Initial Analysis of Characteristics of Textual Genres: Comparison of Lithuanian and English

J. Mandravickaitė, T. Krilavičius and K. L. Man

Abstract—**We report an ongoing study on statistical characteristics of texts written in different genres. At this stage, we compared Lithuanian and English texts of different genres. We used 16 indices which describe frequency structure of text as well as measure vocabulary richness. Initial study showed significant differences of indices calculated for genre pairs of the same language. Analysis of indices showed that the correlation between the various indices is high.**

Index Terms—**text analysis, genre, vocabulary richness, frequency structure of text, English, Lithuanian.**

I. Introduction

WE report an ongoing study on statistical characteristics of texts written in different genres. It has been suggested that genres resonate with people because they provide familiarity and the shorthand of communication [20], [19], [2]. Also, genres tend to shift hand-in-hand with public opinion and reflect widespread culture of certain period(s) [9], [10]. From NLP perspective, genres come in use in text classification (e.g. [5]) and categorization (e.g. [16]), natural language generation (e.g. [18], [8]), etc.

At this stage, we present preliminary statistical analysis of Lithuanian and English texts of different genres (or super-genres, to be more accurate [17]; however, for the sake of simplicity, “genre” was used). As the main point of interest was frequency structure of text taking genre aspect into consideration, we used 16 indices proposed by [11], [12], [13] and implemented in QUITA - Quantitative Index Text Analyzer [7].

II. Materials & Methods

We used part of Corpus of the Contemporary Lithuanian Language [14] (~1,5 million words) and Freiburg-LOB Corpus of British English (F-LOB) (~1 million words) [4] 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.

For our experiment in characterizing genres, we applied the following 16 indices (see [7] for explanations omitted here for brevity): 1) Type-Token Ratio (TTR), 2) h-Point, 3) Entropy, 4) Average Tokens Length (ATL), 5) R1, 6) Repeat Rate (RR), 7) Relative Repeat Rate of McIntosh (RRmc), 8) Λ (Lambda), 9) Adjusted Modulus (AM), 10) Gini’s coefficient (G), 11) R4, 12) Hapax Legomena Percentage (HP), 13) Curve Length (L), 14) Writers View (WV), 15) Curve Length Indicator (R), 16) Token Length Frequency Spectrum (TLFS).

Significance of calculated indices in terms of genres was tested with asymptotic *u-test* [3]. Also, assuming that different indices measure different things, we performed a hierarchical cluster analysis of the individual indices. This showed which indicators calculate similar things and provided another check of individual measures in terms of their robustness when taking genres in consideration.

III. Results

Results of significance test (asymptotic u-test) of calculated indices in terms of genres are provided in Table 1. The suffix “_LT” indicates Lithuanian part of experimental material, while suffix “_EN” indicates English part of experimental data. Of the indices studied, most of them 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.

Results of hierarchical cluster analysis are provided in Figures 1 & 2. First we used Pearson’s χ^2 to calculate correlation between indices. Cluster analysis set up consisted of Euclidean distance and "average" method of linkage. We note that the correlation between the various indices is high. Thus it appears that they measure similar things. The exceptions are small cluster of 3 measures (RR, G and WV) for Lithuanian part, which are outliers. However, English part does not have such notable outliers,

J. Mandravckaitė is with Vilnius University, Lithuania and Baltic Institute of Advanced Technology, Lithuania (corresponding author, e-mail: justina@ bpti.lt).

T. Krilavičius is with Vytautas Magnus University, Lithuania and Baltic Institute of Advanced Technology, Lithuania (e-mail: t.krilavicius@bpti.lt).

K. L. Man is with Xi’an Jiaotong-Liverpool University, China and Swinburne University of Technology Sarawak, Malaysia
(e-mail: ka.man@xjtlu.edu.cn).

there are 2 main clusters that have 8 indices in them each.

TABLE I RESULTS OF SIGNIFICANCE TEST: GENRE PAIRS.

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

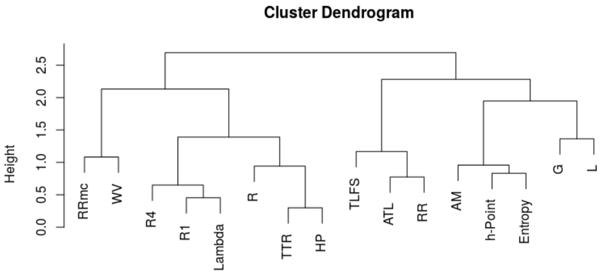

Fig. 1. Cluster analysis of correlations of indices studied: English.

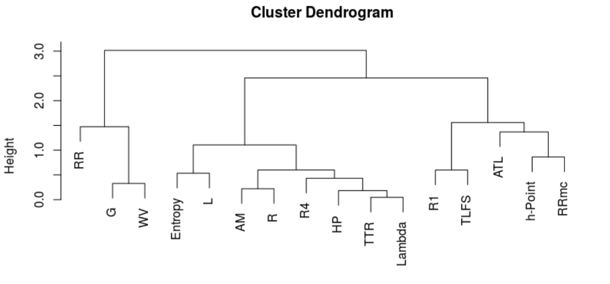

Fig. 2. Cluster analysis of correlations of indices studied: Lithuanian.

IV. Conclusion and Future Plans

We presented an ongoing work on characteristics of texts written in different genres for English and Lithuanian. Features used for this study seemed promising for characterization of genres as there were significant differences for genre pairs in terms of calculated indices, however, for more substantial conclusions additional research is necessary. Thus we plan to extend this work to larger text collections as well as additional genres. We also plan to examine other languages to see whether similar effects found in this study would persist.

References

[1] R. Čech, “Frequency structure of New Year’s presidential speeches in Czech. The authorship analysis”, *Issues in Quantitative Linguistics*, vol. 2, 2011, pp. 82-94.

[2] A. J. Devitt, “Generalizing about genre: new conceptions of an old concept”, *College Composition and Communication*, vol. 44, 1993, pp. 573–586.

[3] M. P. Fay & M. A. Proschan, “Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules”, *Statistics surveys*, vol. 4, 2010, pp. 1-39.

[4] M. Hundt, A. Sand, & R. Siemund, *Manual of information to accompany the Freiburg-LOB Corpus of British English ('FLOB')*, Albert-Ludwigs-Universität Freiburg, 1998.

[5] Y. Kim & S. Ross, “Variation of word frequencies across genre classification tasks”, In: C. Thanos, F. Borri, & A. Launaro (eds.), *Second DELOS Conference on Digital Libraries: Pisa, Italy, 5-7 December 2007*, Series: The DELOS network of excellence on digital libraries, GEIE-ERCIM: Sophia Antipolis, Nice, France, 2007.

[6] M. Kubát, & R. Čech, “Thematic Concentration and Vocabulary Richness”, In *Issues in Quantitative Linguistics*, vol. 4, 2016, pp. 150-159.

[7] M. Kubát, V. Matlach, & R. Čech, “QUITA - Quantitative Index text Analyzer”, *Studies in Quantitative Linguistics*, vol. 18, Lüdensheid: RAM, 2014.

[8] C. van der Lee, E. Krahmer & S. Wubben, “Pass: A dutch data-to-text system for soccer, targeted towards specific audiences”, In *Proc. INLG’17*, 2017, pp. 95–104.

[9] C. R. Miller, “Genre as social action”, *Quarterly Journal of Speech*, vol. 70, 1984, pp. 151–167.

[10] C. R. Miller, “Genre as a social action”, *Genre and the New Rhetoric*, A. Freedman and P. Medway (eds.), London: Taylor and Francis, 1994, pp. 20–35.

[11] I.-I. Popescu, G. Altmann, P. Grzybek, B.D. Jayaram, R. Köhler, V. Krupa, J. Mačutek, R. Pustet, L. Uhlířová & M. N. Vidya, *Word frequency studies*, Berlin-New York: Mouton de Gruyter, 2009.

[12] I.-I. Popescu, Čech, R., Altmann, G., *The lambda-structure of texts*, Lüdenscheid: Ram-Verlag, 2011.

[13] I.-I. Popescu, J. Mačutek & G. Altmann, “Word forms, style and typology”, *Glottotheory*, vol. 3(1), 2010, pp. 89-96.

[14] E. Rimkutė, J. Kovalevskaitė, V. Melninkaitė, A. Utka, & D. Vitkutė-Adžgauskienė, “Corpus of Contemporary Lithuanian Language – the Standardised Way”, *Proc. of the 4th International Conference HLT – The Baltic Perspective*, 2010, pp. 154–160.

[15] M. D. Rohangiz, “Authorship attribution and statistical text analysis”, *Metodološki zvezki*, vol. 4(2), 2007, pp. 149–163.

[16] E. Stamatatos, N. Fakotakis & G. Kokkinakis, “Automatic text categorization in terms of genre and author”, *Computational Linguistics*, vol. 26(4), 2001, pp. 471–495.

[17] G. Steen, “Genres of discourse and the definition of literature”, *Discourse Processes*, vol. 28, 1999, pp. 109-120.

[18] O. Stock & C. Strapparava, “The act of creating humorous acronyms”, *Applied Artificial Intelligence*, vol. 19 (2), 2005, pp. 137–151.

[19] J. M. Swales, *Genre Analysis: English in Academic and Research Settings*, Cambridge: Cambridge University Press, 1990.

[20] A. Tereszkiewicz, “Lead, headline, news abstract? – Genre conventions of news sections on newspaper websites”, *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, vol. 129, 2012, pp. 211–224.