Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200598

Similarities and Differences of Lithuanian Functional Styles: A Quantitative Perspective

Justina MANDRAVICKAITĖ [a,c,1] and Tomas KRILAVIČIUS [b,c]
[a] *Vilnius University, Lithuania*
[b] *Vytautas Magnus University, Lithuania*
[c] *Baltic Institute of Advanced Technology, Lithuania*

Abstract. We report an analysis of similarities and differences in terms of selected characteristics of 3 Lithuanian functional styles (FS): administrative, scientific, and publicistic. We combined 8 quantitative indicators and multivariate statistical analysis for this task. We also analyzed tendencies of indicators to be more or less pronounced in particular FS.

Keywords. Functional styles, Lithuanian, multivariate statistical analysis, quantitative indicators

1. Introduction

We report analysis of similarities and differences in terms of selected characteristics of 3 Lithuanian FS: administrative, scientific, and publicistic. 8 quantitative indicators and multivariate statistical analysis were chosen.

We define functional style as a variety of standard language that is characterized by domain, contents, functions, stylistic devices, and linguistic means [1]. Although there are 5 FS in the Lithuanian language: colloquial, administrative, fictional, publicistic and scientific (colloquial and fictional styles were not included). Colloquial was not included, because we analyze only written language, and fictional because of the impact of the author's style [2].

2. Data

We use 3 corpora, one corpus per FS: a corpus of administrative style (A), a corpus of publicistic style (P), and a corpus of scientific style (S). Corpus A is based on the administrative part of the Corpus of the Contemporary Lithuanian Language [3]. Corpus P is based on delfi.lt corpus [4], consisting of news articles. Finally, Corpus S is based on the non-fiction part of Corpus of the Contemporary Lithuanian Language, consisting

[1] Corresponding Author: Justina Mandravickaitė, Vilnius University; Lithuania; Baltic Institute of Advanced Technology, Lithuania; E-mail: justina.mandravickaite@bpti.eu

of educational and popular science texts. The latter texts were supplemented with these summaries of doctoral dissertations. Thus, all in all, A has 5.8 million words (4,527 texts), S – 20.2 million words (1,025 texts) and P – 10.4 million words (13,450 texts).

3. Methods

3.1. Indicators as Features

As characteristics of FS, based on [5–7, 9] and others, we chose 8 indicators. They address different linguistic characteristics and have lower or almost no significant dependence on the length of text [5, 6]. Also, these indicators have mathematical as well as linguistic explanation, which leads to an easier interpretation of the results. Selected indicators are the following:

1. **Average Token Length (ATL)** is used as a simple readability measure in linguistics.
2. **Indicator *a*** measures proportion of high frequency (usually function words) and lower frequency words in a text [6], [7].
3. **Indicator** R_1 was developed as a measure of vocabulary richness that is focused on less frequent words [5, 7]. We use word forms instead of lemmas, thus R_1 is more likely to measure a diversity of less frequent word forms, which belong to the main parts of speech, such as verbs, nouns, adjectives, or adverbs.
4. **Relative Repeat Rate of McIntosh (**RR_{mc}**)** measures vocabulary concentration of the text [8, 9]. RR_{mc} is normalized RR [8, 9] and this is more suitable for comparison with other indicators.
5. **Moving Average Type-Token Ratio (MATTR)** is a modification of Type-Token Ratio (TTR), which is independent of text length [10, 11].
6. **Thematic Concentration (TC)** measures the degree a text is concentrated over its topic [8, 11, 12]. TC is based on thematic words – words that are normally less frequent, but in particular text has a frequency close to the most frequent words which are usually function words – which define topic of the text.
7. **Activity (Q)** indicator expresses the dynamism of the text in terms of proportion of verbs and adjectives [8, 11, 13].
8. **Verb Distances (VD)** measure how many words on average occur in the text between 2 consecutive verbs [8], which can be interpreted, in a simplified way, as measuring complexity of syntactic structure of the text [5].

3.2. Multivariate Statistical Analysis

To analyze similarities and differences of FS, non-parametric multivariate analysis of variance was applied. We chose a non-parametric variety of analysis because of a significant number of outliers as well as different amount of texts in the corpora [14]. We applied Kruskal-Wallis test in order to test whether administrative, scientific, and publicistic functional styles have statistically significant differences among each other. Dunn's test [15] was used to evaluate the differences between pairs of functional styles in terms of each indicator. We then calculated relative treatment effects scores to estimate the scope of differences [16].

Table 1. Results of Dunn test

Indicator	Corpora pair	*Z*-value	*p*-value (adapted to multiple comparisons)
ATL	A-S	18.36141	8.027132e-75
	A-P	98.41522	0.000000e+00
	S-P	32.20224	4.925667e-227
a	A-S	-14.32894	4.329066e-46
	A-P	-93.11607	0.000000e+00
	S-P	-33.71111	1.191995e-248
R_1	A-S	3.496798	0.001412636
	A-P	-91.122320	0.000000000
	S-P	-51.653576	0.000000000
RR_{mc}	**A-S**	**0.350791**	**1**
	A-P	-89.598949	0
	S-P	-47.500637	0
MATTR	A-S	-19.85568	2.952350e-87
	A-P	-101.12481	0.000000e+00
	S-P	-32.03546	1.050061e-224
TC	A-S	45.73392	0.000000e+00
	A-P	71.06957	0.000000e+00
	S-P	-11.34295	2.411528e-29
Q	A-S	18.98684	6.574064e-80
	A-P	-26.51339	2.038068e-154
	S-P	-34.17354	1.793499e-255
VD	A-S	17.86002	7.246323e-71
	A-P	77.51638	0.000000e+00
	S-P	21.74412	2.355814e-104

4. Results

Kruskal-Wallis test showed, that all the analyzed FS differ by all the indicators significantly ($p < 0.05$). Dunn's test revealed that differences between all pairs of FS in terms of each indicator were statistically significant, except for the A-S pair in terms of indicator RR_{mc} (see Table 1). To estimate the scope of these differences, the relative treatment effects scores of those differences are presented in Table 2. A higher score indicates a higher probability of higher values for certain indicator in the texts of certain FS, i.e.:

- higher ATL values indicate longer words (more difficult to read);
- higher a values indicate lesser proportion of high frequency words;
- higher R_1 values indicate higher diversity of less frequent word forms;
- higher RR_{mc} values indicate higher vocabulary concentration;
- higher MATTR values indicate on average higher numbers of unique word forms in comparison to all word forms;
- higher TC values indicate higer thematic concentration;
- higher Q values indicate more dynamic texts (more verbs in comparison to adjectives);
- higher VD values indicate more complex syntactic structure (longer distance between 2 consecutive verbs).

Thus, longer word forms are more probable in A than in S and P. However, in this case, S is closer to A than to P. For P, lower proportion of high frequency words is more probable than for A and S. However, A and S, in this case, are closer to each other than to P. Similarly, higher diversity of less frequent words and word forms is more probable in P, however, it is less probable in A and S, which, according to relative treatment effects score, have rather similar probability. Furthermore, higher vocabulary concentration is more probable in P, than in A and S, which have the same lower probability. Higher

Table 2. Relative treatment effects

Indicator	Corpus	Relative treatment effects
ATL	A	0.85
	S	0.67
	P	0.37
a	A	0.17
	S	0.31
	P	0.63
R_1	A	0.19
	S	0.15
	P	0.63
RR_{mc}	A	0.19
	S	0.19
	P	0.63
MATTR	A	0.14
	S	0.34
	P	0.64
TC	A	0.77
	S	0.31
	P	0.42
Q	A	0.42
	S	0.23
	P	0.55
VD	A	0.78
	S	0.60
	P	0.40

proportion of different words and word forms is more probable in P and less probable in A. In this case, S is more similar to A than to P. Also, higher thematic concentration is more probable in A and less in S. P stands in between, although is more similar to S than to A. Additionally, very dynamic texts are more probable in P, although this tendency is not highly pronounced as relative treatment effects score for A in terms of this indicator is not much lower. In this case, S have a lower probability of very dynamic texts. Finally, texts with more complex syntactic structure are more probable in A. S stands slightly closer to A than to P in this matter, while P has a lower probability for syntactically complex texts.

5. Conclusions and Future Plans

We report an analysis of similarities and differences in terms of certain characteristics of 3 Lithuanian FS: administrative, scientific, and publicistic. We combined 8 quantitative indicators and multivariate statistical analysis for this task. Administrative and scientific style are closer each other in terms of indicators ATL, a, R_1, RR_{mc}, MATTR and VD. Administrative and publicistic functional styles are closer to each other in terms of indicator Q. Scientific and publicistic FS are closer to each other in terms of indicator TC.

Our future plans include experimenting with different variety of quantitative indicators as well as cross-lingual comparison in terms of scope of characteristics of FS. Future plans also include some practical applications, such as automatic text classification according to FS.

References

[1] Župerka KR. Stilistika. III pataisytas ir papildytas leidimas. Šiauliai: VšĮ Šiaulių univ. leid. 2012.

[2] Bitinienė A. Grožinio stiliaus prozos tekstų tiesioginės kalbos sakinio ilgis ir struktūra. Kalbotyra. 2001;50:17-28.
[3] Kovalevskaitė J. Dabartinės lietuvių kalbos tekstynas–10 metų kaupimo ir naudojimo patirtis. Prace Baltystyczne. 2006; 3:231-41.
[4] Bumbulienė I, Mandravickaitė J, Boizou L, Krilavičius T. An overview of Lithuanian internet media n-gram corpus. In CEUR Workshop Proc.: SYSTEM 2017, Proc. of the symposium for Young Scientists in Technology, Engineering and Mathematics, Kaunas, Lithuania, Apr 28, 2017. Aachen: CEUR-WS, 2017, Vol. 1853 2017.
[5] Kubát M. Kvantitativní analýza žánrů. Disertační práce Univerzita Palackého v Olomouci, Filozofická fakulta, Olomouc. 2015.
[6] Zörnig P, Kelih E, Fuks L. Classification of Serbian texts based on lexical characteristics and multivariate statistical analysis. Glottotheory. 2016 Jun 1; 7(1):41-66.
[7] Popescu II. Word frequency studies. Walter de Gruyter; 2009 Jun 2.
[8] Kubát M, Matlach V, Čech R. QUITA. Quantitative Index Text Analyzer. Lüdenscheid: RAM-Verlag. 2014.
[9] Popescu II, Cech R, Altmann G. Vocabulary richness in Slovak poetry. Glottometrics. 2011;22:62-72.
[10] Covington MA, McFall JD. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). Jrnl of quant. ling. 2010 May 1;17(2):94-100.
[11] Kubát M, Cech R. Quantitative Analysis of US Presidential Inaugural Addresses. Glottometrics. 2016 Jan 1;34:14-27.
[12] Čech R. Tematická koncentrace textu v češtině. ÚFAL, Ústav formální a aplikované lingvistiky; 2016.
[13] Zörnig P, Altmann G. Activity in Italian presidential speeches. Glottometrics. 2016 Jan 1;35:38-48.
[14] Field A, Miles J, Field Z. Discovering statistics using R. Sage pub.; 2012 Mar 31.
[15] Dinno A. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. The Stata Jrnl. 2015 Apr;15(1):292-300.
[16] Ellis AR, Burchett WW, Harrar SW, Bathke AC. Nonparametric inference for multivariate data: the R package npmv. Jrnl. of Stat. Soft. 2017 Jan 1;76(4):1-8.