Abstract

Name student: Isolde van Dorst

Thesis title: You, thou and thee: A statistical analysis of Shakespeare's use of pronominal address terms

In recent decades a lot of research on Shakespeare's use of the singular second person pronouns *you*, *thou* and *thee* has been done. However, the results so far are inconclusive as to which features influence the choice of pronoun. This study aims to create a prediction model to find which linguistic and extra-linguistic features influence the pronoun choice made by Shakespeare. The 23 features used in this study contain speaker and addressee information (e.g. age and status), play and scene data (e.g. play name and genre), and contextual information (e.g. the words used in close proximity of the pronoun).

The three algorithms used in this study, Naive Bayes, decision tree and support vector machine, are selected based on their difference in assumptions and learning biases. Additionally, a binary and trinary prediction was performed. For the trinary classification, the three pronouns *thou*, *thee* and *you* were kept separated. In the binary classification, *thou* and *thee* were condensed into one category THOU. The latter is common in YOU/THOU research, while the difference in case of the THOU pronouns supports a trinary approach. As predicted, the support vector machine models score best on the four scores assessed in this study: precision, recall, F-measure and accuracy. With 87.3% accuracy, the binary support vector machine model scored 24% better than the baseline.

Additionally, there is one group of features that shows up as the main predictor of the pronoun, namely the words of the n-gram. In particular RW1 and LW1 are important, which show that the direct linguistic context of the pronoun is most important when predicting the pronoun. There are several other features that show a positive influence on the pronoun prediction as well, among which are the names of the speaker and addressee, the status differential, and positive and negative sentiment.