1.01 UDC: 004.934:821.111SHAK(083.41)

Isolde van Dorst*

You, Thou and Thee: A Statistical Analysis of Shakespeare's Use of Pronominal Address Terms

IZVLEČEK

YOU, THOU IN THEE: STATISTIČNA ANALIZA UPORABE IZRAZOV ZAIMKOVNEGA NASLAVLJANJA PRI SHAKESPEARU

Študija se ukvarja z oblikovanjem napovednega modela, namenjenega ugotavljanju, katere jezikovne in nejezikovne značilnosti vplivajo na izbiro zaimkov v Shakespearovih igrah. V angleščini, ki se je uporabljala v Shakespearovem obdobju, je razlikovanje med YOU in THOU, ki je danes arhaično, še obstajalo. Običajno se navaja, da sta ga določala relativni družbeni status ter osebna bližina govorca in naslovljenca. Vendar pa je treba še ugotoviti, ali bo statistično strojno učenje potrdilo to tradicionalno razlago. Proučuje se 23 značilnosti, izbranih z različnih jezikoslovnih področij, kot so pragmatika, sociolingvistika in analiza pogovora. Trije uporabljeni algoritmi – naivni Bayesov klasifikator, odločitveno drevo in metoda podpornih vektorjev – so izbrani kot ilustrativni nabor možnih modelov zaradi njihovih kontrastnih predpostavk in učne pristranskosti. Opravita se dve napovedi, prva o binarnem (you/thou) razlikovanju in druga o trinarnem (you/thou/thee) razlikovanju. Od vseh treh algoritmov daje najboljše rezultate metoda podpornih vektorjev. Po ugotovitvah so značilnosti, ki najbolje napovejo izbiro zaimka, besede iz neposrednega jezikovnega konteksta. Izkazalo se je, da na napoved zaimka vpliva tudi več drugih značilnosti, vključno z imenom govorca in naslovljenca, razliko v statusu ter pozitivnim ali negativnim mnenjem.

Ključne besede: izrazi zaimkovnega naslavljanja, Shakespeare, korpusno jezikoslovje, digitalna humanistika, statistično modeliranje

* **ESRC Centre for Corpus Approaches to Social Science, Lancaster University, Faculty of ICT, University of Malta, Faculty of Arts, University of Groningen, isoldevd@hotmail.com**

ABSTRACT

This study creates a prediction model to identify which linguistic and extra-linguistic features influence pronoun choices in the plays of Shakespeare. In the English of Shakespeare's time, the now-archaic distinction between you and thou persisted, and is usually reported as being determined by relative social status and personal closeness of speaker and addressee. However, it remains to be determined whether statistical machine learning will support this traditional explanation. 23 features are investigated, having been selected from multiple linguistic areas, such as pragmatics, sociolinguistics and conversation analysis. The three algorithms used, Naive Bayes, decision tree and support vector machine, are selected as illustrative of a range of possible models in light of their contrasting assumptions and learning biases. Two predictions are performed, firstly on a binary (you/thou) distinction and then on a trinary (you/thou/thee) distinction. Of the three algorithms, the support vector machine models score best. The features identified as the best predictors of pronoun choice are the words in the direct linguistic context. Several other features are also shown to influence the pronoun prediction, including the names of the speaker and addressee, the status differential, and positive and negative sentiment.

Keywords: pronominal address terms, Shakespeare, corpus linguistics, digital humanities, statistical modelling

Introduction

For several decades much research has been undertaken on the use of *you*, *thou* and *thee* in Shakespeare's works. However, the results so far have yet to arrive at an exact and conclusive answer regarding how these pronouns were used.

This study combines the strengths of multiple research fields in an effort to determine via hitherto unused methods which linguistic and extra-linguistic features influence the choice of second person singular pronoun (*you* versus *thou* or *thee*) in the plays of William Shakespeare. Prior findings in literary and linguistic studies are utilised to find which features could be relevant in this choice, and tools and applications created for corpus linguistics and computer science are exploited to analyse the data in a more exact way than has so far been accomplished. Through these techniques, I hope to identify which features can contribute to a more accurate prediction of pronoun choice, in a model to mimic the pronoun use of Shakespeare.

It is worth observing at this point that it has not yet been determined whether it is even possible to predict the pronoun based on linguistic features. Part of the aim of this paper is to make a determination on this point. In other words, is it possible to create a computational model that can predict which pronoun will be used based on a set of linguistic and extra-linguistic features taken from the text itself and selected on the basis of knowledge that we have of English in the late 1500s and early 1600s? To

accomplish this, all occurrences of *you*, *thou* and *thee* are extracted from Shakespeare's plays, and every instance is manually coded for 23 linguistic and extra-linguistic features, creating data which will serve to ascertain the answer to this primary question. A second question to be addressed is whether some features perform better as predictors of the pronoun choice than others. Thirdly, the issue of whether the use of different algorithms affects the prediction outcomes will be considered.

Throughout this paper, italicised *you*, *thou* and *thee* refer to specific pronoun forms. However, whereas *you* – in Early Modern English as in contemporary English – does not exhibit any formal variation for pronoun case, *thou* is strictly a nominative form with *thee* as its accusative/dative form. *Thou* and *thee* are therefore related inflectional forms of a single pronoun lemma; *you* exists in variation with both. Small capitals are used to indicate the pronoun lemmas, thus: you and thou, where thou includes both *thou* and *thee*. Whenever discussing pronouns in this paper, I am strictly referring to the singular second-person pronouns *you*, *thou* and *thee* that are examined in this study.

Background

Digital Humanities

Over the past few years, computational research has branched out into other research fields that are not necessarily closely connected to computer science. Digital Humanities (DH) is an umbrella term for all research that is computational but approaches the datasets investigated within, and/or addresses questions or problems that are of importance to, the disciplines of the humanities.

The popularity of Digital Humanities, a cross-domain field of study, is attributable to the fact that it does not diminish the differences between fields but rather operationalises this difference to solve difficulties that could not be dealt with within a single discipline. The role of computational methods in the humanities can be considered as that of a supporting character; in any DH computer modelling research, it should be kept in mind that the interpretation is as important at the suitability of a computational model and its outcomes.

Early Modern English and YOU/THOU

In Early Modern English (EModE), two different second person singular pronouns were used, namely the formally singular thou and the formally plural (but pragmatically also respectful-singular) you, with only the latter surviving the EModE period (Taavitsainen and Jucker 2003). The difference between the uses of these two pronouns is evident from multiple literary studies that have addressed Shakespeare's

work, work of his contemporaries, and other documents from this era, such as Walker (2003) and Busse (2002). These studies suggest that unwritten social rules governed the use of these pronouns, abiding by which rules was necessary in order to speak according to society's standards. The use of the two different pronouns acted as a sign of relative status: you would be used to superiors and thou towards inferiors. The choice of pronoun can thus also operate as a subtle means of showing respect or disrespect; using the pronouns in this way would have been natural and easy to English native speakers of the period.

Shakespeare lived during the Early Modern English period, and thus used both you and thou in his writing. His work was written less than 100 years before *thou* and *thee* disappeared from the standard language (surviving in dialects and archaicised registers, such as pious addresses to the divinity). Thus we may straightforwardly posit that the disappearance of thou was likely already in progress around his time. Though obviously heightened in its use of emotional and dramatic language and style to accommodate to the genre of the play script, the language of Shakespeare – including the usage of the two second-person pronouns – can be assumed to be a reasonably good representation of the language used generally in social interaction and conversation at that time (Calvo 1992).

Prior studies on YOU/THOU

Most studies of Shakespeare's use of YOU and THOU so far have been literary and nonnumeric studies (Brown and Gilman 1960; Quirk 1974; Calvo 1992); the relative few to have used data-based or quantitative techniques did not implement any method beyond directly comparing raw frequency counts (Busse 2003; Mazzon 2003; Stein 2003). Moreover, these studies did not look at all the extant Shakespeare plays, but instead chose a few plays to focus on. Nonetheless, these studies have demonstrated some patterns in the use of YOU and THOU and thus provide a workable foundation for a more in-depth study of the usage of those two pronouns.

These prior studies support in the overall conclusion that the pronouns YOU and THOU appear to be used to support the explicit expression of respect, social status, and familiarity. Quirk (1974) and Mazzon (2003) characterise the role of the pronoun as a linguistic marker, whose usage can be seen as either marked or unmarked. In other words, the use of a particular pronoun can be seen as marked when it is used unexpectedly, for example when YOU is expected based on social status, but THOU is used instead. Thus, in contrast to earlier studies (Brown and Gilman 1960), they do not perceive YOU and THOU to be in direct contrast, and to have a more variable interpretation than was assumed until then, based on the context it occurs in. Calvo (1992) and Stein (2003) expand on this by concluding that markedness of the pronoun is dependent on the context and the situation, in addition to the pronoun choice depending on stable factors such as the social statuses of, and the level of familiarity between, the characters

in Shakespeare's plays; the speakers and addressees in this study – rather than *just* the latter factors (Brown and Gilman 1960). The emotive effect of the utterances within which the YOU/THOU distinction is utilised is of importance as well; feelings such as anger and love for another character may find expression through pronoun choice. This is connected to the notion of respect, as, in an angry remark, marked pronouns can be used to disrespect the addressee based on their social status (Stein 2003).

As Stein (2003) and Busse (2006) already stressed in their studies, a study of YOU and THOU in Shakespeare cannot and should not be limited to a single research discipline. Rather, what is needed is a combination of literature, sociolinguistics, pragmatics and conversation analysis, which are all useful in capturing the complexity of pronominal address and the social constrictions that may have underpinned the choice of one honorific pronoun-form over the other.

Methodology

As has already been mentioned, this is a strictly empirical study which attempts to verify the findings of earlier research through a computational approach. The use of a computational, statistical method is motivated by the goal of creating a more objective representation of Shakespeare's use of YOU and THOU in his plays than has been accomplished so far, since it does not require analysis of meaning-in-context by a human being, but rather proceeds directly from quantitative measurements.

Hypotheses

Three hypotheses were formulated on the basis of the literature:

1. No single model will be able to predict the pronominal address term solely based on linguistic and extra-linguistic features.

This, being a null-hypothesis, is exactly what this study aims to falsify by developing such a model. It is not likely that a single model will be able to predict Shakespeare's original choice of YOU or THOU based on linguistic and extra-linguistic features, because this choice is dependent on so many factors. However, the application of literature, sociolinguistics, pragmatics and conversation analysis all combined into a computational model will be able to successfully predict the pronoun choice as it includes all the factors that might influence the choice for either YOU or THOU.

2. The features of social status, age and sentiment will be better predictors of the pronoun choice than other features.

A hierarchy will be established according to which the linguistic and extra-linguistic features are predicting the pronoun choice in the best performing model. It may be inferred from the literature that social status, age and sentiment are highly likely to be at the top of this hierarchy, among the most influential features; these three features have shown up most reliably in prior research.

3. The best performing algorithm will combine features both dependently and independently.

The different learning biases and assumptions of the three algorithms applied in this study will reveal how the features interact with one another. The first algorithm, Naive Bayes, assumes all features are independent of one another, while the decision tree algorithm assumes that the features are all dependent on each other. Lastly, the support vector machine works with both dependent and independent features. I expect the set of features that will be included in the final model to be a combination of both dependent and independent features, and therefore the support vector machine algorithm to perform best. The three algorithms will be discussed in more detail later in the chapter Classification based on three algorithms.

Data

The data for this study comes from the *Encyclopaedia of Shakespeare's Language* project[1], which is a research project at Lancaster University (UK). The project corpus consists of 38 of Shakespeare's plays, which includes all 36 plays from the First Folio with the addition of *The Two Noble Kinsmen* and *Pericles: Prince of Tyre*. A broadly annotated version of the full Shakespeare corpus can be found online[2]. Some of the annotation and all of the abbreviations used for the titles of the plays follow *The Arden Shakespeare*.

Linguistic and extra-linguistic features

The Encyclopaedia of Shakespeare's Language corpus is richly annotated. However, some additional annotation was necessary to perform a full analysis of what extra-linguistic features could be predictors of the pronominal address term. The full set of features used in this study can be found in Table 1. The added features are briefly described here.

As a referent (such as a second person singular pronoun) is dependent on context, the adjacent part of the utterance is used as a feature to test the effect of co-text. Six

1 More information on this project, which is funded by the Arts and Humanities Research Council (AH/N002415/1), can be found on http://wp.lancs.ac.uk/shakespearelang/.

2 CQPweb Main Page, http://cqpweb.lancs.ac.uk.

Table 1: List of all features used in this study

Feature	Acronym	Annotation
Genre	Genre	Pre-annotated
Play name	Play	Pre-annotated
Play, act, scene	Scene	Pre-annotated
Speaker ID	S_ID	Pre-annotated
Speaker gender	S_Gender	Pre-annotated
Speaker status	S_Status	Pre-annotated
Production date	Prod_Date	Pre-annotated
N-gram	LW1-3, RW1-3	Automatic
Positive sentiment	Pos_Sent	Automatic
Negative sentiment	Neg_Sent	Automatic
Speaker age	S_Age	Manual
Location	Location	Manual
Addressee ID	A_ID	Automatic
Addressee gender	A_Gender	Pre-annotated
Addressee status	A_Status	Pre-annotated
Addressee age	A_Age	Manual
Status differential	Stat_Diff	Automatic
No. of people addressed	A_Number	Pre-annotated

co-textual words are included, i.e. a 7-gram altogether. "LW" labels the words occurring on the left of the pronoun, and "RW" the words on the right of the pronoun. Each of these words are numbered based on their distance from the pronoun, e.g. LW3 is the third word on the left of the pronoun. In corpus linguistics, collocations are often examined within a three-word-window, meaning there are three words on either side of the word of interest. While I am not necessarily looking at specific collocations of YOU and THOU, the LW/RW features will look at similarities and differences in co-textual words to see if they can predict the pronoun choice.

Another feature noted as critical in prior studies is sentiment, that is the use of the pronoun to convey positivity or negativity. Sentiment was annotated with the use of the 7-gram described above. *SentiStrength* is a lexicon-based sentiment analysis program that scores phrases with a score for positivity and negativity (Thelwall et al. 2010). Since *SentiStrength* was developed to work with online comments rather than complete sentences as in formal written English, it works well with n-grams too. The scores for positivity and negativity are kept as separate variables.

The corpus already included metadata on the speakers; however, I wanted to include age as well. The age of a character is often not given except for when it is an important attribute of that character, making this difficult to annotate. Therefore, Quennell and Johnson's (2002) character descriptions were used. The characters were

sorted into a trinary classification, with 'adult' as the default category. Any deviations towards 'younger' or 'older' were based on textual references or the character's name, such as for 'Old Man' in *King Lear*. Older characters were occasionally classified as such based on the fact they had adult children with prominent roles in the plays.

A more global feature is the location where the scene is set. This was difficult to annotate, due to the often unreliable stage directions. Instead of a nominal description for each scene location, I used a binary annotation of 'public' and 'private'. The text itself was examined to determine the location based on what characters said about their location, but in addition Bate and Rasmussen's (2007) annotation and Greenblatt, Cohen, Howard and Maus' (1997) annotations were consulted. The use of these three resources enabled the binary manual annotation of location for every scene.

Besides the information about the speaker and the scene, information regarding the addressee is essential when analysing character interaction from a conversation analysis perspective. As a manual annotation for addressee would be incredibly time consuming, I instead used an automatic method which identifies the previous speaker as the addressee of any given utterance. This is in line with the last-as-next bias used in conversation analysis (Mazeland 2003). This means that, even in larger group conversations, it is often expected that the last speaker before the current speaker will also be the next speaker, thus making it likely that the current speaker is addressing the last speaker. If the utterances were interrupted by the start of a new scene or other stage directions (e.g. someone walking into the scene), the annotated addressee would be the *next* speaker rather than the previous speaker for the first utterance after the interruption.

Using the data for the social status of the speaker and the addressee, I also created a status differential. As the status category labels are numeric and ordered, this can be done by taking the difference between the two. For example, a king (status = 0) and a servant (status = 6) are distant in status, and thus will have a high status differential (here: 6). Between a king and a prince (status = 1), the difference is a lot smaller (here: 1). This absolute feature was automatically generated from the already annotated features.

A feature that had to be excluded is familiarity between characters (social distance). This data was not already available, and it was beyond the scope of this study to annotate this for all relevant character pairs. The literature has shown this to be a relevant feature. However, through the use of sentiment analysis, I have attempted to cover the complimentary and insulting aspects that could arise from high familiarity, and any lack thereof arising from low familiarity. Obviously, this does not cover all aspects of familiarity, but it means that this feature is not totally neglected.

Classification based on three algorithms

Three different algorithms are used for the classification task, namely Naive Bayes, decision trees and support vector machines. Whereas it would be ideal to achieve a high precision and recall score, the main goal of this research is to see whether it is even possible to predict the second person singular pronoun choice through a computational application *at all*. If this is indeed the case, what features contribute to this prediction? It is thus more important to verify which features influence the choice and to what extent they do so.

The reason for using three algorithms, and in particular these three, is their differences in learning biases and assumptions. Naive Bayes assumes all features are independent of one another, whereas decision tree attempts to create a dependent, hierarchical structure in the features. Support vector machine (SVM) is more complex and is able to combine both dependent and independent features. The addition of the latter algorithm will be particularly useful if the difference between the two simpler algorithm's models is small.

As well as applying three algorithms, I will also look at the difference between keeping *thou* and *thee* separate and combining them into the one category THOU. For this, I will run both a binary (YOU and THOU) and a trinary (*you*, *thou* and *thee*) classification, to see whether this affects the scores or changes which features are included in the best models.

Overview of implementation

I ran the three algorithms using the Waikato Environment for Knowledge Analysis (Weka[3]) software[4] with the default settings. The algorithms were run using a 10-fold cross-validation to ensure the best model based on training and testing of all folds combined.

The number of relevant instances of *you/thou/thee* extracted from the dataset is 22,932, which makes up 99.5% of the total number of such pronouns in the dataset. The pronouns were extracted using a Python script with simple heuristics. About 0.5% was missed due to noise in the dataset. The number of instances of *you/thou/thee* that were extracted from each play range from 363 (in *Macbeth*) to 811 (in *Coriolanus*).

I attempted to improve or maintain the scores while making the model simpler by excluding features, that is, through feature ablation. When there were conflicting changes in the scores, the scores of precision and F-measure were prioritised. I hoped to identify which features truly help predict the pronoun by building the simplest but best performing model. The baseline that the models were compared to is derived

3 Weka 3 - Data Mining with Open Source Machine Learning Software in Java, http://www.cs.waikato.ac.nz/ml/weka/.

4 In Weka, Naive Bayes is identified as NaiveBayesMultinominal, decision tree as J48, and support vector machine as SMO.

from the distribution of the pronouns in the dataset, thus 62.6% of YOU and 37.4% THOU.

I first took out groups of features that are related, rather than one feature at a time. Among the 23 features, I created six different groups. The first group related to the wider linguistic and social context (play, production date, genre, scene, location), while the second group was the closer linguistic co-text (n-gram). Information on the speaker (name, status, gender, age) and the addressee (name, status, gender, age, number of people) were groups 3 and 4. I kept status differential on its own, because it relates to multiple groups. Finally, the last group was sentiment (positive and negative). After the group ablation, I went back over the features to see if individual feature exclusions would improve the model further. This ensured the simplest and best model for each algorithm. The scores and the features included in each model are given in Tables 2, 3 and 4.

Results

Trinary classification scores

Table 2 shows the results of the trinary classification. As can be seen, each model performed significantly better than the baseline model, on all scores. The F-measure of the best model, the support vector machine model, is highlighted in bold.

Table 2: Scores for precision, recall, F-measure and accuracy for trinary pronoun prediction

Algorithm		Precision	Recall	F-measure	Accuracy
Baseline	Weighted Avg.	0.392	0.626	0.483	62.6417%
	you	0.626	1.000	0.770	
	thou	0.000	0.000	0.000	
	thee	0.000	0.000	0.000	
Naive Bayes	Weighted Avg.	0.826	0.826	0.826	82.64%
	you	0.880	0.885	0.882	
	thou	0.865	0.850	0.857	
	thee	0.509	0.510	0.510	
Decision Tree	Weighted Avg.	0.732	0.752	0.712	75.2093%
	you	0.738	0.960	0.835	
	thou	0.896	0.574	0.700	
	thee	0.408	0.097	0.157	
Support Vector Machine	Weighted Avg.	0.854	0.857	**0.854**	85.675%
	you	0.871	0.927	0.898	
	thou	0.919	0.836	0.876	
	thee	0.659	0.566	0.609	

Binary classification scores

Table 3 shows the results of the best models for the binary classification. The F-measure of the best model, again the support vector machine model, is highlighted in bold. This is also the best scoring model out of all models presented in this paper.

Table 3: Scores for precision, recall, F-measure and accuracy for binary pronoun prediction

Algorithm		Precision	Recall	F-measure	Accuracy
Baseline	Weighted Avg.	0.392	0.626	0.483	62.6417%
	YOU	0.626	1.000	0.770	
	THOU	0.000	0.000	0.000	
Naive Bayes	Weighted Avg.	0.868	0.868	0.867	86.8306%
	YOU	0.876	0.920	0.897	
	THOU	0.853	0.782	0.816	
Decision Tree	Weighted Avg.	0.818	0.818	0.818	81.8376%
	YOU	0.849	0.863	0.856	
	THOU	0.764	0.744	0.754	
Support Vector Machine	Weighted Avg.	0.872	0.873	0.872	87.2798%
	YOU	0.886	0.914	0.900	
	THOU	0.848	0.803	0.825	

Feature comparison of the models

Overall, the final models contain similar sets of features. The exact compositions are given in Table 4. What is surprising is that the binary classification model for the decision tree is very different from the other models: it does not contain any of the words from the n-gram as a predictor, whereas the others did.

Table 4: Features included in the best model of each algorithm

Algorithm	Type	Features included
Naive Bayes	Trinary	LW1, LW2, RW1, RW2, S_ID
	Binary	LW1, LW2, LW3, RW1, RW2, RW3, A_ID
Decision Tree	Trinary	LW1, LW2, RW1, RW2, S_ID, Stat_Diff, Neg_Sent
	Binary	Scene, S_ID, S_Gender, A_ID, A_Status, A_Age, Stat_Diff, Pos_Sent
Support Vector Machine	Trinary	LW1, RW1, S_ID, S_Age, A_ID, A_Age, A_Number, Stat_Diff, Pos_Sent, Neg_Sent
	Binary	LW1, RW1, S_ID, S_Age, A_ID, A_Age, A_Number, Stat_Diff, Pos_Sent, Neg_Sent

Discussion

This study has given some new insights into the analysis of pronominal address terms. Looking at the second person singular pronoun choice as a binary and a trinary classification problem resulted in slightly different outcomes. Even though the highest scores were achieved in the binary classification, one might still wonder whether this is the best method for addressing the second person singular pronoun choice. Looking back at prior studies on pronoun interpretation and comparing them to the features used in this study, we can conclude that *thee* and *thou* are equal in their opposition to *you*, with the main difference being their grammatical role. From the model comparison, we have seen that the co-text is most important when predicting the pronoun. This is evidence of the purely grammatical difference between *thou* and *thee* and their overall similarity in other aspects. Therefore, both linguistically and computationally, it makes more sense to perform a binary classification.

Differences between the algorithms were observed, but all three algorithms easily outperformed the baseline. The support vector machine models performed best, but the scores for the Naive Bayes models were quite similar to those for the SVM models. A choice between these approaches could be based solely on the scores for accuracy, precision, recall and F-measure, or also by taking into account the complexity, which is significantly higher for the support vector machine models. The more nuanced models that the support vector machine creates, which include more features than the models of the other algorithms, may suggest that the extra complexity of SVM models is indeed beneficial.

The best predicting features were the LW and RW features, which supports the importance of the direct linguistic co-text. In particular RW1 appeared as the most important feature in predicting the second person singular pronominal address term. Other important features were the speaker's name, addressee's name, status differential, positive sentiment and negative sentiment, with additional support from the speaker's gender, addressee's status, addressee's age, speaker's age, and number of people addressed. Only six features were not included in any of the models: genre, play, production date, location, speaker's status and addressee's gender.

I am, therefore, now able to falsify the null-hypothesis that it is not possible to build a reliable prediction model based on linguistic and extra-linguistic features. All six models demonstrate that linguistic and extra-linguistic features substantially improve the prediction of the pronominal address term, as all six outperform the baseline.

The second hypothesis, about which features would be good predictors, was partially correct in predicting that social status, age and sentiment would be included in the best models. However, none of these features were the main predictor of pronoun choice; that was the immediate co-text.

With regard to the final hypothesis, it has been revealed that the features are indeed both dependent on and independent of each other. However, since the Naive Bayes

models perform almost identically to the support vector machine models, we can say that the features are, for the most part, independent of one another.

Conclusions

The primary finding of this study is that it is indeed possible to build a prediction model for the use of YOU versus THOU with a singular referent in the plays of Shakespeare that is based on linguistic and extra-linguistic features. Moreover, in particular, the direct linguistic co-text of the second person singular pronoun is important. Other important features include the speaker's and addressee's names, status differential and both positive and negative sentiment. All in all this suggests that the pronoun choice is influenced by several linguistic and extra-linguistic features.

The best scoring algorithm and model was the support vector machine with 87.3% accuracy through its binary classification model.

For future research, I would recommend an exploration of other algorithms and features that were left out of this study, such as morphology, word embeddings and POS-tags. This will help us gain more information about the linguistic co-text directly surrounding the second person singular pronoun, which will likely give more insight into why this direct co-text is so important in deciding the choice of YOU or THOU. Moreover, including familiarity between characters (social distance) as a feature would be beneficial, as this has been noted multiple times in prior research as an influential factor, but was beyond the scope of this study.

Although this study has not yet provided a comprehensive set of all the linguistic and extra-linguistic features that influence the second person singular pronoun choice in Shakespeare's plays, it has definitely provided a more objective and extensive analysis of the matter that furthers the research into YOU and THOU.

Acknowledgements

The research presented in this article was conducted in collaboration with the Encyclopaedia of Shakespeare's Language project at Lancaster University. This project is funded by the UK's Arts and Humanities Research Council (AHRC), grant reference AH/N002415/1. The Shakespeare corpus will be made publicly available in Summer 2019, first via the CQPweb interface and then through download at a later stage. Many thanks to Jonathan Culpeper and the rest of the team for their advice and support throughout the study.

References

Literature:

- Bate, Jonathan, and Eric Rasmussen, eds. 2007. *William Shakespeare: Complete works*. London: The Royal Shakespeare Company.
- Brown, Roger W., and Albert Gilman. 1960. "The pronouns of power and solidarity." In *Style in language*, edited by Thomas A. Sebeok, 253–76. Cambridge: MIT Press.
- Busse, Beatrix. 2006. *Vocative constructions in the language of Shakespeare*. Amsterdam: John Benjamins.
- Busse, Ulrich. 2003. "The co-occurrence of nominal and pronominal address forms in the Shakespeare Corpus: Who says thou or you to whom?", in *Diachronic perspectives on address term systems*, edited by Irma Taavitsainen and Andreas H. Jucker, 193–221. Amsterdam: John Benjamins.
- Busse, Ulrich. 2002. *The function of linguistic variation in the Shakespeare corpus: A corpus-based study of the morpho-syntactic variability of the address pronouns and their socio-historical and pragmatic implications*. Amsterdam: John Benjamins.
- Calvo, Clara. 1992. "Pronouns of address and social negotiation in As You Like It." In *Language and Literature*, Vol. 1(1), 5–27. London: Longman Group UK Ltd.
- Greenblatt, Stephen, Walter Cohen, Jean E. Howard, and Katherine E. Maus. 1997. *The Norton Shakespeare: Based on the Oxford edition*. New York: W.W. Norton & Company, Inc.
- Mazeland, Harrie. 2003. *Inleiding in de conversatieanalyse*. Bussum: Coutinho bv.
- Mazzon, Gabriella. 2003. "Pronouns and nominal address in Shakespearean English: A socio-affective marking system in transition." In *Diachronic perspectives on address term systems*, edited by Irma Taavitsainen and Andreas H. Jucker, 223–49. Amsterdam: John Benjamins.
- Quennell, Peter, and Hamish Johnson. 2002. *Who's who in Shakespeare*. London: Routledge.
- Quirk, Randolph. 1974. "Shakespeare and the English language." In *The linguist and the English language*, edited by R. Quirk, 46–64. London: Edward Arnold.
- Stein, Dieter. 2003. "Pronomial usage in Shakespeare: Between sociolinguistics and conversation analysis." In *Diachronic perspectives on address term systems*, edited by Irma Taavitsainen and Andreas H. Jucker, 251–307. Amsterdam: John Benjamins.
- Taavitsainen, Irma, and Andreas H. Jucker. 2003. "Introduction." In *Diachronic perspectives on address term systems*, edited by Irma Taavitsainen and Andreas H. Jucker, 1–25. Amsterdam: John Benjamins.
- Thelwall, Mike, Kevan Buckley, Georgious Paltoglou, Di Cai, and Arvid Kappas. 2010. "Sentiment strength detection in short informal text." *Journal of the American Society for Information Science and Technology*, 61(12): 2544–58. https://doi.org/10.1002/asi.21416.
- Walker, Terry. 2003. "*You* and *thou* in Early Modern English dialogues: Patterns of usage." In *Diachronic perspectives on address term systems,* edited by Irma Taavitsainen and Andreas H. Jucker, 309–42. Amsterdam: John Benjamins.

Isolde van Dorst

YOU, THOU AND THEE: A STATISTICAL ANALYSIS OF SHAKESPEARE'S USE OF PRONOMINAL ADDRESS TERMS

SUMMARY

Much research has been undertaken on the use of *you*, *thou* and *thee* in Shakespeare's works. However, the results so far have yet to arrive at an exact and conclusive answer regarding how these pronouns were used. This study combines the strengths of multiple research fields in an effort to determine via hitherto unused computational methods which linguistic and extra-linguistic features influence the second person singular pronoun choices in the plays of Shakespeare. In the English of Shakespeare's time, the now-archaic distinction between YOU and THOU persisted, and is usually reported as being determined by relative social status and personal closeness of speaker and addressee. However, even between studies with similar outcomes, the results vary massively on the degree of influence and by the inclusion or exclusion of a wide range of other potential influencing factors. Therefore, it remains to be determined whether statistical machine learning will support this traditional explanation.

In this study, 23 linguistic and extra-linguistic features are investigated, having been selected from multiple linguistic areas, such as pragmatics, sociolinguistics and conversation analysis. The three algorithms used, Naive Bayes, decision tree and support vector machine, are selected as illustrative of a range of possible models in light of their contrasting assumptions and learning biases. Two predictions are performed, firstly on a binary (YOU/THOU) distinction and then on a trinary (*you*/*thou*/*thee*) distinction, giving six final models to compare. This is a strictly empirical study, which attempts to verify the findings of earlier research through a computational approach. Its aim and main focus is to try and find a pattern or model that best explains the use of second person singular pronominal address terms in Shakespeare, rather than simply achieve the best performing model.

The primary finding of this study is that it is indeed possible to build a prediction model for the use of singular second person pronouns in the plays of Shakespeare based on linguistic and extra-linguistic features. Moreover, in particular, the direct linguistic context of the pronoun is the most important feature in all of the models except one. Several other features are also influencing the pronoun prediction, including the names of the speaker and addressee, the status differential, and positive and negative sentiment. Additionally, all three algorithms easily outperformed the baseline. Out of the three algorithms, the support vector machine models score best. However, the Naive Bayes models perform almost equally well. This reveals that the features are, for the most part, independent of one another. When comparing the binary and trinary classification outcomes, the binary models scored better than the trinary ones.

Looking back at prior studies on pronoun interpretation and comparing them to the features used in this study, we can conclude that *thee* and *thou* are equal in their opposition to *you*, with the main difference being their grammatical role. Therefore, both linguistically and computationally, it makes most sense to use the binary classification.

Isolde van Dorst

YOU, THOU IN THEE: STATISTIČNA ANALIZA UPORABE IZRAZOV ZAIMKOVNEGA NASLAVLJANJA PRI SHAKESPEARU

POVZETEK

O uporabi zaimkov *you*, *thou* in *thee* v Shakespearovih delih je bilo opravljenih veliko raziskav. Vendar rezultati doslej še niso dali natančnega in dokončnega odgovora o tem, kako so se ti zaimki uporabljali. Študija združuje prednosti z različnih raziskovalnih področij, da bi z računalniškimi metodami, ki doslej še niso bile uporabljene, ugotovili, katere jezikovne in nejezikovne značilnosti vplivajo na izbiro osebnega zaimka druge osebe ednine v Shakespearovih igrah. V angleščini, ki se je uporabljala v Shakespearovem obdobju, je razlikovanje med YOU in THOU, ki je danes arhaično, še obstajalo. Običajno se navaja, da sta ga določala relativni družbeni status ter osebna bližina govorca in naslovljenca. Vendar pa se tudi med študijami s podobnimi rezultati ti zelo razlikujejo glede stopnje vplivanja ter upoštevanja ali neupoštevanja številnih drugih mogočih dejavnikov vpliva. Zato je treba še ugotoviti, ali bo statistično strojno učenje potrdilo to tradicionalno razlago.

V tej študiji se proučuje 23 jezikovnih in nejezikovnih značilnosti, izbranih z različnih jezikoslovnih področij, kot so pragmatika, sociolingvistika in analiza pogovora. Trije uporabljeni algoritmi – naivni Bayesov klasifikator, odločitveno drevo in metoda podpornih vektorjev – so izbrani kot ilustrativni nabor možnih modelov zaradi njihovih kontrastnih predpostavk in učne pristranskosti. Opravita se dve napovedi, prva o binarnem (*you/thou*) razlikovanju in druga o trinarnem (*you/thou/thee*) razlikovanju, s čimer dobimo šest končnih modelov, ki jih lahko primerjamo. Študija je strogo empirična, njen cilj pa je z računalniškim pristopom preveriti ugotovitve predhodnih raziskav. Osredotoča se predvsem na iskanje vzorca ali modela, ki bi najbolje pojasnil uporabo izrazov zaimkovnega naslavljanja za drugo osebo ednine pri Shakespearu, in ne le na oblikovanje modela, ki deluje najbolje.

Temeljna ugotovitev te študije je, da je resnično mogoče oblikovati napovedni model za uporabo zaimkov za drugo osebo ednine v Shakespearovih igrah na podlagi jezikovnih in nejezikovnih značilnosti. Poleg tega je neposredni jezikovni kontekst zaimka najpomembnejša značilnost v vseh modelih razen v enem. Na napoved zaimka

vpliva tudi več drugih značilnosti, vključno z imenom govorca in naslovljenca, razliko v statusu ter pozitivnim ali negativnim mnenjem. Vsi trije algoritmi so tudi z lahkoto dosegli boljše rezultate od izhodišča. Od vseh treh algoritmov daje najboljše rezultate metoda podpornih vektorjev. Vendar tudi modeli naivnega Bayesovega klasifikatorja dosegajo skoraj enako dobre rezultate. Iz tega izhaja, da so značilnosti večinoma neodvisne druga od druge. Primerjava binarne in trinarne klasifikacije je pokazala, da so rezultati binarnih modelov boljši od rezultatov trinarnih. Če primerjamo predhodne študije o interpretaciji zaimkov z značilnostmi, uporabljenimi v tej študiji, lahko ugotovimo, da sta zaimka *thee* in *thou* v opoziciji z zaimkom *you* enakovredna, pri čemer je najpomembnejša razlika njihova slovnična vloga. Zato je z jezikoslovnega in računalniškega stališča najbolj smiselna uporaba binarne klasifikacije.