poem is represented by a sequence of super-units (feet), and one can say that it is a hyper-unit. In this way, one can set up various hierarchies depending on the basic entity. However, one should not mix the aspects, e.g., the material aspects with the grammatical or semantic ones.

We conjecture that there is a very great number of possible hierarchies in texts. The best known ones are the Köhlerian motifs, consisting of sequences of not-repeated entities of

[1] Michal Místecký, University of Ostrava; e-mail: MMistecky@seznam.cz.
[2] Peter Zörnig, University of Brasilia, e-mail: peter@unb.br.

any sort. The super-motif is a sequence of different motifs, the hyper-motif is a restricted sequence of super-motifs. In the future, this research must be deepened.

If we construct hierarchies, then the sample should consist of a number of finished frame units. One can, of course, perform a random sampling of any sort, but before doing it, one must set up a hypothesis in a mathematical form, and after sampling, test it on the data. Sometimes, even the data are not satisfactory and must be assembled in a different form.

Here, we shall simply ask how many individual hyper-syllable types (= line types) of the above list appear in a selected corpus. That is, we simply start from the assumption that the line types do not occur with equal frequencies, but that the frequencies can be ranked. As usual, we start from the hypothesis that the relative rate of the frequency change is a constant for a given poem or language, i.e. –

$$\frac{y'}{y} = -b,$$

yielding –

$$y = ae^{-bx},$$

i.e. the exponential function. The parameter $b$ is negative because we ordered the types according to the decreasing frequencies. That means we conjecture that the frequencies change constantly, depending on the frequency of the given class. As to ranking, it is no "false" procedure because one simply seeks a variable which influences the frequencies, even if it is found and numerically expressed a posteriori. We conjecture that this function will hold true for all hexameters, but there is a possibility that entities other than feet may abide by a different regime.

Other approaches are the Zipfian power function and the Zipf-Mandelbrot function, having three parameters. The samples can be compared using classical tests, and either one compares the frequencies or the ranks of the identical hyper-motifs. We restrict ourselves to the modelling of Drobisch's data (1866) according to the exponential function. The results are presented in Table 1.

**Table 1**
Frequency of individual line types in Latin hexameters (Drobisch 1866)
and the exponential function fit

| Rank | **Vergil** | | **Horace** | | **Lucrece** | | **Manilius** | |
|---|---|---|---|---|---|---|---|---|
| | Frequency | Expon | Frequency | Expon | Frequency | Expon | Frequency | Expon |
| 1 | 78 | 79.25 | 62 | 60.39 | 88 | 87.27 | 93 | 88.05 |
| 2 | 75 | 69.11 | 53 | 55.65 | 72 | 74.94 | 69 | 75.53 |
| 3 | 57 | 60.27 | 49 | 51.28 | 63 | 64.35 | 67 | 64.79 |
| 4 | 52 | 52.56 | 48 | 47.25 | 56 | 55.26 | 57 | 55.58 |
| 5 | 44 | 45.84 | 44 | 43.54 | 51 | 47.46 | 48 | 47.68 |
| 6 | 38 | 39.98 | 38 | 40.12 | 39 | 40.75 | 34 | 40.90 |
| 7 | 33 | 34.86 | 38 | 36.97 | 37 | 35.00 | 33 | 35.09 |
| 8 | 30 | 30.40 | 36 | 34.07 | 36 | 30.05 | 30 | 30.10 |
| 9 | 27 | 26.52 | 35 | 31.39 | 26 | 25.81 | 28 | 25.82 |
| 10 | 27 | 23.12 | 32 | 28.93 | 21 | 22.16 | 22 | 22.15 |
| 11 | 22 | 20.17 | 30 | 26.65 | 17 | 19.03 | 22 | 19.00 |
| 12 | 20 | 17.59 | 22 | 24.56 | 17 | 16.34 | 19 | 16.30 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 13 | 18 | 15.34 | 22 | 22.63 | 12 | 14.03 | 16 | 13.99 |
| 14 | 13 | 13.38 | 20 | 20.86 | 10 | 12.05 | 11 | 11.99 |
| 15 | 12 | 11.66 | 20 | 19.22 | 8 | 10.35 | 11 | 10.29 |
| 16 | 4 | 10.17 | 11 | 17.71 | 7 | 8.89 | 9 | 8.83 |
| | a = 90.8680<br>b = 0.1369<br>$R^2$ = 0.9820 | | a = 65.5357<br>b = 0.0818<br>$R^2$ = 0.9617 | | a = 101.6228<br>b = 0.1523<br>$R^2$ = 0.9903 | | a = 102.6397<br>b = 0.1533<br>$R^2$ = 0.9831 | |

| Rank | **Persius** | | **Juvenal** | | **Lucanus** | |
|---|---|---|---|---|---|---|
| | Frequency | Expon | Frequency | Expon | Frequency | Expon |
| 1 | 118 | 110.88 | 85 | 80.19 | 98 | 94.10 |
| 2 | 96 | 92.78 | 67 | 69.90 | 83 | 79.04 |
| 3 | 68 | 77.64 | 64 | 60.94 | 59 | 66.38 |
| 4 | 62 | 64.97 | 51 | 53.12 | 58 | 55.75 |
| 5 | 48 | 54.37 | 40 | 46.31 | 39 | 46.82 |
| 6 | 39 | 45.49 | 38 | 40.37 | 33 | 39.33 |
| 7 | 35 | 38.07 | 35 | 35.19 | 32 | 33.03 |
| 8 | 35 | 31.85 | 29 | 30.68 | 29 | 27.74 |
| 9 | 32 | 26.66 | 26 | 26.74 | 28 | 23.30 |
| 10 | 30 | 22.30 | 25 | 23.31 | 23 | 19.57 |
| 11 | 27 | 18.66 | 22 | 20.32 | 19 | 16.44 |
| 12 | 19 | 15.62 | 20 | 17.72 | 15 | 13.80 |
| 13 | 13 | 13.07 | 19 | 15.44 | 13 | 11.59 |
| 14 | 12 | 10.94 | 14 | 13.46 | 12 | 9.74 |
| 15 | 9 | 9.15 | 13 | 11.74 | 11 | 8.17 |
| 16 | 5 | 7.66 | 12 | 10.23 | 8 | 6.87 |
| | a = 132.5077<br>b = 0.1782<br>$R^2$ = 0.9707 | | a = 91.9889<br>b = 0.1378<br>$R^2$ = 0.9824 | | a = 112.0432<br>b = 0.1745<br>$R^2$ = 0.9762 | |

The frequencies represent here merely samples, not the complete data by Drobisch.

K.-H. Best (2008, 2009) used several analyses of Drobisch; we show them in Table 2, applying the exponential function. Best applied the negative hypergeometric distribution, which has three parameters; we use here the exponential function, having only two parameters. As has been said many times, the use of a distribution or a function is merely a matter of mathematical choice.

**Table 2**
Fitting the exponential function to hexameters: Vergil
(Source: Drobisch 1868; Best 2008)

| | Vergil, *Aeneis* | | | Vergil, *Georgica* | | | Vergil, *Bucolica* | | |
|---|---|---|---|---|---|---|---|---|---|
| *x* | Type | f(x) | Expon | Type | f(x) | Expon | Type | f(x) | Expon |
| 1 | dsss | 423 | 408.61 | dsss | 344 | 321.89 | ddss | 107 | 100.60 |
| 2 | ddss | 338 | 359.17 | dsds | 262 | 278.69 | dsss | 90 | 90.28 |
| 3 | dsds | 325 | 315.71 | ddss | 249 | 241.30 | dsds | 79 | 81.01 |
| 4 | sdss | 297 | 277.51 | sdss | 213 | 208.92 | dssd | 64 | 72.70 |
| 5 | ddds | 229 | 243.94 | dssd | 152 | 180.89 | sdss | 63 | 65.24 |
| 6 | ssss | 191 | 214.42 | ddds | 136 | 156.61 | ddsd | 59 | 58.55 |
| 7 | ssds | 170 | 188.48 | ssss | 128 | 135.60 | ddds | 57 | 52.54 |
| 8 | dssd | 167 | 165.67 | ssds | 119 | 117.40 | dsdd | 43 | 47.15 |
| 9 | sdds | 167 | 145.63 | sdds | 111 | 101.65 | sdds | 43 | 42.31 |
| 10 | ddsd | 136 | 128.01 | ddsd | 105 | 88.01 | ssss | 40 | 37.97 |
| 11 | dsdd | 117 | 112.52 | sdsd | 88 | 76.20 | sdsd | 38 | 34.07 |
| 12 | sdsd | 106 | 98.91 | dsdd | 70 | 65.97 | ssds | 29 | 30.58 |
| 13 | sssd | 102 | 86.94 | sssd | 56 | 57.12 | dddd | 27 | 27.44 |
| 14 | dddd | 66 | 76.42 | ssdd | 52 | 49.46 | sssd | 26 | 24.62 |
| 15 | sddd | 60 | 67.17 | dddd | 49 | 42.82 | ssdd | 23 | 22.10 |
| 16 | ssdd | 58 | 59.05 | sddd | 42 | 37.07 | sddd | 21 | 19.83 |
| | a = 464.8523<br>b = 0.1290<br>$R^2$ = 0.9822 | | | a = 371.7745<br>b = 0.14409<br>$R^2$ = 0.9763 | | | a = 12.1025<br>b = 0.1083<br>$R^2$ = 0.9803 | | |

**Table 3**
Fitting the exponential function to hexameters: other Latin authors
(Source: Drobisch 1868: 32; 1866: 93, 95; Best 2008)

| | Ennius, *Fragments* | | | Cicero, *Arat* (translation) | | | Ovidius, *Metamorphoses* | | |
|---|---|---|---|---|---|---|---|---|---|
| *x* | Types | f(x) | Exp | Types | f(x) | Exp | Types | f(x) | Exp |
| 1 | ssss | 64 | 50.69 | dsss | 92 | 96.33 | dssd | 78 | 87.56 |
| 2 | sdss | 39 | 45.55 | sdss | 77 | 80.92 | dsss | 76 | 75.47 |
| 3 | dsss | 39 | 40.94 | ssss | 74 | 67.98 | dsds | 63 | 65.05 |
| 4 | ssds | 35 | 36.80 | ddss | 74 | 57.10 | ddss | 60 | 56.07 |
| 5 | sssd | 25 | 33.07 | sdds | 40 | 47.97 | ddsd | 57 | 48.33 |
| 6 | dssd | 24 | 29.72 | ssds | 35 | 40.29 | dsdd | 54 | 41.66 |
| 7 | dsds | 24 | 26.71 | dsds | 34 | 33.85 | ddds | 45 | 35.91 |
| 8 | ddss | 24 | 24.01 | ddds | 33 | 28.43 | dddd | 33 | 30.95 |
| 9 | sdds | 23 | 21.57 | sdsd | 20 | 23.89 | sdss | 26 | 26.68 |
| 10 | ddds | 21 | 19.39 | sssd | 18 | 20.06 | sdsd | 21 | 23.00 |
| 11 | ddsd | 20 | 17.43 | dssd | 16 | 16.86 | sdds | 13 | 19.82 |
| 12 | dddd | 20 | 15.66 | ddsd | 16 | 14.16 | sddd | 11 | 17.08 |
| 13 | sdsd | 19 | 14.08 | ssdd | 10 | 11.89 | ssss | 6 | 14.73 |
| 14 | sddd | 15 | 12.65 | sddd | 9 | 9.99 | ssds | 6 | 12.69 |
| 15 | dsdd | 12 | 11.37 | dsdd | 8 | 8.39 | ssdd | 6 | 10.94 |
| 16 | ssdd | 10 | 10.22 | dddd | 4 | 7.05 | sssd | 5 | 9.43 |

| | | | |
|---|---|---|---|
| | a = 56.3997<br>b = 0.1068<br>$R^2$ = 0.8492 | a = 114.6713<br>b = 0.1743<br>$R^2$ = 0.9589 | a = 101.5840<br>b = 0.1486<br>$R^2$ = 0.9363 |

**Table 4**
Fitting the exponential function to hexameters: other Latin authors (II)
(Source: Drobisch 1866, 1868; Best 2008)

| | Lucrece, *De rer. nat.* | | | Catullus, 2 poems | | | Horace, *Satires* | | |
|---|---|---|---|---|---|---|---|---|---|
| *x* | Type | f(x) | Exp | Type | f(x) | Exp | Type | f(x) | Exp |
| 1 | dsss | 88 | 87.27 | dsss | 124 | 112.97 | dsss | 285 | 269.14 |
| 2 | ddss | 72 | 74.94 | sdss | 65 | 82.98 | sdss | 228 | 240.44 |
| 3 | sdss | 63 | 64.35 | ddss | 55 | 60.97 | dsds | 205 | 214.81 |
| 4 | ddds | 56 | 55.26 | dsds | 51 | 44.78 | ddss | 203 | 191.91 |
| 5 | dsds | 51 | 47.46 | ssss | 43 | 32.89 | ssss | 174 | 171.45 |
| 6 | ssss | 39 | 40.75 | ssds | 25 | 24.16 | ssds | 137 | 153.17 |
| 7 | sdds | 37 | 35.00 | dssd | 15 | 17.75 | dssd | 134 | 136.84 |
| 8 | dssd | 36 | 30.05 | ddds | 15 | 13.04 | ddds | 115 | 122.25 |
| 9 | ssds | 26 | 25.81 | sdsd | 8 | 9.58 | sdds | 108 | 109.22 |
| 10 | sssd | 21 | 22.16 | sdds | 7 | 7.04 | sssd | 102 | 97.58 |
| 11 | sdsd | 17 | 19.03 | ddsd | 6 | 5.17 | sdsd | 93 | 87.17 |
| 12 | ddsd | 17 | 16.34 | dsdd | 5 | 3.80 | ddsd | 92 | 77.88 |
| 13 | dsdd | 12 | 14.03 | sddd | 4 | 2.79 | dsdd | 79 | 69.58 |
| 14 | sddd | 10 | 12.05 | sssd | 3 | 2.05 | ssdd | 63 | 62.16 |
| 15 | dddd | 8 | 10.35 | ssdd | 3 | 1.50 | dddd | 48 | 55.53 |
| 16 | ssdd | 7 | 8.89 | dddd | 1 | 1.11 | sddd | 46 | 49.61 |
| | a = 101.6228<br>b = 0.1523<br>$R^2$ = 0.9903 | | | a = 153.7832<br>b = 0.3085<br>$R^2$ = 0.9617 | | | a = 301.2523<br>b = 0.1127<br>$R^2$ = 0.9809 | | |

**Table 5**
Fitting the exponential function to hexameters: other Latin authors (III)
(Source: Drobisch 1866; Best 2008)

| | Manilius, *Astronomica* | | | Persius, *Satires* | | | Juvenal, *Satires* | | |
|---|---|---|---|---|---|---|---|---|---|
| *x* | Type | f(x) | Exp | Type | f(x) | Exp | Type | f(x) | Exp |
| 1 | dsss | 93 | 85.68 | dsss | 118 | 110.79 | dsss | 85 | 80.19 |
| 2 | sdss | 67 | 73.67 | ddss | 96 | 92.75 | sdss | 67 | 69.90 |
| 3 | dsds | 60 | 63.35 | dsds | 68 | 77.65 | dsds | 64 | 60.94 |
| 4 | ddss | 57 | 54.47 | sdss | 62 | 65.01 | ddss | 51 | 53.12 |
| 5 | ssss | 48 | 46.84 | dssd | 48 | 54.43 | dssd | 40 | 46.31 |
| 6 | ssds | 34 | 40.27 | ddds | 39 | 45.57 | ssds | 38 | 40.37 |
| 7 | dssd | 33 | 34.63 | sdsd | 35 | 38.15 | ddds | 35 | 35.19 |
| 8 | ddsd | 30 | 29.78 | ddsd | 35 | 31.94 | ddsd | 29 | 30.67 |
| 9 | sdds | 28 | 25.60 | dsdd | 32 | 26.74 | sdsd | 26 | 26.74 |
| 10 | sdsd | 22 | 22.02 | ssss | 30 | 22.39 | ssss | 25 | 23.31 |
| 11 | ddds | 22 | 18.93 | ssds | 27 | 18.74 | sdds | 22 | 20.32 |
| 12 | dsdd | 19 | 16.28 | sdds | 19 | 15.69 | sddd | 20 | 17.72 |
| 13 | ssdd | 16 | 14.00 | dddd | 14 | 13.14 | dsdd | 19 | 15.44 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14 | sssd | 11 | 12.03 | sssd | 12 | 11.00 | ssdd | 14 | 13.46 |
| 15 | dddd | 11 | 10.35 | sddd | 9 | 9.21 | sssd | 13 | 11.74 |
| 16 | sddd | 9 | 8.90 | ssdd | 5 | 7.71 | dddd | 12 | 10.23 |
| | a = 99.6452<br>b = 0.1509<br>$R^2$ = 0.9780 | | | a = 132.3358<br>b = 0.1777<br>$R^2$ = 0.9705 | | | a = 91.9889<br>b = 0.1373<br>$R^2$ = 0.9824 | | |

**Table 6**
Fitting the exponential function to hexameter types: other Latin authors (IV)
(Source: Drobisch 1866; Best 2008)

| | Lucanus, *Pharsalia* | | | Silius Italicus, *Punica* | | | Valerius Flaccus, *Argonautica* | | |
|---|---|---|---|---|---|---|---|---|---|
| *x* | Type | f(x) | Exp | Type | f(x) | Exp | Type | f(x) | Exp |
| 1 | dsds | 98 | 94.06 | dsss | 75 | 73.42 | dsds | 131 | 111.68 |
| 2 | dsss | 83 | 79.03 | sdss | 63 | 65.30 | ddss | 75 | 93.21 |
| 3 | ddss | 59 | 66.39 | ssds | 54 | 58.08 | ddds | 64 | 77.80 |
| 4 | sdss | 58 | 55.77 | ssss | 53 | 51.66 | dsss | 63 | 64.93 |
| 5 | dssd | 39 | 46.85 | dsds | 47 | 45.95 | ddsd | 54 | 54.19 |
| 6 | ddds | 33 | 39.35 | ddss | 47 | 40.87 | dssd | 52 | 45.23 |
| 7 | ssds | 32 | 33.07 | sssd | 34 | 36.35 | dsdd | 49 | 37.75 |
| 8 | ddsd | 29 | 27.77 | sdsd | 32 | 32.33 | ssds | 30 | 31.51 |
| 9 | sdds | 28 | 23.33 | dssd | 28 | 28.75 | sdsd | 24 | 26.30 |
| 10 | sdsd | 23 | 19.59 | ddds | 26 | 25.57 | dddd | 24 | 21.95 |
| 11 | dsdd | 19 | 16.46 | sdds | 25 | 22.75 | sdss | 22 | 18.32 |
| 12 | ssss | 15 | 13.83 | ddsd | 24 | 20.23 | sdds | 21 | 15.29 |
| 13 | dddd | 13 | 11.61 | ssdd | 18 | 17.99 | ssss | 12 | 12.76 |
| 14 | sssd | 12 | 9.76 | dsdd | 16 | 16.00 | sddd | 11 | 10.65 |
| 15 | sddd | 11 | 8.20 | dddd | 11 | 14.24 | ssdd | 5 | 8.89 |
| 16 | ssdd | 8 | 6.88 | sddd | 7 | 12.66 | sssd | 3 | 7.42 |
| | a = 112.0032<br>b = 0.1743<br>$R^2$ = 0.9762 | | | a = 82.5477<br>b = 0.1172<br>$R^2$ = 0.9768 | | | a = 133.8160<br>b = 0.1808<br>$R^2$ = 0.9296 | | |

**Table 7**
Fitting the exponential function to hexameter types: other Latin authors (V)
(Source: Drobisch 1866, 1868; Best 2008)

| | Statius, *Thebais* | | | Claudianus, *Raptus Proserpinae* | | | Horace, *Epistulae* | | |
|---|---|---|---|---|---|---|---|---|---|
| *x* | Type | Freq | Exp | Type | Freq | Exp | Type | Freq | Exp |
| 1 | dsds | 83 | 82.70 | dsds | 102 | 105.27 | dsss | 237 | 226.17 |
| 2 | dsss | 76 | 71.77 | ddss | 83 | 86.83 | sdss | 198 | 206.03 |
| 3 | ddds | 57 | 62.29 | dsss | 75 | 71.62 | dsds | 189 | 187.69 |
| 4 | ddss | 53 | 54.06 | sdss | 67 | 59.07 | ddss | 168 | 170.98 |
| 5 | ssds | 43 | 46.92 | ssds | 51 | 48.72 | dssd | 147 | 155.76 |
| 6 | dsdd | 40 | 40.72 | ddds | 38 | 40.18 | ddsd | 132 | 141.89 |
| 7 | ddsd | 39 | 35.34 | sdds | 34 | 33.14 | ssss | 125 | 129.26 |
| 8 | sdss | 34 | 30.67 | dssd | 30 | 27.34 | ssds | 124 | 117.75 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 9 | sdds | 32 | 26.61 | sdsd | 24 | 22.55 | ddds | 109 | 107.26 |
| 10 | dssd | 24 | 23.10 | ddsd | 21 | 18.60 | sdds | 109 | 97.71 |
| 11 | dddd | 17 | 20.05 | dsdd | 14 | 15.34 | ddsd | 97 | 89.01 |
| 12 | sdsd | 16 | 17.40 | ssdd | 8 | 12.65 | dsdd | 94 | 81.09 |
| 13 | ssss | 14 | 15.01 | dddd | 5 | 10.44 | sssd | 89 | 73.87 |
| 14 | ssdd | 14 | 13.10 | sssd | 3 | 8.61 | ssdd | 59 | 67.29 |
| 15 | sssd | 10 | 11.37 | sddd | 3 | 7.10 | sddd | 54 | 61.30 |
| 16 | sddd | 8 | 9.81 | ssss | 2 | 5.86 | dddd | 36 | 55.84 |
| | a = 95.2947<br>b = 0.1417<br>$R^2$ = 0.9829 | | | a = 127.6315<br>b = 0.1926<br>$R^2$ = 0.9841 | | | a = 248.2765<br>b = 0.0932<br>$R^2$ = 0.9661 | | |

**Table 8**
Fitting the exponential function to hexameter types: Leibniz
(Source: Drobisch 1866, 1868; Best 2008)

| Leibniz, *Epicedium* | | |
|---|---|---|
| Type | Freq | Exp |
| dsss | 59 | 59.29 |
| ddss | 57 | 49.37 |
| dsds | 33 | 41.11 |
| ddds | 31 | 34.23 |
| sdss | 30 | 28.50 |
| dssd | 24 | 23.73 |
| sdds | 20 | 19.76 |
| dsdd | 16 | 16.46 |
| ssss | 13 | 13.70 |
| ddsd | 12 | 11.41 |
| sdsd | 10 | 9.50 |
| sddd | 10 | 7.91 |
| ssds | 10 | 6.59 |
| ssdd | 6 | 5.48 |
| dddd | 5 | 4.57 |
| sssd | 2 | 3.80 |
| a = 71.2106 , b = 0.1831<br>$R^2$ = 0.9645 | | |

**Table 9**
Fitting the exponential function to hexameter types: Greek authors
(Sources: Drobisch 1868: 44, 50, 57; Best 2008)

| | Homer, *Iliad* | | | Homer, *Odyssey* | | | Theocrit, *Idyll 1* | | |
|---|---|---|---|---|---|---|---|---|---|
| *x* | Type | Freq | Exp | Type | Freq | Exp | Type | Freq | Exp |
| 1 | dddd | 350 | 375.69 | dddd | 410 | 402.40 | dsdd | 31 | 28.23 |
| 2 | dsdd | 320 | 309.04 | dsdd | 323 | 328.90 | dddd | 21 | 21.82 |
| 3 | sddd | 296 | 254.22 | sddd | 277 | 268.82 | sddd | 13 | 16.87 |
| 4 | ddds | 196 | 209.12 | ddds | 185 | 219.72 | dssd | 11 | 13.04 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5 | sdds | 155 | 172.02 | ssdd | 176 | 179.59 | ssdd | 10 | 10.08 |
| 6 | ssdd | 149 | 141.50 | sdds | 161 | 146.78 | ddsd | 9 | 7.79 |
| 7 | dsds | 145 | 116.40 | dsds | 149 | 119.97 | ssds | 7 | 6.02 |
| 8 | ssds | 78 | 95.75 | ddsd | 92 | 98.06 | dsds | 5 | 4.65 |
| 9 | ddsd | 76 | 78.76 | ssds | 82 | 80.15 | sdds | 5 | 3.60 |
| 10 | sdsd | 73 | 64.79 | dssd | 71 | 65.51 | sssd | 5 | 2.78 |
| 11 | dssd | 56 | 53.30 | sdsd | 65 | 53.54 | sdsd | 4 | 2.15 |
| 12 | sdss | 28 | 43.84 | sssd | 35 | 43.76 | ddds | 2 | 1.66 |
| 13 | ddss | 22 | 36.06 | ddss | 34 | 35.77 | dsss | 1 | 1.28 |
| 14 | dsss | 21 | 29.67 | sdss | 20 | 29.24 | | | |
| 15 | sssd | 19 | 24.40 | dsss | 18 | 23.90 | | | |
| 16 | ssss | 8 | 20.07 | ssss | 5 | 19.53 | | | |
| | a = 456.7120<br>b = 0.1953<br>$R^2$ = 0.9743 | | | a = 492.3214<br>b = 0.2017<br>$R^2$ = 0.9858 | | | a = 26.5219<br>b = 0.2575<br>$R^2$ = 0.9515 | | |

**Table 10**
Fitting the exponential function to hexameter types: Theognis
(Source: Drobisch 1872: 10; Best 2008)

| x | Type | f(x) | Exp |
|---|---|---|---|
| 1 | dsdd | 117 | 117.10 |
| 2 | sddd | 99 | 95.26 |
| 3 | dddd | 78 | 77.48 |
| 4 | ssdd | 66 | 63.03 |
| 5 | dsds | 38 | 51.27 |
| 6 | dssd | 36 | 41.70 |
| 7 | sdds | 34 | 33.92 |
| 8 | ddds | 28 | 27.59 |
| 9 | ssds | 26 | 22.44 |
| 10 | ddsd | 25 | 18.26 |
| 11 | sdsd | 23 | 14.85 |
| 12 | sssd | 21 | 12.08 |
| 13 | ddss | 7 | 9.83 |
| 14 | sdss | 4 | 7.99 |
| 15 | dsss | 3 | 6.50 |
| 16 | ssss | 2 | 5.29 |
| | a = 143.9660<br>b = 0.2065<br>$R^2$ = 0.9731 | | |

As can be seen, the ranking is an appropriate method: the decrease of the ranked frequencies is regular. The parameter *b* is almost constant, this meaning that it can be replaced by a sort of regular, language-, author-, or genre-determined figure. Nevertheless, other languages in which the hexameter occurs may yield other values. In order to test the hypothesis, we use again Drobisch's (1968b) data and test the model on various German poems. The results are listed in the following tables.

**Table 11**
Fitting the exponential function to hexameter types:
*Reinecke Fucks* and *Hermann und Dorothea* by Goethe
(Drobisch 1868, 149, 152: Best 2009)

| | Goethe, *Reinecke Fuchs* | | | Goethe, *Hermann und Dorothea* | | |
|---|---|---|---|---|---|---|
| *x* | Type | f(x) | Exp | Type | f(x) | Exp |
| 1 | sdss | 204 | 198.04 | sdss | 200 | 193.53 |
| 2 | sdds | 181 | 160.65 | sdds | 149 | 160.51 |
| 3 | ssds | 96 | 130.33 | sdsd | 142 | 133.12 |
| 4 | ddss | 93 | 105.72 | ddss | 104 | 110.40 |
| 5 | sdsd | 86 | 85.77 | sddd | 98 | 91.56 |
| 6 | sddd | 71 | 69.58 | ddsd | 73 | 75.93 |
| 7 | ddds | 65 | 56.44 | ddds | 63 | 62.98 |
| 8 | dsds | 46 | 45.79 | dddd | 41 | 52.23 |
| 9 | dddd | 43 | 37.14 | ssds | 40 | 43.32 |
| 10 | ddsd | 42 | 30.13 | ssdd | 35 | 35.92 |
| 11 | ssdd | 38 | 24.44 | dsds | 34 | 29.79 |
| 12 | dsdd | 22 | 19.83 | dsss | 31 | 24.71 |
| 13 | ssss | 14 | 16.09 | dssd | 25 | 20.49 |
| 14 | sssd | 7 | 13.05 | dsdd | 25 | 16.99 |
| 15 | dsss | 6 | 10.59 | ssss | 15 | 14.09 |
| 16 | dssd | 5 | 8.59 | sssd | 11 | 11.69 |
| | a = 244.1189<br>b = 0.2092<br>$R^2$ = 0.9557 | | | a = 233.3580<br>b = 0.1871<br>$R^2$ = 0.9866 | | |

**Table 12**
Fitting the exponential function to hexameter types: other German texts (I)
(Drobisch 1868: 140, 143, 146; Best 2009)

| | Klopstock, *Messias* | | | Voss, *Odyssee* | | | Voss, *Luise* | | |
|---|---|---|---|---|---|---|---|---|---|
| *x* | Types | f(x) | Exp | Types | f(x) | Exp | Types | f(x) | Exp |
| 1 | sddd | 129 | 142.70 | dsds | 125 | 140.32 | sddd | 188 | 211.08 |
| 2 | sdds | 128 | 125.09 | sdds | 123 | 125.12 | dsdd | 173 | 185.84 |
| 3 | ddds | 125 | 109.66 | sddd | 114 | 111.57 | ddsd | 164 | 163.63 |
| 4 | dddd | 102 | 96.13 | dddd | 98 | 99.49 | dddd | 161 | 144.07 |
| 5 | dsds | 76 | 84.26 | dsdd | 96 | 88.71 | sdsd | 154 | 126.85 |
| 6 | ddss | 66 | 73.87 | ddds | 91 | 79.10 | ddds | 135 | 111.68 |
| 7 | dsdd | 65 | 64.75 | sdsd | 86 | 70.53 | sdds | 131 | 98.33 |
| 8 | sdsd | 60 | 56.76 | ddsd | 67 | 62.89 | dsds | 95 | 86.58 |
| 9 | sdss | 60 | 49.76 | sdss | 65 | 56.08 | ddss | 70 | 76.23 |
| 10 | ssds | 52 | 43.62 | ddss | 59 | 50.00 | sdss | 51 | 67.12 |
| 11 | ddsd | 46 | 38.23 | ssdd | 53 | 44.59 | dssd | 44 | 59.09 |
| 12 | ssdd | 45 | 33.52 | ssds | 39 | 39.76 | ssdd | 35 | 52.03 |
| 13 | dssd | 26 | 29.38 | dssd | 31 | 35.45 | ssds | 32 | 45.81 |
| 14 | dsss | 13 | 25.75 | dsss | 16 | 31.61 | dsss | 9 | 40.33 |
| 15 | ssss | 6 | 22.58 | sssd | 4 | 28.19 | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 16 | sssd | 3 | 19.79 | ssss | 1 | 25.13 | | | |
| | a = 162.7932<br>b = 0.1314<br>$R^2$ = 0.9334 | | | a = 157.3716<br>b = 0.1147<br>$R^2$ = 0.9045 | | | a = 239.7334<br>b = 0.1273<br>$R^2$ = 0.8910 | | |

It is to be noted that in the second poem by Voss, two types ("sssd" and "ssss") are missing.

**Table 13**
Fitting the exponential function to hexameter types: other German texts (II)
(Drobisch 1875: 9, 11; Best 2009)

| | Goethe, *Elegien* | | | Schiller, Compilation | | |
|---|---|---|---|---|---|---|
| *x* | Type | f(x) | Exp | Type | f(x) | Exp |
| 1 | sdss | 96 | 103.53 | sdds | 79 | 81.93 |
| 2 | sdds | 92 | 83.85 | sddd | 68 | 69.65 |
| 3 | sdsd | 72 | 67.91 | sdss | 65 | 59.21 |
| 4 | ssds | 51 | 55.00 | sdsd | 51 | 50.34 |
| 5 | ddss | 45 | 44.54 | ddds | 47 | 42.79 |
| 6 | sddd | 37 | 36.07 | dddd | 31 | 36.38 |
| 7 | ddds | 32 | 29.22 | ddss | 30 | 30.93 |
| 8 | ddsd | 28 | 23.66 | ddsd | 25 | 26.29 |
| 9 | dsds | 19 | 19.16 | dsds | 24 | 22.35 |
| 10 | dddd | 11 | 15.52 | ssds | 21 | 19.00 |
| 11 | ssdd | 8 | 12.57 | dsdd | 20 | 16.15 |
| 12 | ssss | 6 | 10.18 | dsss | 16 | 13.73 |
| 13 | sssd | 6 | 8.24 | dssd | 13 | 11.67 |
| 14 | dssd | 6 | 6.68 | ssdd | 8 | 9.92 |
| 15 | dsss | 5 | 5.41 | ssss | 1 | 8.44 |
| 16 | dsdd | 4 | 4.38 | sssd | 1 | 7.17 |
| | a =127.8306<br>b = 0.2109<br>$R^2$ = 0.9830 | | | a = 96.3798<br>b = 0.1624<br>$R^2$ = 0.9746 | | |

**Table 14**
Fitting the exponential function to hexameter types: other German texts (III)
(Drobisch 1875: 26, 28; Best 2009)

| | Goethe, *Distichen* (without *Elegien*) | | | Goethe, all *Distichen* | | |
|---|---|---|---|---|---|---|
| *x* | Type | f(x) | Exp | Type | f(x) | Exp |
| 1 | sdds | 92 | 78.52 | sdds | 184 | 178.94 |
| 2 | sdsd | 46 | 62.28 | sdss | 141 | 144.42 |
| 3 | sdss | 45 | 49.40 | sdsd | 118 | 116.55 |
| 4 | sddd | 44 | 39.18 | sddd | 81 | 94.07 |
| 5 | ddds | 24 | 31.08 | ssds | 73 | 75.92 |
| 6 | ddss | 24 | 24.65 | ddss | 69 | 61.27 |
| 7 | ssds | 22 | 19.55 | ddds | 56 | 49.45 |
| 8 | dddd | 19 | 15.51 | ddsd | 42 | 39.91 |
| 9 | ddsd | 14 | 12.30 | dsds | 31 | 32.21 |
| 10 | dsds | 12 | 9.76 | dddd | 30 | 26.00 |
| 11 | dsdd | 12 | 7.74 | ssdd | 19 | 20.98 |
| 12 | ssdd | 11 | 6.14 | dsdd | 16 | 16.93 |
| 13 | dsss | 7 | 4.87 | dsss | 12 | 13.67 |
| 14 | ssss | 4 | 3.86 | ssss | 10 | 11.03 |
| 15 | dssd | 4 | 3.06 | dssd | 10 | 8.90 |
| 16 | sssd | 2 | 2.43 | sssd | 8 | 7.18 |
| | a = 98.9897<br>b = 0.2317<br>$R^2$ = 0.9243 | | | a = 221.7090<br>b = 0.2143<br>$R^2$ = 0.9915 | | |

Besides German poems, the exponential fit was also tested on *Václav Živsa*, a quantitative-meter idyll written by Czech author Svatopluk Čech. The analysed material comprised the first 100 lines of the poem.

**Table 15**
Fitting the exponential function to hexameter types: *Václav Živsa* by Svatopluk Čech

| | Svatopluk Čech, *Václav Živsa* | | |
|---|---|---|---|
| Rank | Types | Frequ | Expon |
| 1 | dssd | 14 | 14.27 |
| 2 | ddsd | 12 | 12.47 |
| 3 | ddds | 11 | 10.90 |
| 4 | ddss | 9 | 9.52 |
| 5 | sdsd | 9 | 8.32 |
| 6 | dsds | 9 | 7.27 |
| 7 | dsdd | 6 | 6.35 |
| 8 | dsss | 6 | 5.55 |
| 9 | sdss | 5 | 4.85 |
| 10 | sddd | 4 | 4.24 |
| 11 | ssds | 4 | 3.71 |
| 12 | ssdd | 3 | 3.24 |
| 13 | sdds | 3 | 2.83 |

| | | | |
|---|---|---|---|
| 14 | dddd | 2 | 2.47 |
| 15 | sssd | 1 | 2.16 |
| 16 | ssss | 1 | 1.90 |
| | a = 16.3304, b = 0.1348, $R^2$ = 0.9715 | | |

Now, having this results, we may conjecture that not only the hexameter (= a sequence of feet), but any other poem may abide by this regularity.

The greater the parameter *b* is, the more concentrated the hexameter is to a smaller number of line-types, i.e., the more rhythmically monotonous the poem is. Again, one could select the parameters *b* from individual poems, and order them to see at least the tendency. The results collected from the above hexameters is presented in Table 16.

**Table 16**
Parameters of the exponential functions in the hexameters of pieces of study
(ranked according to the decreasing parameter *b*)

| Poet | Work | a | b |
|---|---|---|---|
| Catullus | 2 poems | 153.7832 | 0.3085 |
| Theokrit | *Idyll 1* | 26.5219 | 0.2575 |
| Goethe | *Distichen* (without *Elegien*) | 98.9897 | 0.2317 |
| Goethe | All *Distichen* | 221.709 | 0.2143 |
| Goethe | *Elegien* | 127.8306 | 0.2109 |
| Goethe | *Reinecke Fucks* | 244.1189 | 0.2092 |
| Theognis | *Elegische Dichtungen* | 143.966 | 0.2065 |
| Homer | *Odyssey* | 192.3214 | 0.2017 |
| Homer | *Iliad* | 456.712 | 0.1953 |
| Claudianus | *Raptus Proserpinae* | 127.6315 | 0.1926 |
| Goethe | *Hermann und Dorothea* | 232.3574 | 0.1871 |
| Leibnitz | *Epicedium* | 71.2105 | 0.1831 |
| Persius | *Satires* | 132.3358 | 0.1777 |
| Cicero | *Arat* (translation) | 114.6713 | 0.1743 |
| Lucanus | *Pharsalia* | 112.0032 | 0.1743 |
| Schiller | *Compilation* | 96.3797 | 0.1624 |
| Lucrece | *De rerum natura* | 101.6228 | 0.1523 |
| Valerius Flaccus | *Argonautica* | 99.6452 | 0.1509 |
| Manilius | *Astronomica* | 99.6452 | 0.1509 |
| Ovidius | *Metamorphoses* | 101.584 | 0.1486 |
| Vergil | *Georgica* | 371.7745 | 0.1441 |
| Statius | *Thebais* | 95.2947 | 0.1417 |
| Juvenal | *Satires* | 91.9899 | 0.1373 |
| Svatopluk Čech | *Václav Živsa* | 16.3304 | 0.1348 |
| Klopstock | *Messias* | 162.7932 | 0.1314 |
| Vergil | *Aeneis* | 464.8523 | 0.129 |
| Voss | *Luise* | 239.7334 | 0.1273 |

| | | | |
|---|---|---|---|
| Voss | *Odysee* | 157.3716 | 0.1197 |
| Silius Italicus | *Punica* | 82.5477 | 0.1172 |
| Horace | *Satires* | 301.2523 | 0.1127 |
| Vergil | *Bucolica* | 12.1025 | 0.1083 |
| Ennius | *Fragments* | 56.3997 | 0.1068 |
| Horace | *Epistulae* | 248.2765 | 0.0932 |

The hexameter takes into account the length of syllables. A foot is a super-syllable, i.e., a higher entity composed of syllables. The hexameter line is a kind of a hyper-syllable because it consists of feet. The individual types of lines can be ordered according to their frequencies, one can examine the distance between equal lines, or the number of dactyls and spondees in the line. We conjecture that there are fixed functions behind all of these entities and properties. Further, there is surely a relation between the parameters of the exponential function. All these properties may be studied in all languages using the hexameter, the evolution in one language can be analyzed, and the authors can be compared. Following the same principles, other verse systems can be investigated, too.

**Hexameter motifs**

Motifs have been introduced into linguistics by R. Köhler (2008, 2015). They may be quantitative, consisting of non decreasing numbers, or qualitative, in which a new motif begins with a hexameter type already present in the immediately previous motif, but no two types can be repeated. For example, if we have a sequence [A, B, C, D, A, B, G], then the first motif is [A, B, C, D], the second is [A], and the third is [B, G] – i.e., [B] cannot be placed in the second motif. In hexameters, we have always types consisting of 4 feet, which makes it simple to establish the motifs.

Now, length measured in terms of lines in a motif is a characteristic property of motifs and can easily be computed. If one has the observed values of motif lengths, one can characterize the rhythmic variegation of the poem. For example, the greater the mean of lengths, the more variegated the poem rhythmically is. In this way, various indicators can be interpreted.

This procedure can be applied to any type of poems, but one should also consider trochees, iambs, etc. Automatically, several questions arise, leading to various hypotheses: Is the given distribution of lengths characteristic for a given language, or is it general? – Is the given rhythmic character of a poem typical of a writer? – Is there a development in a language concerning the motif length? –

For the Czech hexameter, we obtain the distribution of lengths presented below. The mean of these lengths is 3.1212 – that means, on average, 3.1 motifs represent a different rhythm.

d = [DDDS, DSDD, SDDS, DDSD, DDDS, DDSD, DSDS, DDSS, SDSD, DSSD, DDDS, DSSD, DDSD, DDSD, SSDD, SDSD, DDSD, DSSS, SDSS, DDDS, DSSD, DDSD, DSDS, DDSS, DSSD, DDDS, DDSD, DDSD, DDSD, SDSS, SSSD, DDSS, DDDS, SDSD, SSDD, DDSD, DSDS, SDSD, DSDS, DSSS, DSSS, SDSD, SDDS, DDSS, DSSD, DSSD, DSSS, DDSS, SDDS, SSDS, SDSS, DSSD, DDDD, DSSD, DSSD, SDSS, SDSS, DDSS, DSSD, DDSS, DDDS, DDSD, SDSD, SDDD, DDDS, SSDD, SDSD, DSDD, DSSS, DSSD, DSDD, DSDD, DDSS, DSSD, DSDS, SDDD, DDDD, DDDS, SDSS, SDDD, DDDS, DDSS, DSSD, DSDS, DSDS, SDSD, SDSD, DSDS, DDSD, DSDD, SSDS, DDDS, SDDD, DSSS, DSSD, SSSS, SSDS, DSDD, SSDS, DSDS].

The motifs of the Czech text are presented in Table 17.

**Table 17**
The hexameter motifs in Svatopluk Čech's *Václav Živsa*

| |
|---|
| [DDDS, DSDD, SDDS, DDSD] |
| [DDDS] |
| [DDSD, DSDS, DDSS, SDSD, DSSD, DDDS] |
| [DSSD] |
| [DDSD] |
| [DDSD, SSDD, SDSD] |
| [DDSD, DSSS, SDSS, DDDS, DSSD] |
| [DDSD, DSDS, DDSS] |
| [DSSD, DDDS, DDSD] |
| [DDSD] |
| [DDSD, SDSS, SSSD, DDSS, DDDS, SDSD, SSDD] |
| [DDSD, DSDS] |
| [SDSD, DSDS, DSSS] |
| [DSSS] |
| [SDSD, SDDS, DDSS, DSSD] |
| [DSSD, DSSS] |
| [DDSS, SDDS, SSDS, SDSS, DSSD, DDDD] |
| [DSSD] |
| [DSSD, SDSS] |
| [SDSS, DDSS] |
| [DSSD, DDSS, DDDS, DDSD, SDSD, SDDD] |
| [DDDS, SSDD] |
| [SDSD, DSDD, DSSS, DSSD] |
| [DSDD] |
| [DSDD, DDSS, DSSD, DSDS, SDDD, DDDD, DDDS, SDSS] |
| [SDDD] |
| [DDDS, DDSS, DSSD, DSDS] |
| [DSDS, SDSD] |
| [SDSD] |
| [DSDS, DDSD, DSDD, SSDS, DDDS, SDDD, DSSS, DSSD, SSSS, SSDS] |
| [DSDD] |
| [SSDS, DSDS] |

Several functions can be used to express the given regularity. We apply only three of them. Needless to say, one should collect a number of data in order to come to a theoretical result.

**Table 18**
Lengths of hexameter motifs in *Václav Živsa*

| Length | Frequency | Exponential + 1 | Zipf-Alekseev + 1 | Menzerath |
|---|---|---|---|---|
| 1 | 10 | 10.18 | 9.94 | 10.04 |
| 2 | 7 | 6.84 | 7.31 | 6.92 |
| 3 | 5 | 4.72 | 4.72 | 4.92 |
| 4 | 4 | 3.36 | 3.21 | 3.54 |

| | | | | |
|---|---|---|---|---|
| 5 | 1 | 2.50 | 2.36 | 2.57 |
| 6 | 3 | 1.96 | 1.87 | 1.87 |
| 7 | 1 | 1.61 | 1.57 | 1.36 |
| 8 | 1 | 1.39 | 1.39 | 1.00 |
| 10 | 1 | 1.16 | 1.19 | 0.54 |
| | | a = 14.4347<br>b = 2.2108,<br>$R^2$ = 0.9459 | a = -0.0003<br>b = -0.7271<br>c = 8.9435<br>$R^2$ = 0.9458 | a = 13.5328<br>b = -0.1061<br>c = 0.2940<br>$R^2$ =0.9475 |

As can be seen, in this case the motifs abide by the same regularity as hexameter lines.
That means that in investigating a higher level of entities – here, the hexameter line is the lower level, the motifs of hexameter lines are the higher level –, one must first fit the same function (distribution) in order to find a hierarchy. The syllable is positioned in several hierarchies, and their research will take years. In the present paper, we simply showed some directions; evidently, the examination is not yet finished.

Moreover, besides the theoretical applications, there is a vast field of practical use of the counts in literary studies, as the texts can be ranged according to the proportion of spondaic and dactylic feet, and authors, poems, periods, and poetic styles can be evaluated on the basis of their rhythmical complexity.

## References

**Best, K.-H.** (2008). Zur Diversifikation lateinischer und griechischer Hexameter. *Glottometrics* 17, 43–50.

**Best, K.-H.** (2009) Zur Diversifikation deutscher Hexameter. *Naukovyj Visnyk Cernivec'koho Universytetu: Herman'ska filolohija. Vypusk* 431, 172–180.

**Drobisch, M. V.** (1866). Ein statistischer Versuch über die Formen des lateinischen Hexameters. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft über die Wissenschaften zu Leipzig. Philologisch-historische Classe* 18, 73–119.

**Drobisch, M. V.** (1868a). Weitere Untersuchungen über die Formen des Hexameters des Vergil, Horaz und Homer. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft über die Wissenschaften zu Leipzig. Philologisch-historische Classe* 20, 16–53.

**Drobisch, M. V.** (1868b). Über die Formen des deutschen Hexameters bei Klopstock, Voss und Goethe. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft über die Wissenschaften zu Leipzig. Philologisch-historische Classe* 20, 138–160.

**Drobisch, M. V.** (1872). Statistische Untersuchungen des Distichon (von Hrn. Dr. Hultgren). *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen* 24, 1–33.

**Drobisch, M. V**. (1875). Über die Gesetzmässigkeit in Goethe's und Schiller's Distichen. In: *Königlich-Sächsische Gesellschaft der Wissenschaften zu Leipzig, Philologisch-Historische Classe, Berichte über die Verhandlungen* 27, 8–34.

**Grotjahn, R. (ed.)** (1981). *Hexameter Studies*. Bochum: Brockmeyer.

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

**Köhler, R.** (2015). Linguistic motifs. In: Mikros, G. K., Mačutek, J. (eds.) (2015). *Sequences in Language and Text*, 89–108. Berlin / Boston: de Gruyter Mouton.

*Glottometrics 46, 2019, 98-104*

# Tense and Person in English: Modelling Attempts

*Michal Místecký*[1]
*Gabriel Altmann*

**Abstract.** In the article, ready-made data are used to find a simple model. The verbs are ordered in semantic classes, and for the ranked frequencies of tenses and verbal persons, some models are found. It is recommended to use as few parameters as possible. Here, the exponential and the Zipf-Alekseev functions are used, in one case the power function, too.

**Keywords:** *English, verbs, tenses, semantic classes, models*

## 1. Tense

In the present investigation, we try to find a model for the ranking of tenses, persons, and semantic verb classes in English as presented in the article by Levickij and Lučak (2005), and Scheibman (2001). The first authors analyze 9 fiction texts and several journalistic and scientific texts. They subdivide the tenses as follows:

1. Present Indefinite
2. Present Continuous
3. Present Perfect
4. Present Perfect Continuous
5. Past Indefinite
6. Past Continuous
7. Past Perfect
8. Past Perfect Continuous
9. Future Indefinite
10. Future Continuous
11. Future Perfect
12. Future Perfect Continuous
13. Future Indefinite in the Past
14. Future Continuous in the Past
15. Future Perfect in the Past
16. Future Perfect Continuous in the Past

They concentrate to the comparison of text types, but do not model the frequencies. Here, we shall try to solve this problem. First, we order the frequencies of individual tenses and fit the exponential function, in which the relative rate of change is considered in its relation to the previous class, i.e. –

[1] Michal Místecký, University of Ostrava; e-mail: MMistecky@seznam.cz.

$$\frac{dy}{(y-1)\,dx} = -a$$

a function with constant decrease. The result of the equation is

$$y = 1 + a * e^{-bx} .$$

The respective numbers are presented in Table 1.

**Table 1**
Frequencies of individual tenses in English texts
(Levickij and Lučak 2005)

| Rank | Fiction | | Scientific texts | | Journalistic texts | |
|---|---|---|---|---|---|---|
| | Frequency | Exponential | Frequency | Exponential | Frequency | Exponential |
| 1 | 5387 | 5407.77 | 2290 | 2285.17 | 1639 | 1626.08 |
| 2 | 2349 | 2220.38 | 447 | 486.72 | 685 | 722.76 |
| 3 | 726 | 912.01 | 159 | 104.29 | 309 | 321.56 |
| 4 | 269 | 374.95 | 136 | 22.96 | 189 | 143.37 |
| 5 | 263 | 154.50 | 31 | 5.67 | 118 | 64.23 |
| 6 | 258 | 64.01 | 16 | 1.99 | 37 | 29.08 |
| 7 | 218 | 26.86 | 3 | 1.21 | 20 | 13.47 |
| 8 | 201 | 11.62 | | | 17 | 6.54 |
| 9 | 20 | 5.36 | | | 13 | 3.46 |
| 10 | 18 | 2.79 | | | 4 | 2.09 |
| 11 | 13 | 1.73 | | | 3 | 1.49 |
| 12 | 7 | 1.30 | | | 1 | 1.22 |
| 13 | | | | | 1 | 1.10 |
| | a = 13171.7872<br>b = 0.8904<br>$R^2$ = 0.9933 | | a = 10741.5528<br>b = 1.5481<br>$R^2$ = 0.9956 | | a = 3658.9526<br>b = 0.8116<br>$R^2$ = 0.9973 | |

As can be seen, the simple exponential function fits the data satisfactorily, with a very high determination coefficient. In all data, we omitted the zero frequencies and used the exponential function with added 1. One can find other functions expressing the trends slightly better, but with the disadvantage of an extra, third parameter.

The authors subdivided the verbs in 20 classes – namely Exchange Verbs, Measure Verbs, Change of Ownership Verbs, Change of Position Verbs, Change of Physical State Verbs, Circumstance Verbs, Impact/Effect Verbs, Directed Motion Verbs, Verbs of Existence, Ingestion Verbs, Verbs of Mental Process, Load/Spray Verbs, Manner of Motion Verbs, Verbs of Ownership, Verbs of Perception and Communication, Position Verbs, Verbs of Removing, Orientation Verbs, Verbs of Psychological State, and Verbs of Sound Emission.

The authors analyze the occurrences of these classes only in fiction and ascribe to each class the tenses in which the verbs occurred. Again, each semantic class has a special trend, and the tenses are not equally distributed. Here, we can analyze the individual semantic classes and study the rank-order of tenses. We took into account only classes represented at least by three different tenses. In two cases (Change of Ownership Verbs and Exchange Verbs), the data were fitted by the exponential function (see above); in others, we used the usual Zipf-Alekseev function, defined in terms of a differential equation as

$$\frac{y'}{y-1} = \frac{a + k * \ln x}{x},$$

where $a$ is the state of the language, $k * \ln x$ is the contribution of the writer, and $x$ in the denominator is the general breaking of the movement. The resulting function is, after reparametrisation, –

$$y = c * x^{a+b*\ln x} + 1 .$$

The results are presented in Tables 2a–d.

**Tables 2a–d**
Rank-order of tenses of the semantically classified verbs
(Levickij and Lučak 2005)

| Rank | Existence | | Perception, Communication | | Circumstance | | Directed Motion | | Mental Process | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Freq | Comp | Freq | Comp | Freq | Comp | Freq | Comp | Freq | Comp |
| 1 | 1092 | 1092.90 | 1321 | 1321.45 | 736 | 736.50 | 657 | 654.76 | 493 | 493.03 |
| 2 | 572 | 563.11 | 327 | 319.38 | 412 | 409.46 | 118 | 149.72 | 405 | 404.83 |
| 3 | 103 | 149.12 | 122 | 134.58 | 200 | 197.22 | 103 | 74.95 | 38 | 40.57 |
| 4 | 100 | 38.66 | 55 | 71.97 | 77 | 98.99 | 75 | 49.55 | 18 | 3.83 |
| 5 | 49 | 11.22 | 54 | 44.05 | 58 | 52.76 | 59 | 37.57 | 13 | 1.24 |
| 6 | 31 | 4.01 | 41 | 29.44 | 47 | 29.81 | 35 | 30.82 | 10 | 1.02 |
| 7 | 30 | 1.96 | 40 | 20.95 | 23 | 17.77 | 26 | 26.60 | 7 | 1.00 |
| 8 | 29 | 1.33 | 18 | 15.62 | 21 | 11.15 | 22 | 23.75 | 5 | 1.00 |
| 9 | 6 | 1.12 | 7 | 12.09 | 5 | 7.35 | 6 | 21.73 | 1 | 1.00 |
| 10 | 2 | 1.05 | 3 | 9.65 | | | 5 | 20.25 | | |
| 11 | | | | | | | 3 | 19.13 | | |
| | a = 0.5130<br>b = -2.1221<br>c = 1091.8963<br>$R^2$ = 0.9915 | | a = -1.9956<br>b = -0.0817<br>c = 1320.4541<br>$R^2$ = 0.9992 | | a = -0.2430<br>b = -0.8736<br>c = 735.4976<br>$R^2$ =0.9980 | | a = -2.3967<br>b = 0.3759<br>c = 653.7559<br>$R^2$ =0.9897 | | a = 3.1488<br>b = -4.9555<br>c = 492.0263<br>$R^2$ = 0.9983 | |

| Rank | Ownership | | Manner of Motion | | Change of Ownership | | Change of Position | | Psychological State | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Freq | Comp | Freq | Comp | Freq | Exp | Freq | Comp | Freq | Comp |
| 1 | 177 | 177.26 | 283 | 282.69 | 126 | 123.57 | 157 | 156.90 | 108 | 108.00 |
| 2 | 135 | 133.26 | 27 | 34.86 | 30 | 43.75 | 23 | 25.74 | 71 | 70.98 |
| 3 | 24 | 33.42 | 22 | 15.07 | 26 | 23.83 | 19 | 12.11 | 9 | 9.28 |
| 4 | 21 | 7.87 | 17 | 9.87 | 26 | 15.49 | 6 | 8.13 | 3 | 1.78 |
| 5 | 19 | 2.50 | 17 | 7.80 | 20 | 11.08 | 6 | 6.43 | 2 | 1.08 |
| 6 | 18 | 1.35 | 9 | 6.81 | 10 | 8.44 | 6 | 5.55 | 2 | 1.01 |
| 7 | 7 | 1.09 | 7 | 6.31 | 3 | 6.70 | 3 | 5.04 | 2 | 1.00 |
| 8 | 1 | 1.02 | 7 | 6.06 | | | | | | |
| 9 | | | 2 | 5.97 | | | | | | |
| 10 | | | 2 | 5.99 | | | | | | |

| 11 | | | 1 | 6.08 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | a = 1.5121<br>b = -2.7792<br>c = 174.2580<br>$R^2$ = 0.9727 | | a = -3.6183<br>b = 0.8104<br>c = 281.6912<br>$R^2$ = 0.9955 | | a = 122.5740<br>b = -1.4981<br>$R^2$ = 0.9608 | | a = -3.0870<br>b = 0.6218<br>c = 155.9048<br>$R^2$ = 0.9966 | | a = 2.3215<br>b = -4.2331<br>c = 107.0028<br>$R^2$ = 0.9996 | |

| Rank | Position | | Change of Physical State | | Impact/ Effect | | Removal | | Ingestion | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Freq | Comp | Freq | Comp | Freq | Comp | Freq | Comp | Freq | Comp |
| 1 | 73 | 73.00 | 71 | 71.04 | 68 | 67.82 | 46 | 46.00 | 41 | 40.99 |
| 2 | 68 | 68.00 | 21 | 20.49 | 23 | 25.09 | 19 | 19.00 | 14 | 14.00 |
| 3 | 6 | 5.97 | 11 | 11.47 | 16 | 12.14 | 7 | 6.98 | 6 | 6.82 |
| 4 | 1 | 1.25 | 7 | 8.13 | 9 | 6.96 | 3 | 3.13 | 6 | 4.08 |
| 5 | 1 | 1.01 | 7 | 6.47 | 2 | 4.51 | 2 | 1.82 | 3 | 2.82 |
| 6 | 1 | 1.00 | 6 | 5.50 | 1 | 3.21 | | | 1 | 2.15 |
| 7 | | | 5 | 4.87 | 1 | 2.47 | | | 1 | 1.77 |
| 8 | | | | | 1 | 2.01 | | | | |
| | a = 3,8776<br>b = -5.7440<br>c = 71.9998<br>$R^2$ = 1.0000 | | a = -2.0432<br>b = 0.2853<br>c= 70.0376<br>$R^2$ = 0.9993 | | a = -1.2009<br>b = -0.3913<br>c = 66.8210<br>$R^2$ = 0.9897 | | a = -0.4425<br>b = -1.2690<br>c = 45.0004<br>$R^2$ = 1.0000 | | a =-1.3950<br>b = -0.3269<br>c = 39.9902<br>$R^2$ = 0.9948 | |

| Rank | Exchange | | Measure | |
|---|---|---|---|---|
| | Freq | Exp | Freq | Comp |
| 1 | 15 | 15.03 | 2 | 2.00 |
| 2 | 8 | 7.87 | 2 | 2.00 |
| 3 | 4 | 4.12 | 1 | 1.00 |
| 4 | | | 1 | 1.00 |
| | a = 28.7097<br>b = 0.6470<br>$R^2$ = 0.9995 | | a = 34.5737<br>b = -49.8792<br>c = 1.0000<br>$R^2$ = 1.0000 | |

As can be seen, up to "very regular" cases, all can be captured by the Zipf-Alekseev function. Needless to say, many other texts must be analyzed in order to obtain reliable results.

The fact that there is some regularity in the rank frequency of tenses in individual classes can be shown comparing the parameters *a* and *b* in Table 3. We omit the classes which are not Zipf-Alekseev. As can be seen (cf. Figure 1), the greater *a* is, the smaller *b* is, and the observed course is convex. The trend can be expressed by a modified Menzerathian function, in the form of

$$y = c * x * e^{-dx} - 2\,;$$

that is, the power of $x$ is 1, and since the function begins with a very small number and ends with a negative number, we reduce it by 2.

**Table 3**
The relation between parameters *a* and *b* in the rank-frequency ordering of tenses

| a | b | Comp |
|---|---|---|
| -3.6183 | 0.8104 | 1.21015 |
| -3.0870 | 0.6218 | 0.75580 |
| -2.3967 | 0.3759 | 0.15685 |
| -2.0432 | 0.2853 | -0.15368 |
| -1.9956 | -0.0817 | -0.19569 |
| -1.3950 | -0.3269 | -0.72986 |
| -1.2009 | -0.3913 | -0.90411 |
| -0.4425 | -1.2690 | -1.59261 |
| -0.2430 | -0.8736 | -1.77576 |
| 0.5130 | -2.1221 | -2.47759 |
| 1.5121 | -2.7792 | -3.42422 |
| 2.3215 | -4.2331 | -4.20732 |
| 3.1488 | -4.9555 | -5.02294 |
| 3.8776 | -5.7440 | -5.75437 |
| 34.5737 | -49.8792 | -49.87857 |
| c = -0.9254, d = -0.0117, $R^2$ = 0.9990 | | |
| | | |

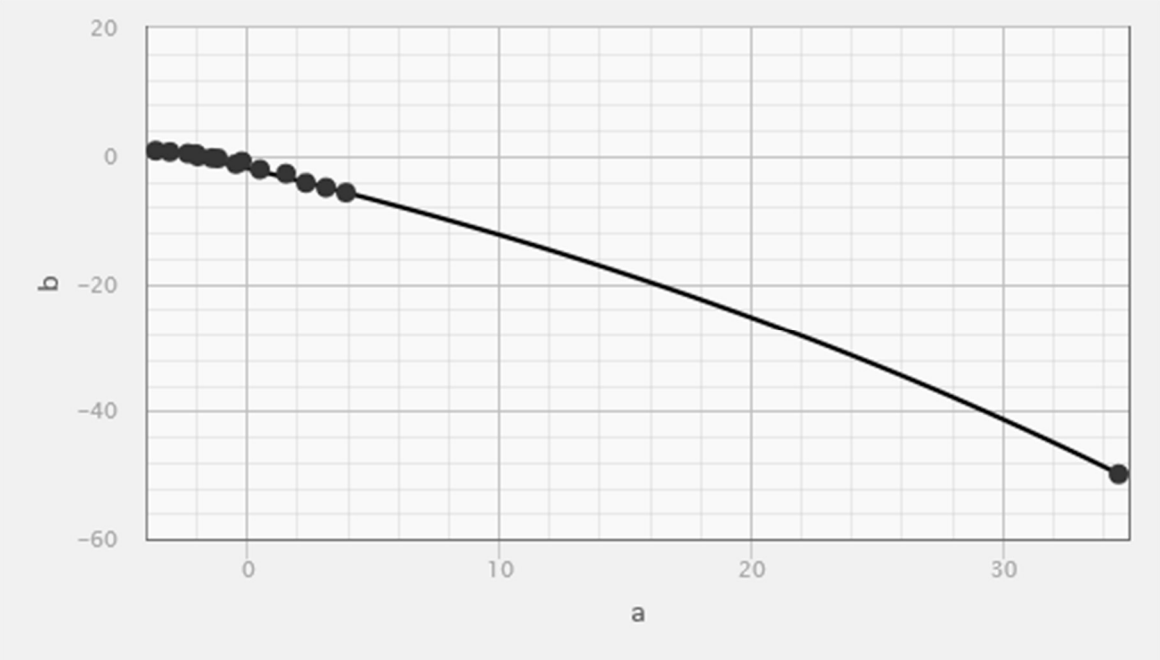


**Figure 1**. Dependence of *b* on *a* in the rank-frequency order of tenses

The computations show that in English prose, the use of tenses is somewhat regular. Needless to say, in other languages – e.g. isolating ones –, the situation would be very different. Besides, one must specify whether the tenses may be expressed by some separate words, or by affixes.

## 2. Persons

Another problem described by J. Scheibman (2001) is the use of persons with the verbs in semantic classes. It is to be noted that the semantic classes stated by Scheibman are different, namely: Cognition, Corporeal, Existential, Feeling, Material, Perception, Perception/ Relational, Possessive/Relational, Relational, Verbal Needless to say, the individual persons are not equal, e.g. the third-person plural forms are used more frequently than the first- or second-person plural ones.

Scheibman used the Corpus of Spoken American English (university of California) Again, it must be noted that the results cannot be used for comparison of text types because they represent spoken English. A difference could be discovered if one compared, e.g., scientific texts with fiction.

Again, we rank the frequencies without caring for the given person, in order to obtain a decreasing function. The results are presented in Tables 3a–b. In order to see whether there are some regularities, we use the Zipf-Alekseev function, and – occasionally – the exponential function, too.

**Tables 3a–b**

Ranking the use of persons with verbs of various classes (Scheibman 2001)

.

| Rank | Cognition | | Corporeal | | Existential | | Feeling | | Material | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Freq | Comp | Freq | Comp | Freq | Exp | Freq | Comp | Freq | Comp |
| 1 | 195 | 200.08 | 30 | 32.15 | 62 | 61.56 | 19 | 18.57 | 176 | 186.60 |
| 2 | 110 | 87.56 | 24 | 17.59 | 12 | 15.19 | 10 | 11.67 | 141 | 131.15 |
| 3 | 15 | 38.32 | 7 | 9.63 | 8 | 3.75 | 9 | 7.33 | 100 | 92.18 |
| 4 | 14 | 16.77 | 3 | 5.27 | 6 | 0.93 | 5 | 4.61 | 90 | 64.79 |
| 5 | 6 | 7.34 | 1 | 2.88 | 3 | 0.23 | 2 | 2.90 | 30 | 45.54 |
| 6 | | | 1 | 1.58 | | | | | 2 | 32.01 |
| | a = 457.1996<br>b = 0.8264<br>$R^2$ = 0.9606 | | a = 58.7515<br>b = 0.6029<br>$R^2$ = 0.9955 | | a = 249.4308<br>b = 1.3992<br>$R^2$ = 0.9746 | | a = 29.5435<br>b = 0.4645<br>$R^2$ = 0.9596 | | a = 265.4827<br>b = 0.3526<br>$R^2$ = 0.9045 | |

| Rank | Perception | | Perception/Relational | | Relational | | Verbal | |
|---|---|---|---|---|---|---|---|---|
| | Freq | Comp | Freq | Comp | Freq | Exp + 1 | Freq | Comp |
| 1 | 27 | 27.90 | 31 | 33.80 | 497 | 496.33 | 128 | 129.72 |
| 2 | 19 | 16.91 | 29 | 25.35 | 50 | 60.60 | 71 | 63.74 |
| 3 | 10 | 10.24 | 21 | 19.01 | 45 | 8.17 | 22 | 31.32 |
| 4 | 6 | 6.21 | 16 | 14.25 | 41 | 1.86 | 21 | 15.39 |
| 5 | 2 | 3.76 | 5 | 0.69 | 6 | 1.10 | 3 | 7.56 |
| 6 | | | | | 2 | 1.01 | | |
| | a = 46.0316<br>b = 0.5009<br>$R^2$ = 0.9796 | | a = 45.0741<br>b = 0.2878<br>$R^2$ = 0.8633 | | a = 4116.7741<br>b = 2.1176<br>$R^2$ = 0.9836 | | a = 263.9883<br>b = 0.7105<br>$R^2$ = 0.9812 | |

Only in the case of relational verbs, we used the exponential function with added 1, otherwise all results are acceptable.

Frequently, it is not easy to decide to which semantic group the verb belongs. One usually applies one's own language intuition, not caring for the environment of the verb. One presupposes that there are the same persons in all languages, but in many Austronesian

languages, there is even a difference between the inclusive and exclusive first person plural forms.

Evidently, there is still much work needed to decide which semantic classification is better than the other ones. The appropriateness of a classification can be decided only if the given state is set in a relation with other properties and, perhaps, inserted in the Köhlerian self-organizing cycle (2005).

## References

**Köhler, R**. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistics*. Berlin/New York: de Gruyter, 760–774.

**Levickij, V., Lučak, M.** (2005). Category of Tense and Verb Semantics in the English Language. *Journal of Quantitative Linguistics* 12(2–3), 212–235.

**Scheibman, J.** (2001). Local parameters of subjectivity in person and verb type in American English conversation. In: Bybee, J., Hopper, P. (eds.), *Frequency and the emergency of linguistics structure*. Amsterdam: John Benjamins Publishing Company, 61–89.

Other linguistic publications of RAM-Verlag:

## Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

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

23. E. Kelih, E. Knight, J. Mačutek, A. Wilson (eds.), *Issues in Quantitative Linguistics Vol 4.* 2016, 287 pp.
24. J. Léon, S. Loiseau (eds). *History of Quantitative Linguistics in France*. 2016, 232 pp.
25. K.-H. Best, O. Rottmann, *Quantitative Linguistics, an Invitation*. 2017, V+171 pp.
26. M. Lupea, M. Rukk, I.-I. Popescu, G. Altmann, *Some Properties of Rhyme.* 2017, VI+125 pp.
27. G. Altmann, *Unified Modeling of Diversification in Language*. 2018, VIII+119 pp.
28. E. Kelih, G. Altmann, *Problems in Quantitative Linguistics, Vol. 6.* 2018, IX+118 pp.
29. S. Andreev, M. Místecký, G. Altmann, *Sonnets: Quantitative Inquiries.* 2018, 129 pp.