Comparison of distance and similarity measures for stylometric analysis of Lithuanian texts

Daumantas Stanikūnas[1], Justina Mandravickaitė[2], Tomas Krilavičius[3]

[1]Department of Mathematics and Statistics, Vytautas Magnus University, Lithuania
Email: daumantas.stanikunas@fcis.vdu.lt

[2]Baltic Institute of Advanced Technology, Vilnius University, Lithuania
Email: justina@bpti.lt

[3] Baltic Institute of Advanced Technology, Vytautas Magnus University, Lithuania
Email: t.krilavicius@bpti.lt

***Abstract*—Constant developments in information and computer technologies make it possible to handle constantly increasing amount of data, thereby expanding the research possibilities. In this article, we discuss and compare distance and similarity measures used in stylometric analysis which could be applied to analyze Lithuanian texts. As corpus for the analysis, transcripts of parliamentary debates by two politicians of the Lithuanian Parliament were chosen. Furthermore, comparison of distance measures, stylometric analysis and visualization were performed. Objective of the experiment was to identify what measures would perform better when executing stylometric analysis of Lithuanian texts and explore where these differences in the performance occur. Summarizing the experiment results, the recommendations are as follow: number of Most Frequent Words used should be at least 1000, Eder's Simple Delta measure can be used in general stylometric analysis of transcriptions of parliamentary debates of Lithuanian Parliament, in a case when Most Frequent Words are limited to 2000, Binomial Index shows an increase in performance over Eder's Simple Delta and thus it is more suitable.**

***Index Terms*—stylometry; computational stylistics; parliamentary speech; R; statistical analysis; distance measure; similarity measure; data visualization**

I. Introduction

Stylometry refers to the study of linguistic style, usually to written language. It uses variety of statistical methods to analyze a text to determine the text's author. Common technique used is to calculate distances or similarities between texts and process the output using different visualization methods. There have already been significant number of experiments performed by various researchers in order to figure out what measures show better results in different cases of stylometric analysis. F. Jannidis and S. Evert analyzed three collections of novels in English, French and German languages and have shown that Cosine Delta measure outperforms all other measures on our three collections [2], [3], while J. Mandravickaitė in her research she explained that such measures like Burrows Delta would not work well with highly inflected languages (Latin, Polish) and suggested using Eder's Delta measure [10].

This paper presents an on-going experimental work on identifying the most suitable measures used in stylometry when analyzing Lithuanian texts. The objective of these experiments is to compare the performance of distance and similarity measures already used in stylometry and other fields of research [4], [5], [11] using R language with the focus on the transcripts of speeches of politicians in the Lithuanian Parliament.

These experiments cover a domain of transcriptions of parliamentary debates of Lithuanian Parliament which is only a small fraction of Lithuanian language, however they represent richness and variety of language quite well, and hence, we expect, that the results could be useful in analysis of other Lithuanian texts.

II. Data and Methods

A. Data Preparation

Data for the analysis will be taken from corpora, collected in ASTRA project. The data consists of transcriptions of parliamentary debates of Lithuanian Parliament. For this investigation we use transcripts of two politicians from the term 2008-2012. The criteria used for selecting the right data for comparison of measures were chosen in such a way that measures would provide the biggest difference between texts of different authors (or speakers, in our case). Considering this, authors for experiments were chosen of different gender and political standing (position and opposition) to strengthen the dissimilarity. This is a different type of approach compared to a similar research which was done by J. Kapočiūtė-Dzikienė where her goal was to identify best methods and features (MFW were not used) for authorship attribution based on machine learning [15]. Her approach was to analyze as similar data sets as possible, when authors are actually different for these data sets.

Automatic extraction of style applied to individual authors and groups of authors, http://dangus.vdu.lt/~jkd/?page_id=2

Table I: Distance/similarity measures and their formulas

Distance/Similarity measure	**Formula**
Manhattan Distance	$\sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Euclidean Distance	$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
Canberra Distance	$\sum_{i=1}^{n} \frac{\lvert x_i - y_i \rvert}{\lvert x_i \rvert + \lvert y_i \rvert}$
Cosine Distance	$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$
Burrows' Delta	$\frac{1}{n} \sum_{i=1}^{n} \left\lvert \frac{x_i - \mu_i}{\sigma_i} - \frac{y_i - \mu_i}{\sigma_i} \right\rvert$
Argamon's Linear Delta	$\frac{1}{n} \sum_{i=1}^{n} \sqrt{\left\lvert \frac{(x_i - y_i)^2}{\sigma_i^2} \right\rvert}$
Eder's Delta	$\frac{1}{n} \sum_{i=1}^{n} \left(\left\lvert \frac{x_i - y_i}{\sigma_i} \right\rvert \cdot \frac{n - n_i + 1}{n} \right)$
Eder's Simple Delta [9]	$\sum_{i=1}^{n} \left\lvert \sqrt{x_i} - \sqrt{y_i} \right\rvert$
Argamon's Quadratic Delta	$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (x_i - y_i)^2$
Bray-Curtis Dissimilarity	$\frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{\sum_{i=1}^{n} (x_i + y_i)}$
Kulczynski Distance	$\frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{\sum_{i=1}^{n} \min(x_i, y_i)}$
Jaccard Index	$\frac{2 \frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{\sum_{i=1}^{n} (x_i + y_i)}}{1 + \frac{\sum_{i=1}^{n} \lvert x_i - y_i \rvert}{\sum_{i=1}^{n} (x_i + y_i)}}$
Gower Similarity	$\frac{1}{n} \sum_{i=1}^{n} \frac{\lvert x_i - y_i \rvert}{\max_i - \min_i}$
Alternative Gower Similarity	$\frac{1}{n_0} \cdot \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Horn's modification of Morisita's Overlap Index	$\frac{2 \sum_{i=1}^{n} x_i y_i}{\left(\frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i\right)^2} + \frac{\sum_{i=1}^{n} y_i^2}{\left(\sum_{i=1}^{n} y_i\right)^2} \right) \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}$
Mountford Index	$\frac{1}{\alpha}$, where α is the parameter of Fisher's log-series
Binomial Index [5]	$\sum_{i=1}^{n} \frac{x_i \cdot \ln \frac{x_i}{2n} + y_i \cdot \ln \frac{y_i}{2n} - 2n \cdot \ln \frac{1}{2}}{2n}$

Table II: Where x_i and y_i are corresponding i values of vectors $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, n is the size of the compared vectors, σ_i and μ_i are standard deviation and mean of i values of all vectors used in comparison, n_i is queue number of i value in a vector (usually $n_i = i$), min_i and max_i are minimum and maximum i values between all compared vectors, n_0 is a number of pairs between corresponding X and Y vector values when at least one value in a pair is equal to 0.

The selected authors were Virginija Baltraitienė (Labor Party Political Group (Darbo partijos frakcija, DPF), who belonged to opposition at the given time) and Donatas Jankauskas (Homeland Union - Lithuanian Christian Democrat Political Group (Tėvynės sąjungos-Lietuvos krikščionių demokratų frakcija, TS-LKDF)). The mean word count for both authors are 190.39 and 203.96 accordingly, which means that their texts have to be concatenated so that one text object would contain around 5000 words. This is because Lithuanian language have many different words with many different word forms, and in this experiment we are not converting words to their base form. In addition, we will consider these 5000 words to be sufficient for a text object, because documents themselves contain as few as 100 words and it was shown by M. Eder that after more than 5000 words the quality of author attribution barely increases [13]. After this step, for further analysis, 16 text objects were created (8 for each politician). The analysis was performed using as features the most frequent words (MFW) in the sub-sample of 2 selected authors sorted in descending order by frequency. More statistics for the corpora and selected authors can be seen in Table III.

B. Methods

To evaluate the measures, Z-index was used, which in other words is a difference between means of standardized values.

$$z = |\mu_s - \mu_v| = \left| \frac{1}{m_s} \sum_{l=1}^{m_s} \frac{s_l - \mu}{\sigma} - \frac{1}{m_v} \sum_{k=1}^{m_v} \frac{v_k - \mu}{\sigma} \right|, \quad (1)$$

where μ_s and μ_v are out-group and in-group means of standardized values, m_s and m_v are out-group and in-group number of comparisons, s_l and v_k are out-group and in-group distance/similarity measurement for the corresponding comparison, μ and σ are the overall mean and standard deviation

Table III: Corpora statistics

Global Parameter	**Value**			
Period	March of 1990 – December of 2013			
Number of authors	147 (18 women and 129 men)			
Minimum word count in one text	100			
Selected Author	**Number of Texts**	**Number of Words**	**Number of Different Words**	**Mean Number of Words in One Text**
Virginija Baltraitienė	418	79584	11711	190.39
Donatas Jankauskas	468	95453	12981	203.96

of all measurements for comparisons, excluding comparisons when compared documents are the same.

In-group data set contains comparisons of documents of speeches by the same politician, while out-group data set contains comparisons of documents of speeches by different politicians. In this experiment we assumed that the bigger difference between means of standardized values, the better performance of certain measure was.

After inspecting results of the experiment, Cluster Analysis and Multidimensional Scaling was used to visualize relation among speeches of selected politicians, using distance measure with better performance and with regards to the quantity of MFW. In addition to Z-index, a dependency on the quantity of MFW (from 100 MWF to 5000 MFW) to the results was analyzed as well.

For stylometric analysis (calculating word frequencies, values of the distance measure, and plotting the relations among documents) R, free software environment for statistical computing and graphics, was used [1]. R language and its environment was chosen because it has all the necessary tools for textual data processing, computations, statistical analysis, visualization capabilities and good performance in general, e.g., one R script can be executed to provide all the required results from raw textual data without any additional software.

All distance/similarity measure computations and stylometric analysis were performed using "stylo" [9] and "vegdist" [16] scripts for R. In order to have a more efficient process, both scripts were merged together. This way all computations and analysis could be executed with one script.

III. Experimental Results

The objective of this investigation was to identify which measures perform better with different number of MFW in stylometric analysis of Lithuanian texts. All the evaluated measures are presented in Table II. Hierarchical Clustering [6], Multidimensional Scaling [7] and Heat Map [8] were applied with parameters described in Table IV.

A. Experiment 1: Initial Experiments

Initial experiments where performed with 100, 1000 and 5000 MFW. In each case all distance measures were sorted according to the calculated Z-index. Naturally, a distance measure with the highest Z-index should have been considered the best, but in our experiments there were no clear consistency in the measures regarding their performances, e.g., the measure that showed better performance, i.e., higher Z-index with 100 MFW, would perform worse with 1000 and 5000 MFW and vice versa. To investigate further, a second experiment was executed.

B. Experiment 2: Analysis of Distance Measures

To find out exactly how every distance measure behaves when increasing the number of MFW, a graph was plotted where all the measures were presented in Z-indexes, taking into consideration quantities of MFW taken for experimentation, see Fig. 1. All Z-indexes were calculated for 100, 200,..., 5000 MFW. In general, every distance measure was displayed as a function. A plot was generated with total of 50 points for every measure, which was enough to detect their behavior when quantity of MFW increased.

In the Fig. 1 we can see that most distance measures possess high values in Z-index throughout the plot, but four of them performed very poorly. These four distance (or similarity) measures are Alternative Gower Similarity, Horn's modification of Morisita's Overlap Index, Cosine Distance and Euclidean Distance (see Table II for details). In order to investigate further, we removed these distance measures from the experiment and concentrated on the remaining ones.

After removing distance measures with bad performance, we generated a new graph where remaining measures could be compared more precisely,see Fig. 2. We can see that up to around 1000 MFW, Z-index is always increasing in value and after this point it either decreases or shows similar results when MFW number is increased. Analyzing the plot further, we can observe three main groups of distance (or similarity) measures:

1) Measures which show very good performance until around 2000 MFW were reached. After that, the performance downgraded very fast. Distance measure that performed best was Binomial Index. Other distance measures which behaved similarly were:
   a) Argamon's Quadratic Delta,
   b) Burrows' Delta,
   c) Argamon's Linear Delta,
   d) Gower Similarity.
2) Measures which slowly reach their performance peak only at around 2000 MWF and then their performance very slowly declined. In general, these distance measures showed a very good performance and stability. In this

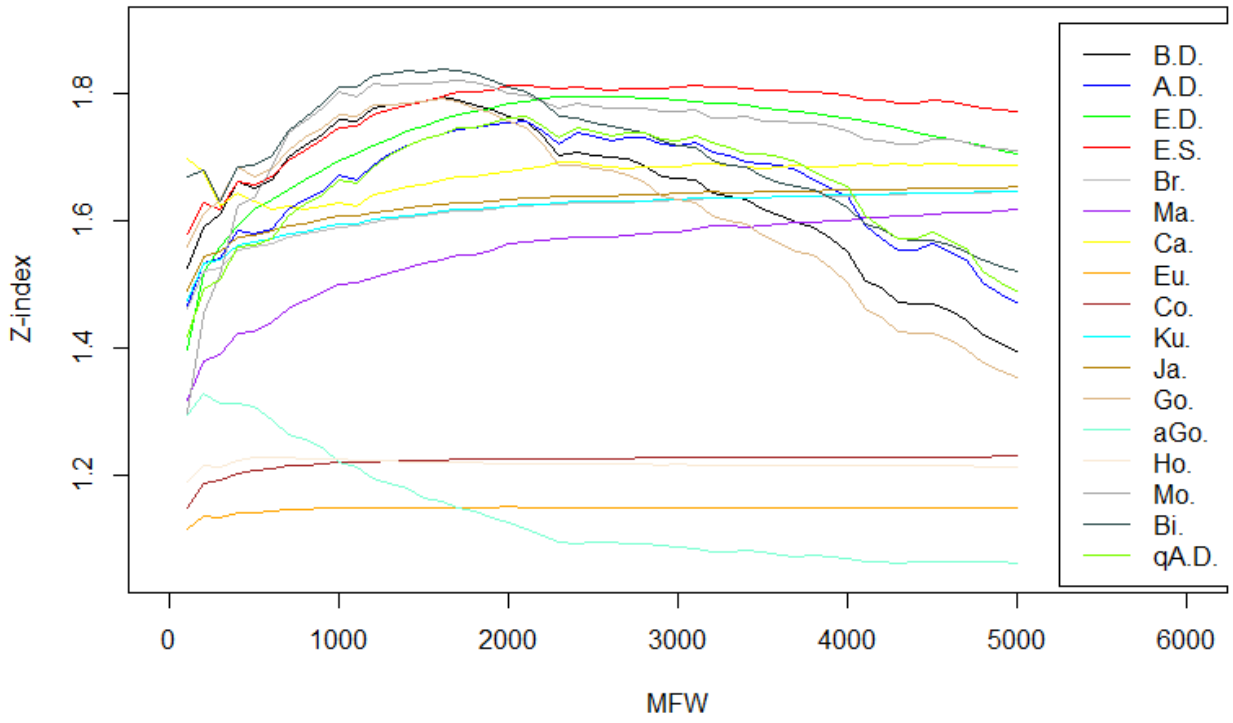


Figure 1: Difference between means of standardized values of in-group and out-group distances as a function of MFW amount used. Indicated for initial distance and similarity measures on the Lithuanian texts.

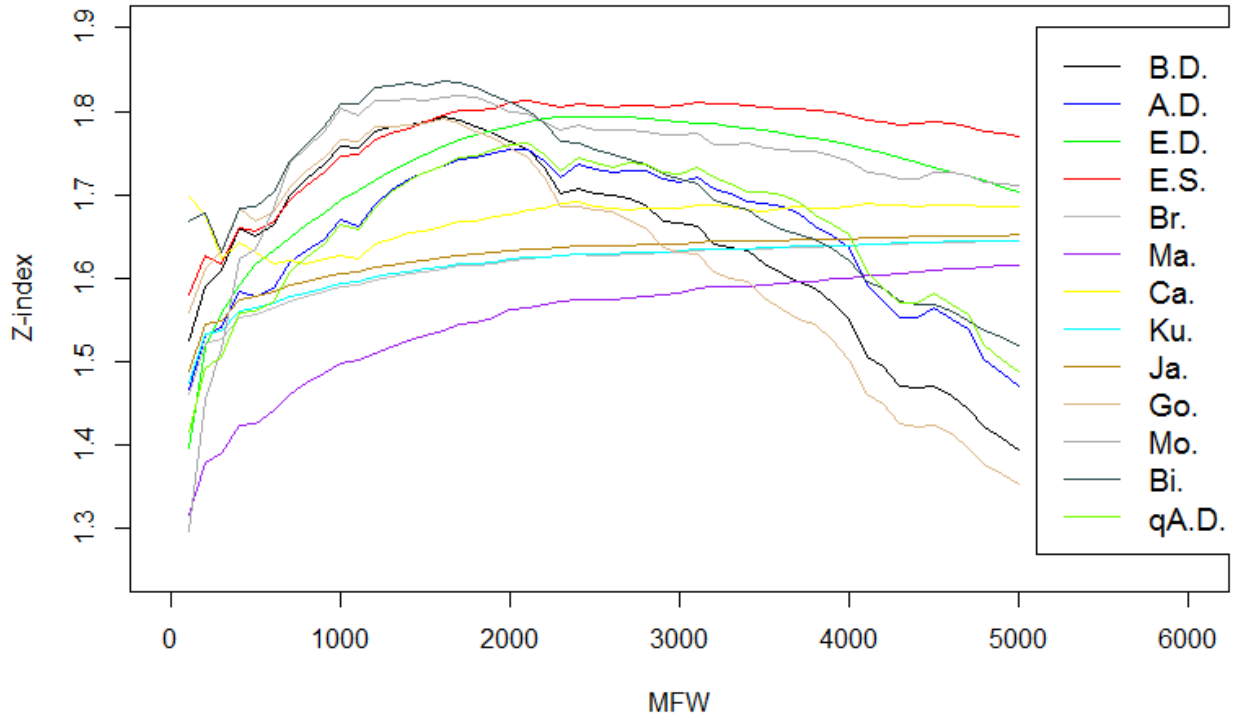


Figure 2: Difference between means of standardized values of in-group and out-group distances as a function of MFW amount used. Indicated for selected distance and similarity measures on the Lithuanian texts.

Table IV: Settings used for stylometric analysis

Parameter	Value
Corpus format	Plain text
Corpus language	English (ALL)
Analyzed features	Words
N-gram size	1
Most Frequent Words (MFW)	Amount that showed better results
Start at freq. rank	1
Culling min	0
Culling max	0
Analysis types	Cluster Analysis, Multidimensional Scaling (MDS), Heat map
Distance/Similarity measure	A measure that showed better results
Sampling	No sampling

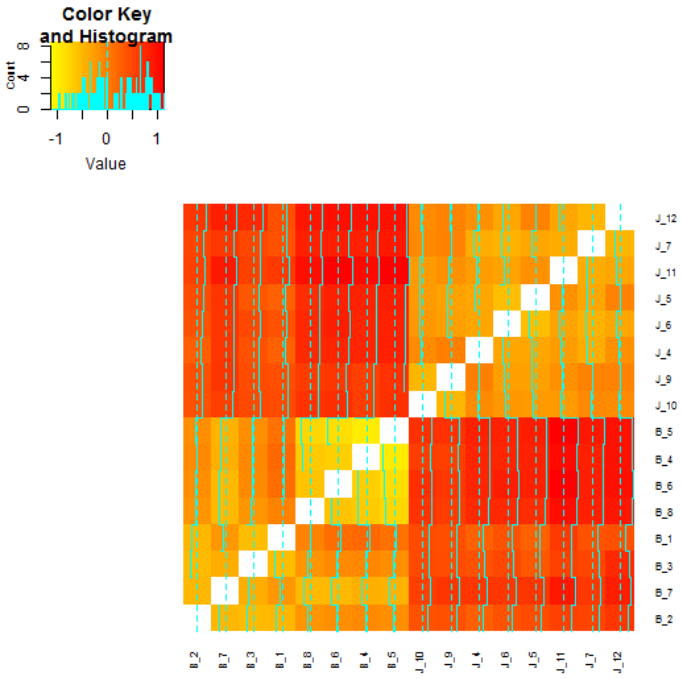


Figure 3: Heat map was generated using measurement matrix received with Eder's Simple Delta and 2000 Most Frequent Words. In the upper left corner, a color palette can be seen which is used to visualize data. Furthermore, histogram is drawn on top to show distribution of matrix values.

group, Eder's Simple Delta had the best results. Other measures which behaved similarly were:

a) Eder's Delta,
b) Mountford Index,
c) Canberra Distance.

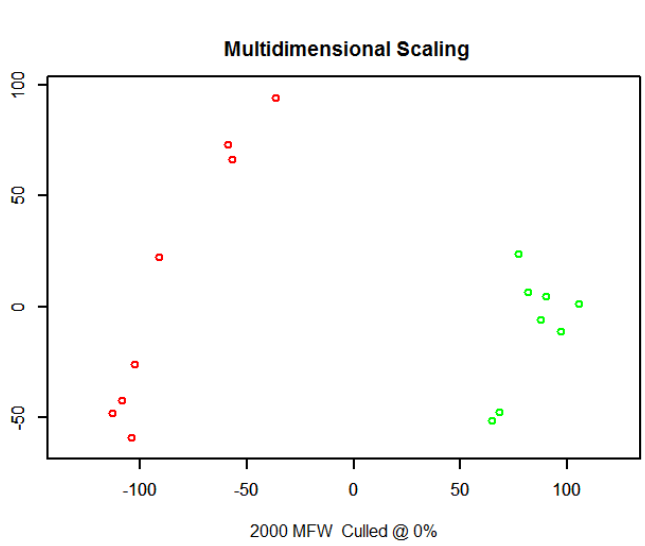
3) Measures which had a continuously improving performance together with the increasing MFW quantity, though this improvement was extremely gradual. In comparison to other groups, this group performed worse than the first and the second groups until reached 3000 MFW and after that it still performed worse than the second group, but it showed better performance over the first group. Distance measure that performed best in this group was Jaccard Index. Other measures which behaved similarly were:

a) Manhattan Distance,
b) Bray-Curtis Similarity,
c) Kulczynski Distance.

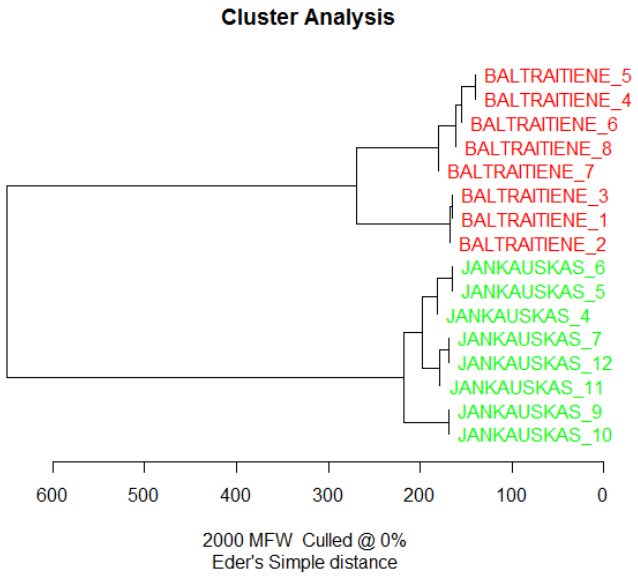
We noticed that Eder's Simple Delta measure performed really good: between the first group, only around 2-3% worse than the Binomial Index and between the second group, where Eder's Simple Delta showed the highest Z-index that was 1-3% better than Eder's Delta. So not only it could be used for bigger quantities of MFW, but for smaller ones as well. For this reason, in this experiment we considered Eder's Simple Delta as the best performing distance (or similarity) measure when analyzing Lithuanian texts in general.

C. Experiment 3: Exemplary Stylometric Analysis

We used Eder's Simple Delta measure for exemplary stylometric analysis in order to visualize the difference in lexical style (usage of words) between the selected members of the Lithuanian Parliament. For this analysis 2000 MFW was used, as it showed the best performance for the chosen distance measure. The first method used in order to map relations among the transcribed speeches made by before mentioned politicians was Hierarchical Cluster Analysis (see in Fig. 4b). In this research an agglomerative hierarchical clustering with Ward linkage, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function [14], was used. As a result, clusters'



(a) Multidimensional Scaling method.



(b) Hierarchical Cluster Analysis method.

Figure 4: Eder's Simple Delta measure and 2000 Most Frequent Words was used. Different colors represent different text authors.

hierarchy was generated and visualized as a dendrogram, where on the right side we have separate documents which are being linked into clusters according their similarity until all documents are merged into one cluster. The results showed clear differentiation between the authors which also contributed to our conclusion that Eder's Simple Delta performed well.

In Fig. 4a visualization and relation mapping among documents were performed using Multidimensional Scaling. The colors represent different authors of the parliamentary speeches, corresponding to the previous Figure (Fig. 4b). The division among parliamentary speeches made by different politicians was executed well, but in this case distribution of points was extended from the perspective of the vertical axis. This behavior might lead to different results if a lot more authors would be analyzed at once. In general, the results were good, which also contributed to our conclusion that the distance measure we used had an overall good performance.

In addition to the two methods we have already used for relation/position mapping of the documents as well as visualization, we tried to display the results with a Heat Map. Fig. 3 shows one of the simple ways to visualize stylistic differences among the documents without any further computations. In this example the color palette used for the plot was gradient from white to dark red, where white means a complete match. Since documents were sorted by author in the horizontal and vertical axes, the rectangle shapes formed out of light and darker colors. Light color showed that the documents were written (or speech spoken, in our case) probably by the same author and dark color – by a different author. Since the brightness of the color matched every pair of documents well enough, Heat Map could be considered to show positive results in regards to the Eder's Simple Delta measure.

To summarize, every method for visualization and relation/position mapping for stylometric analysis that we used did not object our statement that Eder's Simple Delta performed well in analysis of Lithuanian texts. But by no means this is the only measure that produces very good results. Other measures surely could also provide similar performance (as was seen in Fig. 2). Different corpus, parameters for text analysis, selection of Most Frequent Words - these options could still affect the performance of every distance (or similarity) measure. However, to reach solid conclusion, further research is needed.

IV. Conclusions

In this experiment we focused on Lithuanian texts, when corpus was composed of parliamentary speeches from Lithuanian Parliament. Hence, the following recommendations could be applied to the experimentation with the latter corpus.

1) Distance measure can be selected according to the quantity of the Most Frequent Words used in the analysis.
2) At least 1000 of Most Frequent Words should be used. After this point Z-index value either decreases or shows similar results.
3) If quantity of MFW does not exceed 5000 by a wide margin, Eder's Simple Delta measure performs well.
4) If Most Frequent Words are limited to 2000, Binomial Index shows an increase in performance over Eder's Simple Delta and thus it is more suitable in this case.

Finally, these recommendations can be applied for stylometric analysis of generic Lithuanian texts, but precautions must be taken and therefore we plan to experiment with texts belonging to different domains in the future.

REFERENCES

[1] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2008.

[2] F. Jannidis, S. Pielström, C. Schöch and T. Vitt, *Improving Burrows' Delta – An empirical evaluation of text distance measures*, Digital Humanities 2015, 2015.

[3] S. Evert, T. Proisl, C. Schöch, F. Jannidis, S. Pielström and T. Vitt *Explaining Delta, or: How do distance measures for authorship attribution work?*, 2015.

[4] H. S. Horn, *Measurement of "overlap" in comparative ecological studies*, 1966.

[5] M. J. Anderson and R. B. Millar, *Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern New Zealand*, 2004.

[6] L. Rokach and O. Maimon, *Clustering methods*, 2005.

[7] I. Borg and P. Groenen, *Modern Multidimensional Scaling: theory and applications*, 2005.

[8] *Heat map (heatmap)*, http://searchbusinessanalytics.techtarget.com/definition/heat-map, accessed: 2017-03-12.

[9] M. Eder, J. Rybicki and M. Kestemont, *Stylometry with R: a package for computational text analysis*, R journal, 2016.

[10] J. Mandravickaitė and T. Krilavičius, *Language usage of members of the Lithuanian Parliament considering their political orientation*, Deeds and Days, 2015.

[11] T. Krilavičius and V. Morkevičius, *Mining Social Science Data: a Study of Voting of the members of the Seimas of Lithuania by Using Multidimensional Scaling and Homogeneity Analysis*, Intellectual Economics, 2011.

[12] S. Argamon, *Interpreting Burrows's Delta: Geometric and Probabilistic Foundations*, 2007.

[13] M. Eder, *Does size matter? Authorship attribution, small samples, big problem*, Digital Humanities 2010, 2010.

[14] J. H. Ward Jr., *Hierarchical Grouping to Optimize an Objective Function*, Journal of the American Statistical Association, 1963.

[15] J. Kapočiūtė-Dzikienė, A. Utka and L. Šarkutė *Feature Exploration for Authorship Attribution of Lithuanian Parliamentary Speeches*, 2014.

[16] J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. Henry, H. Stevens and H. Wagner, *vegan: Community Ecology Package*, 2016.