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

Justina Mandravickaite\textsuperscript{1,2}, Michael Oakes\textsuperscript{3}

justina@bpti.lt, Michael.Oakes@wlv.ac.uk

\textsuperscript{1} Vilnius University, Lithuania

\textsuperscript{2} Baltic Institute of Advanced Technology, Lithuania

\textsuperscript{3} University of Wolverhampton, United Kingdom

Introduction

- Understanding gender differences in a professional environment \rightarrow more balanced atmosphere
- Stylometry is based on two hypotheses:
  1. the human stylome hypothesis, i.e., each individual has a unique writing style;
  2. the unique writing style of an individual can be measured.
- Information retrieval perspective: stylometry allows the derivation of metaknowledge \rightarrow what can be learned from the text about the author
- Task: detecting stylistic variation based on gender in political debates.
- Assumption: people have different reserves of prefabricated word sequences \rightarrow multiword expressions used as distinguishing features.

Data and Method

Data:
- Corpus contains 23,908,302 words (2,357,596 by female MPs and 21,550,706 by male MPs).

Method:
- Stylistic features
  - multiword expressions - a sequence of $\geq 2$ words that are frequently used together (broad sense).
- Statistical measure
  - Eder’s Delta \[ \Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{f_i(A) - f_i(B)}{\sigma_i} \right) \times \frac{n - n_i + 1}{n} \]
- Visualization
  - Hierarchical agglomerative clustering & Bootstrap consensus tree

Results

Female MPs:
- more morphological collocations, such as dėl to (“therefore”), iki šiol (“by now”), be abejo (“undoubtedly”);
- more subjunctive constructions, e.g. aš siūlyčiau (“I would suggest”), aš manyčiau (“I would think”), galėtų būti (“\[it\] could be”).

Male MPs:
- more references to other MPs;
- more sequences related to power/military, e.g., gynybos štabas (“defence headquarters”), economics/finance, e.g., finansinė atskaitomybė (“financial accountability”);
- more verbs in the first person plural, e.g., ar pritariame (“do we agree [?]”), galime sutarti (“we can agree”).

Conclusions and Future Work

- Multiword expressions can be used as linguistic features to discriminate between male and female speeches in the Lithuanian parliament, Lithuanian being an inflected language.
- This approach could contribute to research on different usage of language depending on gender.
- Future work includes more extensive experiments regarding the usefulness of MWEs as features, as well as the number of features and their range.