# Similarities and Differences of Lithuanian Functional Styles: a Quantitative Perspective

2020


AUTHORS:

Justina Mandravickaitė
Vilnius University
Baltic Institute of Advanced Technology
*justina.mandravickaite@bpti.eu*

Tomas Krilavičius
Vytautas Magnus University
Baltic Institute of Advanced Technology
*tomas.krilavicius@bpti.eu*


BPTI
BALTIC INSTITUTE OF ADVANCED TECHNOLOGY

VYTAUTAS MAGNUS UNIVERSITY

## INTRODUCTION

- We report analysis of similarities and differences in terms of selected characteristics of 3 Lithuanian FS – administrative, scientific, and publicist. For that we chose 8 quantitative indicators as features and multivariate statistical analysis.
- Functional style (FS) according to Župerka (2012, 78):

***area of usage + content + functions + stylistic devices + linguistic mean***

| | Domain | Functions | Characteristics |
|---|---|---|---|
| **Administrative style (A)** | Official communication | Message, directive | Formal, formulaic language |
| **Scientific style (S)** | Scientific activities | Message | Accuracy, logic, objectivity |
| **Publicist style (P)** | Mass information | Message, appelative | Direct social assessment |

## DATA

| Corpus | No. of texts | No. of words |
|---|---|---|
| A | 4527 | 5.8 million |
| S | 1025 | 20.2 million |
| P | 13450 | 10.4 million |
| **Total** | **19002** | **36.4 million** |

## QUANTITATIVE INDICATORS USED AS FEATURES

| Indicator | Calculation | *Interpretation* |
|---|---|---|
| Average Token Length (ATL) | $\frac{1}{N}\sum_{i=1}^{N} x_i$ | Simple readability measure |
| *a* | $\frac{N}{h^2}$ | Evaluate size of area of the most frequent words in frequency table |
| $R_1$ | $1-\left(F(h)-\frac{h^2}{2N}\right)$ | Vocabulary richness measure |
| Relative Repeat Rate of McIntosh ($RR_{mc}$) | $\frac{1-\sqrt{RR}}{1-\frac{1}{\sqrt{V}}}$ | Vocabulary concentration measure |
| Moving Average Type-Token Ratio (MATTR) | $\frac{\sum_{i=1}^{N-L} V_i}{N(N-L+1)}$ | Topic deployment measure |
| Thematic Concentration (TC) | $\sum_{r'=1}^{T} 2\frac{(h-r')f(r')}{h(h-1)f(1)}$ | Measures the degree a text is concentrated over its topic |
| Activity (Q) | $\frac{Vrb}{Vrb.+Adj}$ | Measures dynamism of the text |
| Verb Distances (VD) | Average distance of verbs in a text | Measures complexity of syntactic structure of the text |

## MULTIVARIATE STATISTICAL ANALYSIS: METHODOLOGY

**non-parametric multivariate analysis of variance:**

- Kruskal-Wallis test to test whether A, S, and P have statistically significant differences among each other;
- Dunn's test to evaluate the differences between pairs of functional styles in terms of each indicator;
- relative treatment effects to estimate the scope of differences. S, and P have statistically significant differences among each other;
- Dunn's test to evaluate the differences between pairs of functional styles in terms of each indicator;
- relative treatment effects to estimate the scope of differences.

## RESULTS

**Results of Kruskal-Wallis test**

| Indicator | Chi-square | p-vertė |
|---|---|---|
| ATL | 10051 | < 2.2e-16 |
| *a* | 9127,3 | < 2.2e-16 |
| $R_1$ | 9883,2 | < 2.2e-16 |
| $RR_{mc}$ | 9307 | < 2.2e-16 |
| MATTR | 10572 | < 2.2e-16 |
| TC | 5496,2 | < 2.2e-16 |
| Q | 1656,9 | < 2.2e-16 |
| VD | 6139 | < 2.2e-16 |

**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 | 0.87 |
| | S | 0.60 |
| | P | 0.40 |

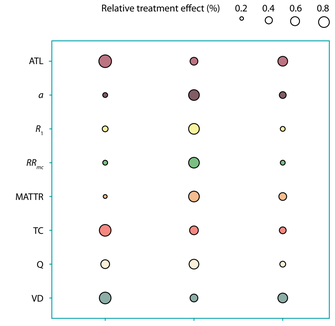
**Results of Dunn's test**

| Indicator | Corpora pair | Z-value | *p*-value *(aded 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 | 2.355813e-104 |

A higher relative treatment effects score indicates a higher probability of higher values for certain indicator in the texts of certain FS:

- higher ATL values indicate longer words *(more difficult to read)*;
- higher a values indicate lesser proportion of high frequency words;
- higher R1 values indicate higher diversity of less frequent word forms;
- higher RRmc values indicate higher vocabulary concentration;
- higher MATTR values indicate on average higher number unique word forms in comparison to all word forms;
- higher TC values indicate higher 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)*.

## SUMMARY




## CONCLUSIONS AND FUTURE PLANS

We report an analysis of similarities and differences in terms of certain characteristics of 3 Lithuanian FS – administrative, scientific, and publicist. We combined 8 quantitative indicators and multivariate statistical analysis for this task.

**Results revealed:**

- Administrative and scientific style are closer each other in terms of indicators ATL, a, R1, RRmc, MATTR and VD.
- Administrative and publicist FS are closer to each other in terms of indicator Q.
- Scientific and publicist FS are closer to each other in terms of indicator TC.

**Our future plans include**

- experimenting with different variety of quantitative indicators;
- cross-lingual comparison in terms of scope of characteristics of FS;
- practical applications, such as automatic text classification according to FS.: