Multiword Expressions for Capturing Stylistic Variation Between Genders in the Lithuanian Parliament

Justina Mandravickaitė[1,2], Michael Oakes[3]
[1]Faculty of Philology, Vilnius University, Vilnius, Lithuania
[2]Baltic Institute of Advanced Technology, Vilnius, Lithuania
[3]Research Institute of Information and Language Processing, University of Wolverhampton, Wolverhampton, United Kingdom

E-mail: justina@bpti.lt, Michael.Oakes@wlv.ac.uk

Abstract

The relation between gender and language has been studied by many authors, but there is no general agreement regarding gender influence on language usage in the professional environment. This could be because in most of the studies data sets are too small or texts of individual authors are too short in order to capture differences of language usage according to gender successfully. This study draws on a larger corpus of transcribed speeches in the Lithu anian Parliament (1990-2013) to explore gender differences in a language with a setting of political debates using stylometric analysis. The experimental set up consists of multiword expressions as features (formulaic language can allow a more detailed interpretation of the results in comparison to character n-grams or even most frequent words) combined with unsupervised machine learning algorithms to avoid the class imbalance problem. MWEs as features in combination with distance measures and hierarchical clustering were successful in capturing and mapping difference in speech according to gender in the Lithuanian Parliament. Our results agree with the experimental outcomes of Hoover (2002) and Hoover (2003), where frequent word sequences and collocations combined with clustering showed more accurate results than just frequent words.

Keywords: multiword expressions, stylometry, parliamentary speeches

1. Introduction

Gender influence on language usage has been studied by many authors, but common agreement has not yet been reached (Lakoff, 1973; Holmes, 2006; Holmes, 2013; Argamon et al., 2003). Understanding gender differences in a professional environment would assist in a more balanced atmosphere (Herring and Paolillo, 2006; Dynel, 2008). Most previous studies relied on relatively small data sets, texts written by the individual authors which were too short to capture the variation in the language usage according to gender (Newman et al., 2008; Herring and Martinson, 2004). Besides, some authors have claimed that gender differences in language depend on the context, e.g., people assume male language in a formal setting and female in an informal environment (Pennebaker, 2011).

In this paper the impact of gender on the language used in a professional setting, i.e., Lithuanian Parliament debates, is explored. We study language with respect to style, i.e., male and female style of the language usage in the Parliament by applying computational stylistics or stylometry. Stylometry is based on two hypotheses: (1) the human stylome hypothesis, i.e., each individual has a unique writing style (Van Halteren et al., 2005); (2) the unique writing style of an individual can be measured (Stamatatos, 2009). From an information retrieval perspective, stylometry allows the derivation of meta-knowledge, i.e., what can be learned from the text about the author (Daelemans, 2013). This can be gender (Luyckx et al., 2006; Argamon et al., 2003; Cheng et al., 2011; Koppel et al., 2002), but also such things as age (Dahllöf, 2012), psychological characteristics (Luyckx and Daelemans, 2008), and political affiliation (Dahllöf, 2012).

As in many other studies of gender and language (Yu, 2014; Herring and Martinson, 2004), biological sex as the criterion for gender was used in this study. Also, we compare differences in the gender related language use at the group level. The Lithuanian language allows an easy distinction between male and female legislators based on their names.

This study seeks not to attribute text samples to female or male MPs (the authorship attribution task), but to explore variation of language use based on gender in political debates of the Lithuanian Parliament (detecting stylistic variation). Since one reason that idiolects differ is that people have different reserves of prefabricated word sequences (Larner, 2014; Johnson and Wright, 2014), in our experiments multiword expressions were used as distinguishing features speeches of female and male MPs. Also, because of the high imbalance in terms of the amount of data (significantly more for male MPs than for female MPs) as well as no gold standard corpus for reference being available, we used unsupervised machine learning methods for detecting stylistic variation between speeches made by female MPs and male MPs. As most stylometric experiments using formulaic language as a feature were performed for English (e. g., Hoover (2003)), the main question this study seeks to answer is whether variability in language use with respect to style can be successfully captured using fixed word sequences, i.e., multiword expressions as features of Lithuanian which is a highly inflected language.

2. Data set

A corpus of parliamentary speeches from the Lithuanian Parliament[1] was used for capturing stylistic variation between genders. It consists of parliamentary speeches from March 1990 till December 2013. 10,727 speeches were made by female members of Parliament (MPs) and 100,181 by male MPs. The whole corpus contains 23,908,302 words (2,357,596 by female MPs and 21,550,706 by male MPs). Further statistics are shown in Table 1 (Kapočiūtė-Dzikienė and Utka, 2014).

	Number of samples	Number of words	Number of unique words	Average length of a sample in words
Female MPs	10727	2357596	93611	219.78
Male MPs	100181	21550706	268030	215.12
TOTAL	110908	23908302	279494	215.57

Table 1: Statistics of the corpus of transcribed Lithuanian parliamentary speeches.

The number of MPs included is 147, being only those included in the corpus, who produced at least 200 speeches of at least 100 words each. Out of 147 MPs, 129 were male and 18 were female.

All the samples were concatenated into two large documents based on gender. Then these two documents for the sake of faster processing were split into parts of equal size (except for the last parts of each big original document), giving 15 smaller documents of transcribed speeches from female MPs and 15 from male MPs.

3. Method

3.1 Stylistic features

Character n-grams are considered to be the most effective features in stylometric analysis (Kestemont, 2014; Stamatatos, 2009; Šarkutė and Utka, 2015; Kapociute-Dzikiene et al., 2014) because they are language-independent, are able to record style and stylistic differences and do not require external linguistic tools such as a part-of-speech tagger or parser. Using the most frequent words or function words (which in most cases have a high frequency (Hochmann et al., 2010; Sigurd et al., 2004)) as linguistic features is the most popular solution (Burrows, 1992; Hoover, 2007; Eder, 2013b; Rybicki and Eder, 2011; Eder and Rybicki, 2013; Eder, 2013a) for stylometric analysis. Most frequent words (MFW) are considered to be topic-neutral and have been relatively successful (Juola and Baayen, 2005; Holmes et al., 2001; Burrows, 2002).

However, we decided to use multiword expressions as linguistic features for our analysis. The choice was based on the assumption that the speech of politicians in their professional setting is rather formalised, and so uses specific expressions. Also, formulaic language can allow a more detailed interpretation (Antonia et al, 2014; Suzuki et al, 2012) of the results in comparison to character n-grams or even most frequent words. In a broad sense, a multiword expression (MWE) is a sequence of at least two words that are frequently used together (Marcinkevičienė, 2001). MWEs have “idiosyncratic interpretations that cross word boundaries (or spaces)” (Sag et al. 2002).

To obtain a list of MWEs to use in our stylometric experiments we used the Ngram Statistics Package[2] (Pedersen et al, 2011). The corpus of parliamentary speeches in the Lithuanian Parliament was split into word bi-grams and then association measures were calculated for each one. Lexical association measures assess the degree of association between components of possible MWE. For our experiment we chose two widely known association measures – Log-likelihood and Dice. Log-likelihood brings out word sequences with the highest degree of valence which ensures strength of association among the MWE components, while Dice gives higher values for word sequences in the corpus with equal frequencies and ignores sequences that are rare (Hunston, 2002). From the MWE candidates for which we calculated Log-likelihood and Dice values we took only the ones with the highest values and then manually eliminated sequences that were definitely not MWE. Eventually for our stylometric analysis we used a list of 4737 bi-gram MWEs. Examples of some MWE found in the corpus are presented in Table 2.

Dice	Log-likelihood
profesinėms sąjungoms (trade unions, dative)	gerbiamieji kolegos (dear colleagues, vocative)

[1] "Automatic Authorship Attribution and Author Profiling for the Lithuanian Language” (ASTRA) (No. LIT-8-69), 2014 – 2015.

[2] http://www.d.umn.edu/~tpederse/nsp.html

šventų atsiminimų (saint memories, genitive)	taip pat (also/as well)
Didžiojoje Britanijoje (Great Britain, locative)	Lietuvos Respublikos (Republic of Lithuania, genitive)
Kristijono onelaičio (Kristijonas Donelaitis, genitive)	bendru sutarimu (by consensus)
chasidų sinagogos(hassidic synagogue, genitive)	įstatymo projektas (bill (law), nominative)
status quo	Seimo nariai (members of the Parliament, nominative)
Drąsiaus Kedžio (Drąsius Kedys, genitive)	iš tikrųjų (indeed/actually)

Table 2: Examples of MWE by lexical association measures (Dice and Log-likelihood).

3.2 Statistical measures and experimental setup

The experiments were performed using the Stylo package for stylometric analysis with R (Eder et al., 2014). For the chosen approach firstly, using the whole corpus, a raw frequency list of features is generated, then normalized using z-scores. The z-scores are calculated by subtracting the mean frequency of a certain feature in one text from its mean frequency in all the texts in the corpus and dividing this difference by the standard deviation (Hoover, 2004a). Using Burrows' Delta measure (Burrows, 2002), the dissimilarity between two texts is the mean of the differences in z-scores over all the features under consideration in those two texts. A distance matrix is generated consisting of all the pairwise dissimilarity scores between the texts. This distance matrix can be visualized using a visualization technique such as a dendrogram produced by hierarchical agglomerative clustering.

Burrows's Delta is possibly the most popular distance measure used for stylometric analysis (Burrows, 2002; Rybicki and Eder, 2011). Delta depends on z-scores, the number of texts and the balance among them in terms of amount, length and number of authors (Stamatatos, 2009). Although this distance measure is effective for English and German texts, it has been less successful for more inflected languages such as Latin and Polish (Rybicki and Eder, 2011). Therefore a variant of Delta was chosen for our experiments. Eder's Delta is a modified standard Burrows's Delta, yet it gives more weight to the frequent features and rescales less frequent features to avoid random infrequent ones (Eder et al., 2014). It was developed for use with highly inflected languages, such as Lithuanian. However, this Delta variant retains sensitivity to the number of samples in the same way as other Delta variations.

The purpose of this paper is to capture stylistic dissimilarities/variations by mapping positions of the text samples in relation to each other according to gender, and therefore (hierarchical) clustering was chosen. Though its sensitivity to changes in the number of features or methods of grouping is well known (Eder, 2013a; Luyckx et al., 2006), in this study it gave rather stable results.

Additionally, the robustness of hierarchical clustering in this study was examined using the bootstrap procedure (Eder, 2013a). This procedure used extensions of Burrows's Delta (Argamon, 2008; Eder et al., 2014) with bootstrap consensus trees (Eder, 2013a) as a way to improve the reliability of cluster analysis dendrograms. Hierarchical clustering analysis lacks standard validation procedures, except for visual examination, and hence we found a combination of hierarchical clustering dendrograms and decision trees a useful tool for the evaluation of results.

4. Results

For our exploration of stylistic variation between female and male MPs from 50 to 4730 most frequent features, in this case MWEs, were chosen. Eder's Delta was combined with hierarchical clustering to visualize the categorization as well as for mapping positions of the samples in relation to each other, i.e. capturing variation in speech according to gender. No culling was applied (Eder et al., 2014; Hoover, 2004b) in our experiments. During the culling procedure words which have most of their occurrences in a single text instead of being distributed throughout the corpus, are eliminated (Stamatatos, 2009).

The results showed that using MWEs as features for stylometric analysis in combination with Delta variants and hierarchical clustering was successful in capturing differences in speech in the Lithuanian Parliament according to gender. The 50 most frequent MWEs were enough to capture the variation between the speeches of female and male MPs.

This recorded variation remained stable up to 1200 most frequent MWEs. This means that the first 1200 MWEs in the list used for analysis were helpful in capturing variation according to gender. The results are shown in Figures 1 and 2, where the data set is clearly divided into clusters corresponding to male and female speakers.

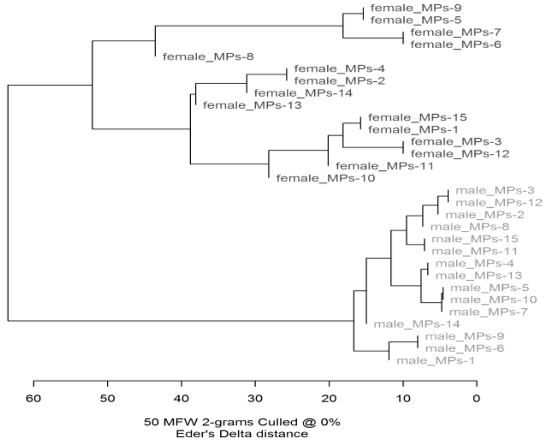

Figure 1: Variation between the speeches of female and male MPs with 50 most frequent MWEs as features.

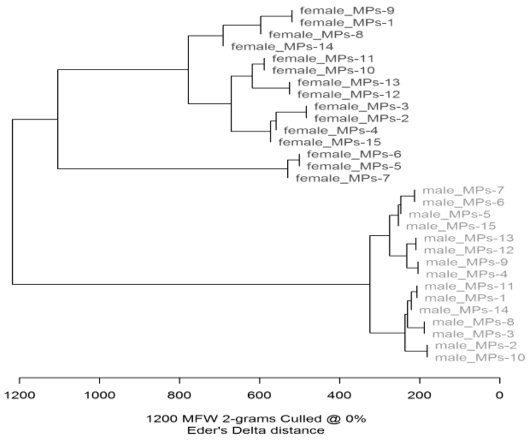

Figure 2: Variation between the speeches of female and male MPs with 1200 most frequent MWEs as features.

The Bootstrap Consensus Tree (BCT) procedure was applied to test the results. It is a combination of hierarchical clustering and decision trees (Eder, 2013a). It works by repeating the clustering with hundreds of different subsets of the original data, and retaining only those linkages between texts which appear in an above threshold proportion of runs. A consensus strength of 0.5 was chosen, i.e., the linkages between two texts retained if they appeared in at least half of the bootstrapping runs. The BCT results for discriminating between male and female legislators in the Lithuanian Parliament are shown in Figure 3.

Among other observations, female MPswere more inclined to use morphological collocations.These are defined as fixed expressions consisting of two or more functional words (inflected or non-inflected) that have a unified common meaning, are non-compositional and also have a syntactic function (Rimkutė, 2009; Rimkutė and Kovalevskaitė, 2010). Of the most frequent MWEs, female MPs used such morphological collocations as *dėl to* (‘therefore‘), *iki šiol* (‘by now‘), *be abejo* (‘undoubtedly‘), etc. more frequently then male MPs. Also, in the transcribed speeches of female MPs there occured more subjunctive constructions (indicating suggestion, certain degree of uncerntainty), for example, *aš siūlyčiau* (‘I would suggest‘), *aš manyčiau* (‘I would think‘), *galėtų būti* (‘[it] could be‘).

Male MPs, among other differences in comparison to female MPs, tended to use more references to other MPs. Moreover, they used more sequences related to power (e.g., *gynybos štabas* (‘defence headquarters‘), *ginkluotosios pajėgos* (‘armed forces‘), *diplomatinis korpusas*, (‘diplomatic corps‘)) economics/finance (e.g., *finansinė atskaitomybė* (‘financial accountability‘), *finansinis tvarumas* (‘financial sustainability‘), *fiskalinis deficitas* (‘fiscal deficit‘)). Also, male MPs used more verbs in the first person plural, for example *ar pritariame* (‘do we agree [?]‘), *galime sutarti* (‘we can agree‘), *būkime biedni* (‘let‘s be poor [but proper]‘ – part of popular Lithuanian saying).

As presented above, using stylometric analysis with MWEs as features, we were able to record certain differences in language usage according to gender. Some of them were topical, others of the nature of lexical or morphosyntactic style. For making more generalisations, further research is needed.

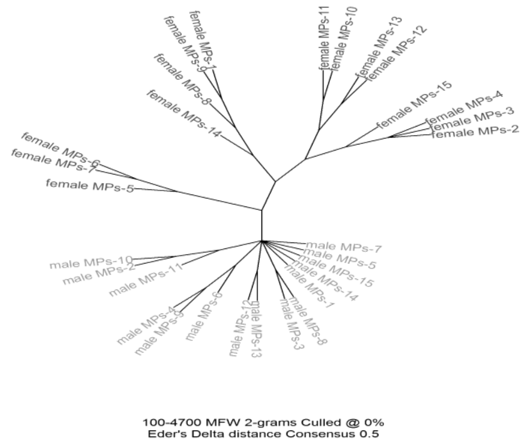

Figure 3: Variation between the speeches of female and male MPs BCT with consensus of 0.5.

5. Conclusions and future work

MWEs as features in combination with distance measures and hierarchical clustering were successful in capturing and mapping difference in speech according to gender in the Lithuanian Parliament. Our results agree with the experimental outcomes of Hoover (2002) and Hoover (2003), where frequent word sequences and collocations combined with clustering showed more accurate results than just frequent words. However, although Eder (2011) reported increased accuracy using bi- and tri-gram collocations for English, word sequences were useless for other languages, especially Latin. Also, we got useful results with far fewer features than some studies (e.g. Eder (2010), Stamatatos (2006)), suggest for successful analysis. Therefore further, more extensive, experiments are required regarding the usefulness of MWEs as features, as well as the number of features and their range. For example, how many features from the beginning of the feature list are useful, and when we should select features from the middle and when from the end of the list of MWEs ordered by frequency. The effect of culling, the elimination of features with the most occurrences in a single text instead of being distributed throughout the corpus, also needs to be explored. We have shown that MWE can be used as linguistic features to discriminate between male and female speeches in the Lithuanian parliament, Lithuanian being an inflected language, and this approach could contribute to research on different usage of language depending on gender.

6. Bibliographical References

Alexis A., Craig, H., Elliott, J. (2014). Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution. *Literary and Linguistic Computing*, 29(2), pp. 147--163.

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2), pp. 131--147.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text-the Hague then Amsterdam then Berlin,* 23(3), pp. 321--346.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), pp. 267--287.

Burrows, J. F. (1992). Not unles you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7(2), pp. 91--109.

Cheng, N., Chandramouli, R., and Subbalakshmi, K. P. (2011). Author gender identification from text. *Digital Investigation*, 8(1), pp. 78--88.

Daelemans, W. (2013). Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing*, Springer, pp 451--462.

Dahllöf, M. (2012). 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*, 27(2), pp. 139--153.

Dynel, M. (2008). Gendered Discourse in the Professional Workplace. *Journal of Pragmatics*, 9(40), pp. 1620--1625.

Eder, M. & Rybicki, J. (2013). Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*, 28(2), pp. 229--236.

Eder, M. (2010). Does size matter? Authorship attribution, small samples, big problem. *Proceedings of Digital Humanities*, pp. 132--135.

Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6(1), pp. 99--114.

Eder, M. (2013a). Computational stylistics and biblical translation: How reliable can a dendrogram be. *The translator and the computer*, pp. 155--170.

Eder, M. (2013b). Mind your corpus: systematic errors in authorship attribution. *Literary and linguistic computing*, 28(4), pp. 603--614.

Eder, M., Rybicki, J., Kestemont, M., and maintainer Eder, M. (2014). Package 'stylo'.

Herring, S. C. & Martinson, A. (2004). Assessing gender authenticity in computer mediated language use evidence from an identity game. *Journal of Language and Social Psychology*, 23(4), pp. 424--446.

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10 (4), pp. 439--459.

Hochmann, J. R., Endress, A. D, and Mehler, J. (2010). Word frequency as a cue for identifying function words in infancy. *Cognition*, 115(3), pp. 444--457.

Holmes, D. I., Gordon, L. J., and Wilson, C. (2001). A widow and her soldier: Stylometry and the American civil war. *Literary and Linguistic Computing*, 16(4), pp. 403--420.

Holmes, J. (2006). Sharing a laugh: Pragmatic aspects of humor and gender in the workplace. *Journal of Pragmatics*, 38(1), pp. 26--50.

Holmes, J. (2013). *Women, men and politeness*. Routledge.

Hoover, D. L. (2002). Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing,* 17(2), pp. 157--180.

Hoover, D. L. (2003). Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3), pp. 261--286.

Hoover, D. L. (2004a). Delta prime? *Literary and Linguistic Computing*, 19(4), pp. 477--495.

Hoover, D. L. (2004b). Testing Burrows's Delta. *Literary and linguistic computing*, 19(4), pp. 453--475.

Hoover, D. L. (2007). Corpus stylistics, stylometry, and the styles of Henry James. *Style*, 41(2), pp. 174--203.

Hunston, S. (2002). Methods in corpus linguistics: Beyond the concordance line. *Corpora in Applied Linguistics*, pp. 36--95.

Johnson, A., Wright, D. (2014). Identifying idiolect in forensic authorship attribution: an n-gram textbite approach. *Language and Law/Linguagem e Direito*, 1(1), pp. 37--69.

Juola, P. & Baayen, R H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, 20(Suppl), pp. 59--67.

Kapočiūtė-Dzikienė J., Šarkutė, L. and Utka, A. (2015). The effect of author set size in authorship attribution for Lithuanian. *Nordic Conference of Computational Linguistics NODALIDA*, pp. 87--96.

Kapočiūtė-Dzikienė, J. & Utka, A. (2014). Seimo posėdžių stenogramų tekstynas autorystės nustatymo bei autoriaus profilio sudarymo tyrimams. Linguistics/Kalbotyra, 66, pp. 27--45.

Kapociute-Dzikiene, J., Sarkute, L., and Utka, A. (2014). Automatic author profiling of Lithuanian parliamentary speeches: exploring the influence of features and dataset sizes. *Proceedings of the Sixth International Conference Baltic HLT 2014*, pp. 99--106.

Kestemont, M. (2014). Function words in authorship attribution from black magic to theory? *EACL 2014*, pp. 59--66.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), pp. 401--412.

Lakoff, R. (1973). Language and woman's place. *Language in society*, 2(01), pp. 45--79.

Larner, S. (2014). A preliminary investigation into the use of fixed formulaic sequences as a marker of authorship. *International Journal of Speech Language and the Law*, 21(1), pp. 1--22.

Luyckx K. & Daelemans, W. (2008). Personae: a corpus for author and personality prediction from text. *LREC 2008*.

Luyckx, K., Daelemans, W., and Vanhoutte, E. (2006). Stylogenetics: Clustering-based stylistic analysis of literary corpora. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.

Marcinkevičienė, R. (2001). Tradicinė frazeologija ir kiti stabilūs žodžių junginiai. *Lituanistica*, 4(48), pp. 81--98.

Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3), pp. 211--236.

Pedersen, T., Banerjee, S. and McInnes, B. T. (2011). The Ngram statistics package (text::nsp): A flexible tool for identifying ngrams, collocations, and word associations. *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (ACL)*, pp. 131--133.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, 211(2828), pp. 42--45.

Rimkutė, E. (2009). Gramatinė morfologinių samplaikų klasifikacija. *Kalbų studijos* 14, pp. 32-38.

Rimkutė, E. & Kovalevskaitė, J. (2010). Sudėtinės ir suaugtinės lietuvių kalbos morfologinės samplaikos. *Kalbų studijos* 16, pp. 79--88.

Rybicki, J. & Eder, M. (2011). Deeper delta across genres and languages: do we really need the most frequent words? *Literary and linguistic computing*, 26(3), pp. 315--321.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, Dan. (2002). Multiword expressions: A pain in the neck for NLP. *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, pp. 1--15.

Sigurd, B., Eeg-Olofsson, M., and Van Weijer, J. (2004). Word length, sentence length and frequency–zipf revisited. *Studia Linguistica*, 58(1), pp. 37--52.

Stamatatos, E. (2006). Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, 15(05), pp. 823--838.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3), pp. 538--556.

Van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M., and Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1), pp. 65--77.

Yu, B. (2014). Language and gender in congressional speech. *Literary and Linguistic Computing*, 29(1), pp. 118--132.