Quantitative Analysis of Textual Genres: Comparison of English and Lithuanian

Justina Mandravickaitė Vilnius University, Lithuania Baltic Institute of Advanced Technology, Lithuania Email: justina@bpti.lt

Abstract—We report an ongoing study on quantitative characteristics of texts written in different genres. At this stage, we compared Lithuanian and English texts in terms of genres. We used 16 indices which describe frequency structure of text as well as indicate several other characteristics of written texts. Initial study showed significant differences of indices calculated for genre pairs of the same language. Hierarchical clustering revealed possible applications in using them as features for text categorization/classification by genre, though better results were achieved for Lithuanian texts.

Index Terms—quantitative genre analysis, frequency structure of text, vocabulary richness, stylometry, English, Lithuanian

I. INTRODUCTION

We report an ongoing study on quantitative characteristics of texts written in different genres. It has been suggested that genres add to familiarity and the shorthand of communication [1], [2], [3] and therefore resonate with people. Also, genres tend to shift in accordance to public opinion and reflect widespread culture of certain time [4]. From NLP perspective, genres are useful for text classification (e.g. [5]) and categorization (e.g. [6]), natural language generation (e.g. [7], [8]), etc.

At this stage, we present initial quantitative analysis of Lithuanian and English texts of different genres (or supergenres, in case of being more precise [9], as the texts were grouped into broad categories or genre groups; however, for the simplicity a term "genre" was used in this paper). As the main point of interest was frequency structure of text considering genre aspect, we used 16 indices proposed by [10], [11], [12] and implemented in QUITA - Quantitative Index Text Analyzer [13].

As we study textual genres wrt style, i.e., *fiction*, *press*, etc. style, we apply computational stylistics or stylometry. Stylometry is based on the two hypotheses:

- *human stylome hypothesis*, i.e., each individual has a unique style [14];
- unique style of individual can be measured [15], and thus stylometry allows gaining meta-knowledge [16], i.e., what can be learned from the text about the author gender [17], [18], age [19], psychological characteristics [20], political affiliation [19], etc.

Genre can be considered as a certain "style", thus we assumed that stylometric analysis could aid in our study of quantitative characteristics of genres. Tomas Krilavičius Vytautas Magnus University, Lithuania Baltic Institute of Advanced Technology, Lithuania Email: t.krilavicius@bpti.lt

II. CORPORA AND METHODS

A. Data Sets and Preprocessing

We used part of Corpus of the Contemporary Lithuanian Language [21] (\approx 1, 5 million words) and Freiburg-LOB Corpus of British English (F-LOB) (\approx 1 million words) [22] for our initial experiment. The composition of Lithuanian material is the following: Fiction (17%), Documents (21%), Scientific (21%) and Periodicals (31%). English material consists of Fiction (25%), General Prose (42%), Learned (16%) and Press (18%). Lithuanian genre category Scientific corresponds to English category Learned, while Lithuanian Periodicals corresponds to English category Press. More detailed constitution of F-LOB corpus by genres described in Table I. Part of the Corpus of the Contemporary Lithuanian Language we used for our study did not have such details available, only genre groups as described above.

As English texts were already concatenated according to their genre, only minimal preprocessing was performed, i.e., lines numbers and tags that marked textual structure were removed. For Lithuanian, as we had individual texts, to get around of "fingerprint" of individual authorship as much as possible, all the samples were concatenated into 4 large documents based on genre group (or super-genre), and then were partitioned into 5 parts each. Thus all in all for Lithuanian part of analysis we had 20 samples.

B. Features for Characterization of Genres

Most frequent words (MFW) as features are one the most popular solutions in stylometric analysis [23], [24], [25], [26], [27], [28] (usually, they coincide with function words [29], [30]). They are considered to be topic-neutral and perform well [31], [32], [33]. As our interest lied in frequency structure of the text as well as vocabulary richness taking genre aspect into consideration, for our experiment we applied 16 indices implemented in QUITA - Quantitative Index Text Analyzer [13]:

- Type-Token Ratio (TTR) ratio between the number of types and the number of tokens in a text, i.e. shows vocabulary variation in a text;
- h-point (h) a fuzzy boundary in the word frequency table where the rank is the same as the frequency;

Genre group	Category	Content of Category
Press	А	Reportage
	В	Editorial
	C	Review
General prose	D	Religion
	Е	Skills, trades and hobbies
	F	Popular lore
	G	Belles lettres, biographies, essays
	Н	Miscellaneous
Learned	J	Science
Fiction	K	General fiction
	L	Mystery and detective Fiction
	М	Science fiction
	Ν	Adventure and Western
	Р	Romance and love story
	R	Humor

Table I F-LOB CORPUS STATISTICS

- R_1 an indicator of vocabulary richness based on the h-point (h);
- Repeat Rate (RR) shows the degree of vocabulary concentration in a text, i.e. inverse measure of vocabulary richness;
- Relative Repeat Rate of McIntosh (RR_{mc}) the relative RR for better comparison with the other indices;
- Hapax Legomenon Percentage (HP) ratio between the number of tokens and number of hapax legomena, i.e. words that occur only once, in a text;
- Lambda (Λ) describes frequency structure of text, i.e. it is related to vocabulary richness, but also considers the relationship between neighbouring frequencies;
- Gini Coefficient (G) measure of statistical dispersion, in linguistics G is used as a measure for vocabulary richness;
- R_4 the reversed Gini coefficient;
- Curve length (L) as a lot of vocabulary richness measures are based on the curve of rank-frequency distribution, L is defined as the sum of the Euclidean distances between all neighbouring points on the curve;
- Curve length R Indicator (R) indicator of vocabulary richness derived from the curve length (L);
- Entropy (E) in linguistics, entropy expresses the degree of vocabulary concentration in the text;
- Adjusted Modulus (AM) frequency structure indicator, independent of text length;
- Writer's View (WV) indicator that is defined by the angle between the h-point and the ends of the rank-frequency distribution, i.e. the golden ratio;
- Average Tokens length (ATL) arithmetic mean of the lengths of tokens;
- Token Length Frequency Spectrum (TLFS) list of all token lengths in a text with their frequency.

Detailed formulas of the indexes (except for TLFS), based on [13] and [10], are presented in Table II. The values of indexes were standartized to make them comparable.

C. Distance Measures

Stylometry refers to the study of linguistic style, usually to written language. It uses variety of statistical methods, although common technique is to calculate distances or (dis)similarities between texts and process the output with different visualization methods. Studies have been performed in order to figure out what distance or similarity measures were more appropriate in different scenarios of stylometric analysis. For example, F. Jannidis and S. Evert found that Cosine Delta measure outperformed all other measures for novels written in English, French and German [36], [37]. Burrows's Delta distance is typically used for stylometric analysis as it proved effective for English and German [33], [26] as well. However, it was less successful for highly inflective languages, e.g., Latin and Polish [26]. Thus in such cases, especially when the most frequent words as features were used, application of Eder's Delta, i.e., a modified Burrows's Delta that gives more weight to the frequent features and rescales less frequent ones in order to avoid random infrequent features, was recommended [38]. Also, taking into consideration variety of possible text lengths, for Lithuanian texts Eder's Simple Delta and Binomial Index were useful (experiments were performed on the transcripts of plenary sittings of the Lithuanian Parliament) [39]. As we aim to compare the performance of distance or (dis)similarity measures already used in stylometry and other fields of research, e.g. ecology [40], biology [35], social sciences [41], we used the variety of them with formulas [39] presented in Table III.

D. Experimental Setup

For stylometric analysis (calculation of distance or (dis)similarity and plotting the relations among text samples) R, free software environment for statistical computing and graphics, was used [42], and its 2 packages - "stylo" [34] and

Index	Formula
Type-Token Ratio (TTR)	$\frac{V}{N}$
h-point (h)	r = f(r)
R_1	$1 - (F(h) - \frac{h^2}{2N})$
Repeat Rate (RR)	$\sum_{r=1}^{V} p_i^2$
Relative Repeat Rate of McIntosh (RR_{mc})	$\frac{1-\sqrt{RR}}{1-1/\sqrt{V}}$
Hapax Legomenon Percentage (HP)	$\frac{N_h}{N}$
Lambda (A)	$\frac{L(\log_{10} N)}{N}$
Gini Coefficient (G)	$\frac{1}{V}(V+1-2m_1')$
R_4	1-G
Curve length (L)	$\sum_{r=1}^{V-1} \sqrt{(f(r) - f(r+1))^2 + 1}$
Curve length R Indicator (R)	$1 - \frac{L_h}{L}$
Entropy (E)	$-\sum_{i=1}^{K} p_i l dp_i$
Adjusted Modulus (AM)	$\frac{\frac{1}{h}(f(1)^2 + V^2)^{1/2}}{\log_{10} N}$
Writer's View (WV)	$\frac{-[(h-1)(f_1-h)+(h-1)(V-h)]}{[(h-1)^2+(f_1-h)^2]^{1/2}[(h-1)^2+(V-h)^2]^{1/2}}$
Average Tokens length (ATL)	$\frac{1}{N}\sum_{i=1}^{N}x_i$

Table II INDEXES AND THEIR FORMULAS

Where V – number of types, N – number of tokens, r – rank/individual rank, f(r) – frequency of the rank, F(h) – cumulative relative frequency up to the h-point, h – h-point, p_i – individual probabilities, estimated by means of relative frequencies, RR – Repeat Rate, N_h – number of hapax legomena, L – arc length of the rank-frequency distribution, m_1 – average frequency distribution, G –Gini coefficient, f – individual frequency, L_h – curve length above h-point, K – inventory size, ld – logarithm to the base 2, f_1 – the highest frequency, x_i – individual length of the token.

"vegan" [43]. For the practical reasons these packages were merged together by [44].

For calculation of indexes (Type-Token Ratio (TTR), h-Point, Entropy, Average Tokens Length (ATL), R_1 , Repeat Rate (RR), Relative Repeat Rate of McIntosh (RR_{mc}), Lambda (Λ), Adjusted Modulus (AM), Gini's coefficient (G), R_4 , Hapax Legomena Percentage (HP), Curve Length (L), Writers View (WV), Curve Length Indicator (R), Token Length Frequency Spectrum (TLFS)) that were taken as features for our stylometric analysis of textual genres, QUITA - Quantitative Index Text Analyzer [13] was used. Also, to check statistical significance of the calculated indices in terms of genres, asymptotic *u-test* [45] was performed.

Then dissimilarity between the text samples was calculated using selected distances or similarity measures, and distance matrix was generated. Then, *hierarchical clustering* was applied to group samples by similarity [46], and dendrograms were used to visualize the results.

The goal of this study was to identify stylistic dissimilarities and map positions of the text samples in relation to each other, not to classify them by genre. Therefore hierarchical clustering with Ward linkage (it minimizes total variance within-cluster [46]) was chosen.

III. RESULTS

A. Statistical Significance of Indicators

Significance (asymptotic *u-test*) of calculated indices in terms of genres are provided in Table IV. The suffix "_LT" indicates Lithuanian part of experimental material, while "_EN" presents English part of our data. Most of calculated indicators achieved significance on at least some conditions. For Lithuanian part 3 indices (TTR, HP and R) were significant under all test conditions. There were no indices that did not achieved significance at any conditions. For English part only 1 indicator (ATL) was significant under all test conditions. Meanwhile, 2 indices (Lambda and HP) did not achieved significance at any conditions.

B. Stylometric Analysis

As it was already mentioned, typically Burrows's Delta distance is used for stylometric analysis [33], [26] with the

Distance/Similarity measure	Formula
Canberra Distance	$\sum_{i=1}^{n} \frac{ x_i - y_i }{ x_i + y_i }$
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 \frac{x_i - \mu_i}{\sigma_i} - \frac{y_i - \mu_i}{\sigma_i} \right $
Argamon's Linear Delta	$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\left \frac{(x_i-y_i)^2}{\sigma_i^2}\right }$
Eder's Delta	$rac{1}{n}\sum_{i=1}^{n}\left(\left rac{x_i-y_i}{\sigma_i} ight \cdotrac{n-n_i+1}{n} ight)$
Eder's Simple Delta [34]	$\sum_{i=1}^n \left \sqrt{x_i} - \sqrt{y_i} \right $
Bray-Curtis Dissimilarity	$\frac{\sum_{i=1}^{n} x_i - y_i }{\sum_{i=1}^{n} (x_i + y_i)}$
Kulczynski Distance	$\frac{\sum_{i=1}^{n} x_i - y_i }{\sum_{i=1}^{n} \min(x_i, y_i)}$
Jaccard Index	$\frac{2\frac{\sum_{i=1}^{i} x_i - y_i }{\sum_{i=1}^{i}(x_i + y_i)}}{1 + \frac{\sum_{i=1}^{i} x_i - y_i }{\sum_{i=1}^{i}(x_i + y_i)}}$
Gower Similarity	$\frac{1}{n}\sum_{i=1}^{n}\frac{ x_i-y_i }{\max_i-\min_i}$
Mountford Index	$\frac{1}{\alpha}$, where α is the parameter of Fisher's log-series
Binomial Index [35]	$\sum_{i=1}^{n} \frac{x_i \cdot \ln \frac{x_i}{2n} + y_i \cdot \ln \frac{y_i}{2n} - 2n \cdot \ln \frac{1}{2}}{2n}$

Table III DISTANCE/SIMILARITY MEASURES AND THEIR FORMULAS

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.

Table IV	
RESULTS OF SIGNIFICANCE TEST: G	GENRE PAIRS.

First variable	Second variable	Significant differences in indexes
Scientific_texts_LT	Documents_LT	TTR, h-Point, Entropy, R ₁ , RR, Lambda, AM, G, R ₄ , HP, L, WV, R.
Scientific_texts_LT	Fiction_LT	TTR, h-Point, Entropy, ATL, RR, Lambda, G, R4, HP, L, WV, R, TLFS.
Scientific_texts_LT	Periodicals_LT	TTR, h-Point, Entropy, ATL, RR, RR, RRmc, Lambda, G, R4, HP, L, R, TLFS.
Documents_LT	Fiction_LT	TTR, h-Point, ATL, R ₁ , Lambda, AM, G, R ₄ , HP, R, TLFS.
Documents_LT	Periodicals_LT	TTR, Entropy, ATL, R1, RR, RRmc, Lambda, AM, G, R4, HP, L, WV, R.
Fiction_LT	Periodicals_LT	TTR, h-Point, Entropy, ATL, RR, RR, RR, AM, HP, L, WV, TLFS.
Press_EN	Learned_EN	h-Point, Entropy, ATL, RR, RRmc, AM, L, WV.
Press_EN	Fiction_EN	Entropy, ATL, RR, RRmc, AM, L, WV, TLFS.
Press_EN	General_prose_EN	ATL, RR, RR_{mc} , R.
Learned_EN	Fiction_EN	ATL, RR, RR_{mc} , WV, R, TLFS.
Learned_EN	General_prose_EN	TTR, h-Point, Entropy, ATL, R ₁ , AM, G, R ₄ , L, WV, R.
Fiction_EN	General_prose_EN	Entropy, ATL, RR, RRmc, AM, L, WV, R, TLFS.

most frequent words (MFW) [23], [24], [25], [26], [27], [28] or function words (they usually occur among MFW [29], [30])as features. However, we achieved the best results with Eder's Delta distance measure (for English dataset; for formula, see III) and Argamon's Linear Delta distance measure (for Lithuanian dataset; for formula, see Table III). Though we experimented with all the distance or similarity measures described in Table III, due to limited space of the paper we present only the latter results (see Fig. 1 and 2).

Hierarchical Clustering [47] of an agglomerative type was used. Ward linkage, where choosing the pair of clusters to merge step-by-step is based on the optimal value of an objective function [48], was applied. This generated hierarchy of clusters, which was visualized as a dendrogram, that is, going from the right side separate documents were linked into clusters by their similarity till all the documents were merged into one cluster.

The results showed clear differentiation of text samples by genre for Lithuanian (all samples were clustered by genre correctly), while clustering of English dataset was somewhat less successful – some samples were attached to incorrect cluster. The reason could be language characteristics (indicators used as features react to the degree of inflection the language posess [10]), i.e. English is analytic language, while

Lithuanian - synthetic, and thus comparison of texts written in different languages becomes a non-trivial issue. Besides, it might have been influenced by grouping of text into genres and genre groups as it seems that this procedure was performed by following different criteria for our datasets in English and Lithuanian, e.g. for English part significantly bigger variety of genres was included into genre groups. Also, construction of comparable datasets for genre analysis might need to be more optimized in terms of sample lengths (even though part of indicators we used in this study was text-lengthindependent [13] and unsupervised machines learning (in this case - hierarchical cluster analysis) allows downscaling class imbalance problem)), samples themselves so that they would represent genre groups and genres best at the same time not forgetting to take authorship into consideration (we need to escape authorial "fingerprint" and concentrate of qualities of textual genres and the means to identify them).

To summarize, stylometric analysis combined with quantitative textual indicators that mark frequency structure or vocabulary richness of the text allowed us to map/position text samples by genre, though results were more successful for Lithuanian part of the experiment. Eder's Delta (for English) and Argamon's Linear Delta (for Lithuanian) distance measures provided the best results, however, by no means this is the only possible configuration. Other measures could also provide similar performance in different experimental setup, e.g. different corpora, parameters for text analysis, selection of features. However, to reach a more solid conclusion, further research is needed.

IV. CONCLUSION AND FUTURE WORK

We presented an ongoing work on quantitative analysis of texts written in different genres for English and Lithuanian. Textual genre in our study was perceived as certain "style" and thus stylometric analysis was performed.

- Features (frequency structure indicators and measures of vocabulary richness) used in this study seemed promising for characterization of genres as there were significant differences for genre pairs in terms of calculated indices.
- 2) As a part of stylometric analysis, 12 distance or (dis)similarity measures were experimented on. Out of them, Eder's Delta (for English dataset) and Argamon's Linear Delta (for Lithuanian dataset) provided the best results for our genre study.
- Cluster analysis allowed groupings of text samples by genre, though results were more successful for Lithuanian dataset in comparison to English one: all Lithuanian samples were grouped correctly.

However, for more substantial conclusions additional research is necessary. Thus we plan to extend this work to larger text collections and additional genres. More extensive study on textual indicators in terms of genre is important as well. We also plan to examine other languages to see whether similar effects found in this study would persist.

REFERENCES

- A. Tereszkiewicz, "Lead, headline, news abstract?-genre conventions of news sections on newspaper websites," *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, no. 129, p. 211, 2012.
- [2] J. Swales, *Genre analysis: English in academic and research settings*. Cambridge University Press, 1990.
- [3] A. J. Devitt, "Generalizing about genre: New conceptions of an old concept," *College composition and Communication*, vol. 44, no. 4, pp. 573–586, 1993.
- [4] C. R. Miller, "Genre as social action (1984), revisited 30 years later (2014)," *Letras & Letras*, vol. 31, no. 3, pp. 56–72, 2015.
- [5] Y. Kim and S. Ross, "Variation of word frequencies across genre classification tasks," 2007.
- [6] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genre and author," *Computational linguistics*, vol. 26, no. 4, pp. 471–495, 2000. [Online]. Available: http: //www.aclweb.org/anthology/J00-4001.pdf
- [7] O. Stock and C. Strapparava, "The act of creating humorous acronyms," *Applied Artificial Intelligence*, vol. 19, no. 2, pp. 137–151, 2005.
- [8] C. van der Lee, E. Krahmer, and S. Wubben, "Pass: A dutch data-to-text system for soccer, targeted towards specific audiences," in *Proceedings* of the 10th International Conference on Natural Language Generation, 2017, pp. 95–104.
- [9] G. Steen, "Genres of discourse and the definition of literature," *Discourse Processes*, vol. 28, no. 2, pp. 109–120, 1999.
- [10] I.-I. Popescu, Word frequency studies. Walter de Gruyter, 2009, vol. 64.
- [11] I.-I. Popescu, J. Mačutek, and G. Altmann, "Word forms, style and typology," *Glottotheory*, vol. 3, no. 1, pp. 89–96, 2010.
- [12] I.-I. Popescu, R. Čech, and G. Altmann, *The lambda-structure of texts*. Ram-Verlag Lüdenscheid, 2011.
- [13] M. Kubát, V. Matlach, and R. Čech, Studies in Quantitative Linguistics 18: QUITA-Quantitative Index Text Analyzer. RAM-Verlag, 2014.
- [14] H. Van Halteren, H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt, "New machine learning methods demonstrate the existence of a human stylome," *Journal of Quantitative Linguistics*, vol. 12, no. 1, pp. 65–77, 2005.
- [15] E. Stamatatos, "A survey of modern authorship attribution methods," Journal of the American Society for information Science and Technology, vol. 60, no. 3, pp. 538–556, 2009. [Online]. Available: http://www.clips.ua.ac.be/stylometry/Lit/Stamatatos_survey2009.pdf
- [16] W. Daelemans, "Explanation in computational stylometry," in Computational Linguistics and Intelligent Text Processing. Springer, 2013, pp. 451–462. [Online]. Available: http://www.clips.ua.ac.be/ ~walter/papers/2013/d13.pdf
- [17] K. Luyckx, W. Daelemans, and E. Vanhoutte, "Stylogenetics: Clusteringbased stylistic analysis of literary corpora," in *Proceedings of the* 5th International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 2006.
- [18] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, "Gender, genre, and writing style in formal written texts," *To appear in Text*, vol. 23, p. 3, 2003. [Online]. Available: http://www.lingcog.iit.edu/wp-content/ papercite-data/pdf/gendertext04.pdf
- [19] M. Dahllöf, "Automatic prediction of gender, political affiliation, and age in swedish politicians from the wording of their speeches - a comparative study of classifiability," *Literary and linguistic computing*, vol. 27, no. 2, pp. 139–153, 2012.
- [20] K. Luyckx and W. Daelemans, "Personae: a corpus for author and personality prediction from text," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC'08*), B. M. J. M. J. O. S. P. D. T. Nicoletta Calzolari (Conference Chair), Khalid Choukri, Ed. Marrakech, Morocco: European Language Resources Association (ELRA), may 2008, http://www.lrecconf.org/proceedings/Irec2008/.
- [21] E. Rimkuté, J. Kovalevskaité, V. Melninkaité, A. Utka, and D. Vitkuté-Adžgauskiené, "Corpus of contemporary lithuanian language-the standardised way," in *Human Language Technologies-The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT* 2010, vol. 219. IOS Press, 2010, p. 154.
- [22] M. Hundt, A. Sand, and R. Siemund, Manual of information to accompany the Freiburg-LOB Corpus of British English ('FLOB'). Albert-Ludwigs-Universität Freiburg, 1998.



Figure 1. Best clustering results for English data: Eder's Delta distance measure. The names of the samples in the cluster analysis were constructed as follows: genre-group_genre; see Table I for the details. As there was only one sample for Learned genre group, it was split into 2 equal samples: J1 and J2.



Figure 2. Best clustering results for Lithuanian data: Argamon's Linear Delta distance measure. The names of the samples in the cluster analysis were constructed as follows: genre-group_number-of-sample, where D = Documents, G = Fiction, M = Scientific texts, and P = Periodicals.

- [23] J. F. Burrows, "Not unles you ask nicely: The interpretative nexus between analysis and information," *Literary and Linguistic Computing*, vol. 7, no. 2, pp. 91–109, 1992.
- [24] D. L. Hoover, "Corpus stylistics, stylometry, and the styles of henry james," *Style*, vol. 41, no. 2, p. 174, 2007.
- [25] M. Eder, "Mind your corpus: systematic errors in authorship attribution," *Literary and linguistic computing*, p. fqt039, 2013. [Online]. Available: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/ mind-your-corpus-systematic-errors-in-authorship-attribution.1.html
- [26] J. Rybicki and M. Eder, "Deeper delta across genres and languages: do we really need the most frequent words?" *Literary* and linguistic computing, vol. 26, no. 3, pp. 315–321, 2011. [Online]. Available: http://dh2010.cch.kcl.ac.uk/academic-programme/ abstracts/papers/pdf/ab-688.pdf
- [27] M. Eder and J. Rybicki, "Do birds of a feather really flock together, or how to choose training samples for authorship attribution," *Literary and Linguistic Computing*, p. fqs036, 2012.
- [28] M. Eder, "Computational stylistics and biblical translation: How reliable can a dendrogram be," *The translator and the computer*, pp. 155–170, 2013.
- [29] J.-R. Hochmann, A. D. Endress, and J. Mehler, "Word frequency as a cue for identifying function words in infancy," *Cognition*, vol. 115, no. 3, pp. 444–457, 2010.
- [30] B. Sigurd, M. Eeg-Olofsson, and J. Van Weijer, "Word length, sentence length and frequency-zipf revisited," *Studia Linguistica*, vol. 58, no. 1, pp. 37–52, 2004.
- [31] P. Juola and R. H. Baayen, "A controlled-corpus experiment in authorship identification by cross-entropy," *Literary and Linguistic Computing*, vol. 20, no. Suppl, pp. 59–67, 2005.
- [32] D. I. Holmes, L. J. Gordon, and C. Wilson, "A widow and her soldier: Stylometry and the american civil war," *Literary and Linguistic Computing*, vol. 16, no. 4, pp. 403–420, 2001.
- [33] J. Burrows, "delta': A measure of stylistic difference and a guide to likely authorship," *Literary and Linguistic Computing*, vol. 17, no. 3, pp. 267–287, 2002.
- [34] M. Eder, J. Rybicki, and M. Kestemont, "Stylometry with r: a package for computational text analysis," *R Journal*, vol. 16, no. 1, 2016.
- [35] M. J. Anderson and R. B. Millar, "Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern new zealand," *Journal* of Experimental Marine Biology and Ecology, vol. 305, no. 2, pp. 191– 221, 2004.
- [36] F. Jannidis, S. Pielström, C. Schöch, and T. Vitt, "Improving burrows' delta-an empirical evaluation of text distance measures," in *Digital Humanities Conference*, 2015.
- [37] 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.
- [38] M. Eder, J. Rybicki, M. Kestemont, and M. M. Eder, "Package 'stylo'," 2014.
- [39] D. Stanikunas, J. Mandravickaite, and T. Krilavicius, "Comparison of distance and similarity measures for stylometric analysis of lithuanian texts," 2017.
- [40] H. S. Horn, "Measurement of" overlap" in comparative ecological studies," *The American Naturalist*, vol. 100, no. 914, pp. 419–424, 1966.
- [41] 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 homegeneity analysis," *Intellectual Economics*, vol. 5, no. 2, pp. 224–243, 2011.
- [42] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.
- [43] J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, R. O'hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner, "vegan: Community ecology package. r package version 1.17-2," *R Foundation for Statistical Computing, Vienna. Available: CRAN. R-project. org/package= vegan.(July 2012)*, 2011.
- [44] D. Štanikūnas, "Matų ir metodų poveikis lietuviškų tekstų stilometrinei analizei," Master's thesis, Vytautas Magnus University, 2017.
- [45] M. P. Fay and M. A. Proschan, "Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules," *Statistics surveys*, vol. 4, p. 1, 2010.
- [46] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Hierarchical clustering," *Cluster Analysis, 5th Edition*, pp. 71–110, 2011.
- [47] L. Rokach and O. Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 321–352.

[48] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.