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

Isolde van Dorst
1606778

MSc. Dissertation

Department of Intelligent Computer Systems
Faculty of Information and Communication Technology
University of Malta
2017

Supervisors:
Dr. Albert Gatt, Institute of Linguistics and Language Technology, University of Malta
Prof. Jonathan Culpeper, Department of Linguistics and English Language, Lancaster University
Dr. Gosse Bouma, Faculty of Arts, University of Groningen

Submitted in partial fulfilment of the requirements for the Degree of
European Master of Science in Human Language Science and Technology

Abstract

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.

M.Sc. (HLST)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITY OF MALTA

Declaration

Plagiarism is defined as “the unacknowledged use, as one’s own work, of work of another person, whether or not such work has been published” (Regulations Governing Conduct at Examinations, 1997, Regulation 1 (viii), University of Malta).

I, the undersigned, declare that the Master’s dissertation submitted is my own work, except where acknowledged and referenced.

I understand that the penalties for making a false declaration may include, but are not limited to, loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Student Name: *Isolde van Dorst*

Course Code CSA5310 HLST Dissertation

Title of work: *Thee, thou and you: A statistical analysis of Shakespeare’s use of pronominal address terms*

Signature of Student:

Date: 23-12-2017

Acknowledgements

I would like to thank the people from the Encyclopaedia of Shakespeare's Language project at Lancaster University for giving me access to their data as well as lending me their expert advice. Besides being thankful to Albert Gatt, Jonathan Culpeper and Gosse Bouma for their guidance as my supervisors, I would like to thank Andrew Hardie for giving feedback on my work too and explaining all the annotations in the Shakespeare corpus.

Table of contents

List of figures

List of tables

1 Introduction

1.1 Aim and goals of the thesis

For several decades a lot of research has been done on the use of *you*, *thou* and *thee* in Shakespeare's work. However, the results so far are inconclusive as to an exact answer to how these pronouns were used, with occasional contradictory results between studies. The aim of this study is to combine the strengths of different research fields and through their combination attempt to find out which linguistic and extra-linguistic features influence the pronoun choice used by Shakespeare. It will use the contributions of literary and linguistic studies to find which features could be relevant in this choice, and will utilise the tools and applications created for corpus linguistics and computer science to analyse the data in a more exact way than so far has been done in research. From a computational perspective, the pronoun choice and its influencing features can be treated as a classification problem, which allows for the use of machine learning techniques. The group of features will be taken as input to represent an instance of a pronoun, with the task to predict the pronoun that should be used. This can be done in different ways, and this study will compare a trinary classification in which all three pronouns are kept separate, as well as a binary classification in which *you* is separated from *thou* and *thee*. An explanation for this is given below. Through these techniques, we hope to find which features contribute to a more accurate prediction of the pronoun, to mimic the pronoun use of Shakespeare.

Since the first hypothesis, which is a null-hypothesis, is at the heart of this thesis, it is presented here as well:

> It will be too complex to build a model that fully predicts the pronominal address term solely based on linguistic and extra-linguistic features.

As I said, this is a null-hypothesis, that will hopefully be falsified through this thesis. In other words, the aim of the thesis is to create a computational model that can predict which pronoun should be used based on a set of linguistic and extra-linguistic features that are taken from the text itself and knowledge that we have of language in this time period. In order to achieve this, I will extract occurrences of *you*, *thou* and *thee* from Shakespeare's plays, and label every instance with 23 linguistic and extra-linguistic features. These features include speaker and addressee details (e.g. age, status, gender), play and scene data (e.g. play name, genre, location), and contextual information (e.g. the words used in close proximity of the pronoun). They cover a range of domains, from the purely linguistic context to the pragmatic and socio-linguistic, which includes considerations of the relative status of the characters. The selection of features is made based on prior studies on pronoun choice, which will be fully discussed in the next chapter. These features will be used in creating a prediction model through the use of three algorithms: Naive Bayes, decision tree and support vector machine.

1.2 Shakespeare as a writer

William Shakespeare, one of the better known writers in English history, was alive in the Early Modern English period. During this period, which runs from approximately 1500 to 1700, the English language lost its use of the words *thou* and *thee* as second person singular pronouns. For a writer of that time, a relatively large amount of Shakespeare's work has survived, approximately 1 million words. In his work, we can still find the use of these pronouns alongside the only remaining variant *you*. The fact that much of his work survived, more than that of his contemporaries, allows for a corpus-based, computational analysis. Additionally, his work was written less than 100 years before *thou* and *thee* had disappeared from the language, which means that the change was likely already in full progress around his time. Therefore, Shakespeare's work is particularly interesting when looking at the usage of these pronouns, since this use has changed since then. This will not only benefit literary studies on Shakespeare and linguistic research on language change, but more specifically, can contribute to the status of interdisciplinary research and show the positive effects of combining fields.

1.3 A note about terminology

For this thesis, I am looking at three pronouns; *you, thou* and *thee*. Whenever I am referring to these pronouns themselves specifically, I use the italics version as shown here. However, whereas *you* occurs in all cases, as it still is used today, *thou* was strictly a nominative form with *thee* as its accusative/dative form. *Thou* and *thee* are therefore related, and both occur as alternative variants for *you*. For this reason, throughout this thesis, I will often refer to the grouped forms THOU and YOU, where THOU includes both *thou* and *thee*. Whenever using the grouped pronoun variants, I will use small caps to signify this.

Additionally, I stray from the standard use of 'utterance' and 'speech act' as well, in particular in the Methodology chapter (chapter 3). With 'utterance' I mean the 7-gram in which the pronoun occurs. With 'speech act', I refer to a single but complete turn of a character as it is written in the play. This is done to be able to differentiate between these two easily.

It should also be noted that I left out any contracted forms of YOU and THOU, such as "prithee", because looking at such discourse and pragmatic markers was beyond the scope of the study.

1.4 Structure of the thesis

This thesis is divided into six parts. After this introduction, we will move onto the literature review in which previous studies will be laid out and compared, as well as a more detailed reasoning of why this topic is interesting to researchers in multiple fields. The thesis will then continue with the methodology of the experiments that are performed in this study, which is also where we will get a closer look at the different hypotheses that this thesis attempts to answer. The fourth chapter will present the results of the experiments explained in the methodology. Those results will be thoroughly discussed and compared to the literature in the discussion. Afterwards, the thesis will be summarised and ideas for future research will be given.

2 Background

What follows now will be a discussion of the field of Digital Humanities and a short introduction to EModE (Early Modern English). This will continue into a literature review in which the main studies on YOU and THOU will be discussed in both chronological order of the data and towards most recent and most relevant researches in their relation to the study laid out in this thesis.

2.1 Digital Humanities

Over the past few years computational research has evolved and with that, also branched out into other research fields that are not necessarily closely connected to computer science. An example of this is Digital Humanities (DH), which is an umbrella term used for all research that is computational and related to the disciplines of the humanities.

2.1.1 Why do we need Digital Humanities?

Since this is a new field, the exact term 'Digital Humanities' has not been fully defined yet:

> Along with the digital archives, quantitative analyses, and tool-building projects that once characteri[s]ed the field, [Digital Humanities] now encompasses a wide range of methods and practices: visuali[s]ations of large image sets, 3D modelling of historical artifacts, 'born digital' dissertations, hashtag activism and the analysis thereof, […], it can at times be difficult to determine with any specificity what, precisely, digital humanities work entails (Klein & Gold, 2016, p. ix).

The reason for the popularity of Digital Humanties, a cross-domain field of study, is that it does not try to diminish the differences between fields but actually operationalises this difference to solve issues that could not be dealt with within fields on their own. An example of a study in which two fields' strengths are combined is topic modelling, which is a type of text mining that traces word clustering and word co-occurrence in a corpus. Based on these clusters, the topic of the text is established (Gillings, 2016). Additionally, this study also stresses the importance of the researcher himself when interpreting the word clusters produced by the computational model. It thus keeps the balance between the computational side as well as the linguistic role in the study.

2.1.2 Approaches to and issues in Digital Humanities

It should perhaps be specified that Digital Humanities is not the simple digitising of humanities, in that it is not concerned with the mere scanning of already existing archives, but has an "increased emphasis on layerings of data" (Jones, 2016, p. 7). This is shown in the use of relatively[1] big data, textual corpora, geographic information system (GIS) mapped data, and the develop of research on new-

[1] 'Relatively' here is used to compare it to data sizes used in most non-digital humanities.

media objects (Jones, 2016). In order words, Digital Humanities is not just the combination of two already existing fields, but also a new field in itself that looks at digital networks.

Due to its wide range of research topics and partially undefined term, it is this not strange that there are critics of Digital Humanities. The main reason behind the concerns towards Digital Humanities is that most data models and infrastructure used in this field actually originated in other fields. In general, people assume that this means that they are not as well-suited for (Digital) Humanities research as for the research they were designed for. Take for example temporal data, which can be dealt with computationally easily, with a lot of precision, but this precision is not usually present in the data that DH works with. Most texts would be dated “ca. 1500s”, rather than having a specific year, let alone day, to use in the models. This, for obvious reasons, will make temporal models work less precisely than when given exact data (Posner, 2016). Another issue that ties in here is the use of *relatively* big data. This is not the same as “big data” as it is interpreted in most computer science studies, which sometimes work with billions of data points, depending on the domain. However, this is often not possible in, let’s say, studies on historical texts. This is for the simple reason that we do not have enough texts, artefacts, or objects to create a dataset of that size. To relate back to the first issue, the computational models are not usually designed for datasets of smaller sizes than those used in computer science or whichever field they were created for. The models might work in that they produce an outcome, but will not necessarily have a meaningful outcome that is reliable. These results will either produce an oversimplified representation of the world, or even be entirely inaccurate (Posner, 2016). It should be noted that, indeed, these are valid points, but they are methods that are used on top of the methods and representations that have already been used in Humanities outside of the digital aspect (Alvarado, in Robertson, 2016). The computational methods applied in Digital Humanities are, for the majority, not specifically designed for the research done in DH, but this does not mean that they cannot be helpful as an additional analysis on top of the traditional non-digital methods as long as the researchers keep this in mind.

In other words, one should address Digital Humanities as an ‘upgrade’ of research in the humanities through the addition of computer science. For example, in linguistic research, it does not make sense to apply machine learning clustering on a linguistic analysis of wordforms if the model has to do all the statistical summarisation of variables based on the words themselves (Hardie, 2017). As linguists, we also have the linguistic information behind those wordforms, such as POS-tags, semantic tags or even lemmatisation that could all be used in the clustering of wordforms. Since we already know that *a* and *the* are determiners, why not combine them already so that the model does not have to figure this out itself? By incorporating linguistic knowledge into the model, the model becomes more accurate from a linguistic perspective. Additionally, the computational modelling allows to fill any gaps left after forming hand-crafted rules, which tend to be incomplete. The final model thus shows the positive effects of combining the fields (Hardie, 2017).

In conclusion, the main idea to take away from this section is that while Digital Humanities is still new, evolving, and appeals to many different fields and research topics, it should be used as an enhancement of the current methods of research in the humanities, rather than replacing them. It is important to think about the combination of certain datasets with computational models, both in terms of the size of the dataset and the information that is actually used in the model. Using a computational model might appeal to researchers due to the 'fancy, new approach' that it allows, but if the researchers themselves are unaware of what a model or an algorithm actually does, it will not benefit their or anyone's research to its full potential. It is, as Andrew Hardie mentioned, important to keep in mind the balance between what the model can do and what the researchers themselves have to do (2017). The model can only provide a singular outcome, which still has to be interpreted by the researchers. This can only be done properly if the researchers have enough understanding of what the model does, and does not do (Hardie, 2017). Thus, the role of computational methods in fields of the humanities is that of a supporting character, and it should be kept in mind that the suitability of the model and its outcomes is as important as its interpretation.

2.2 Early Modern English and YOU/THOU

2.2.1 Early Modern English

The Early Modern English period runs from approximately 1500 until 1700. During this time, a lot of changes were made in the English language. Some of these changes were phonetic, such as the Great Vowel Shift which occurred right before and during this period. This shift caused the pronunciation of most vowels to alter in a push-pull effect, meaning that as one vowel's pronunciation shifted, it caused another to change as well (Baugh and Cable, 2013). This is just one example of a dramatic change that happened in this period.

Another important change in the English language around Shakespeare's time, which is of more importance to this thesis than the first mentioned change, is the standardisation of the language (Baugh and Cable, 2013). This means the introduction of the idea that grammar and spelling should be the same throughout the language, in spite of one's dialect. However, around Shakespeare's time, this was only just being introduced on a wider scale. Standardisation involved the development and spread of a social dialect that has prestige, which, perhaps lucky for Shakespeare, at this time was the most used or most favoured dialect of London and its area.

2.2.2 YOU and THOU in and before the Early Modern English period

Besides in writing and with spelling, people must have relied on social conventions and behaviour for basic conversation as well. An example of this is the difference between the second person pronouns YOU and THOU, which is a distinction that was lost after the Early Modern English period

(Taavitsainen and Jucker, 2003). The difference between the use of these two pronouns is evident from multiple literary studies that have been done both on Shakespeare's work, work of his contemporaries, and other documents from this era, such as Walker's (2003) and Busse's (2002) which will be discussed later in this chapter. These studies suggest that there were unwritten, social rules that have to be abided by in order to speak according to society's standards. This distinction of two forms of pronouns is referred to as the T/V distinction (Brown and Gilman, 1960), and has been studied often for its implications for politeness, which will be described later in this paper. It is named after the Latin pronouns *tu* and *vos*. The personal pronoun THOU seems to originate from the Germanic *þū*, which started as the only singular second person pronoun, while YOU was the plural form in Old English (Hickey, 2003; Adamson, 2001). The change in meaning from singular/plural to a sign of respect and intimacy is most likely caused by the influence of French. This is due to the already established T/V distinction in French and the prestige that it had in medieval and Renaissance England (Burnley, 2003; Mazzon, 2003). The usage of two different pronouns would thus have been a sign of status in which THOU would be used to inferiors and YOU towards superiors (Adamson, 2001).

The choice of pronoun, or address term, is a subtle form of showing respect or disrespect, and is often made with ease by native speakers. However, the criteria that affect the choice of pronoun are difficult to determine as they are culture and language dependent, but also change over time. Moreover, while one feature might become more prominent, another might lose its effect on the choice made (Taavitsainen and Jucker, 2003). It is thus important to keep in mind that a study focussing on pronoun choice, which is bound to its time period and the society it is used in, will provide different outcomes for other time periods or even other societies in the same time period. With society here is referred to both the group of language speakers as a whole, as well as groups of speakers within a language differentiated based on dialect or class.

William Shakespeare was born and wrote his work during the Early Modern English period, and thus used the pronouns YOU and THOU in his writing. His plays can be viewed as a representation of Early Modern English conversation and interaction, though often heightened in their use of emotional and dramatic instances to accommodate to the dramatic genre. In several of his plays, Shakespeare is thought to have specifically exploited the differences in social connotation between the two pronouns specifically for the dramatic effect they would have given to his writing (Hickey, 2003). It is thus not correct to assume that Shakespeare's use of pronominal switching is representative of the usage in common EModE (Calvo, 1992). However, it can be assumed that, though likely to be exaggerated, the pronominal usage in Shakespeare is representing social interaction and conversation from his time as explained above.

2.2.3 T/V distinction in other European languages

As already briefly noted, the T/V distinction was not only used in English, but also in other European languages that do still have these two different pronouns today, such as German *du* and *Sie*, Spanish *tu* and *Usted*, or Finnish *sinä* and *te*. The rough distinction between the two forms in the present day languages is that one is the more familiar, informal form (*du*/*tu*/*sinä*) and the other is more polite and distant (*Sie*/*Usted*/*te*) (Taavitsainen and Jucker, 2003). The loss of the T/V distinction in English is strange in comparison to the other European languages, which, as mentioned, do still have this distinction. Especially since the usage in all these languages appears to be based on formality and respect, the reason for losing the choice of pronoun in English is interesting. There is no absolute consensus on why it happened, but it was suggested that it could be due to simplification of verbal inflection, as an offset to the Quaker belief that THOU should be used as singular and YOU as plural, or that YOU was safer to use as it did not imply social superiority over the addressee (Walker, 2003). As mentioned before, the choice of address term is culture and language dependent, which makes it difficult to accurately compare the development of English second person pronouns to that of Spanish or German pronouns.

2.3 Studies on YOU/THOU

A lot of studies have been done on Shakespeare's use of YOU and THOU. However, most of these studies were literary studies (Barber, 1987; Calvo, 1992; Quirk, 1974), or did not implement any method beyond comparing raw frequency counts directly (Busse, 2003; Mazzon, 2003; Stein, 2003). In addition, these studies did not look at the total amount of plays available, but instead chose a few plays to focus on. Nonetheless, these studies have shown some patterns in the occurrence of YOU and THOU and thus provide a nice foundation for a more exact study on the usage of those two pronouns. One of the first studies on pronoun use is the one by Brown and Gilman (1960), which concludes that the choice mainly depends on power and solidarity. Their notion of pronominal usage is a static one, and although this has changed over time, it was a start towards the more recent approaches in analyses of pronoun use. The terms 'power' and 'solidarity' here refer to the social status of the characters and thus the power they hold over other characters, and the familiarity between characters, thus the relationship between them. The overall idea of power and solidarity influencing the choice of pronoun still holds, but has been shown to be more nuanced than was portrayed by Brown and Gilman (1960).

2.3.1 Middle English

David Burnley, who looked at T/V pronouns in late Middle English, thus right before the Early Modern English period started, constructed a diagram in which a binary choice of a list of features was made to determine whether THOU or *ye*/YOU should be used in the text. This diagram also shows a somewhat hierarchical structure, in which certain features are only used based on the choice of the feature before it. The diagram (Burnley, 2003) is shown below in Figure 1. He mentions that, overall, it worked relatively well, but was not always exact as some factors were more difficult to portray as binary. For example age and status are displayed here as older/younger and higher/lower, but do not take into account how much older/younger or higher/lower. If the age difference is very small, the effect is also likely to be smaller than when there is a large age difference.

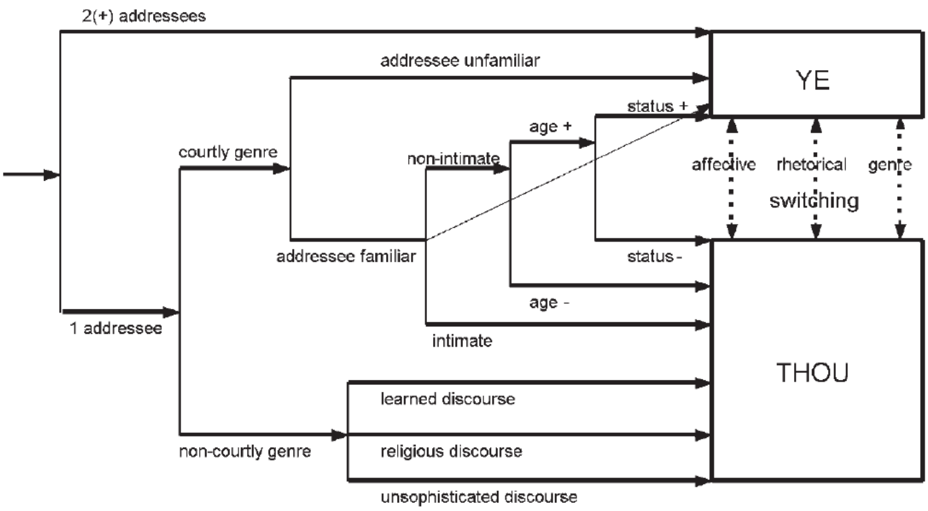

Figure 1: Burnley's diagram on Chaucer's use of second person pronouns

In addition, Burnley mentions that besides the interpersonal relations that this diagram takes into account, there are other factors that influence the choice of pronoun, such as the nature of the discourse or the gender (Burnley, 2003). He also refers to the notion of co-text, the interpretation of linguistic material surrounding the phrase or word in question that help define it. It differs from the context of said phrase or word, which is not necessarily in the text but contains information based on the genre of a text or the world knowledge of the reader. (Burnley, 2003; Wales, 2001) Based on the co-text, Burnley showed that there seem to be patterns in the choice of pronoun based on lexical choices in the phrase. When using *leve brother* or *sonne*, THOU is much more likely to be used, whereas *fader* is usually accompanied with the YOU-form (Burnley, 2003). This is partially shown in his diagram already through the familiarity, intimacy and age relations, but the lexical connection with the choice of pronoun might help predict this choice. Of course, it is unfortunately not as simple as that, as there are still many deviations from this division. Moreover, this study focussed on a different time period, one in which the division of YOU and THOU was not yet as loose and variable as in Early Modern English, which might be why there seemed to be strong evidence for the hierarchical binary structure that Burnley found.

2.3.2 YOU and THOU in the Early Modern English period

Terry Walker's study on the pronouns YOU and THOU in the Early Modern English period focuses on the occurrence of the pronouns in English dialogues (2003). She looks at recorded trials and depositions dialogues, as well as comedies and handbook dialogues. The two periods she uses to compare the pronoun usage on are 1560-1600, which overlaps with Shakespeare's time, and 1680-

1720, which is around the end of the Early Modern English period. She uses these two time periods exactly for this reason, as the occurrence of THOU alongside YOU was still frequent in the earlier time period, but in the later time period THOU had almost entirely disappeared (Walker, 2003). Her corpus contains just over 5,000 occurrences of the two pronouns combined (Walker, 2003), with an approximately even distribution of occurrences over the two time periods. In the earlier time period, the THOU occurrences make up 21.1% of the total pronouns, whereas in the later period, THOU occurred only 16.4% of the time (Walker, 2003).

She additionally found a difference in the occurrences between text types, which shows that genre could indeed matter when looking at the pronoun choices that are made. Perhaps most importantly, she finds a difference in the distribution of the pronouns between the constructed and authentic dialogues. This suggests that constructed dialogues do not necessarily represent true conversation of that time exactly. When comparing the findings based on gender, Walker found that men used THOU more than women, in both time periods. In the first time period, the gender difference is almost indistinguishable, but the second time period shows a massive difference as the usage of THOU by women has almost disappeared: from 25.2% in the first time period to 3.9% in the second time period of the occurrences are THOU.

Her main reasoning behind the findings is that the choice of pronoun is dependent on both genre and gender, which are influenced by the role of social status, both of the speaker and the addressee, and emotional changes in the conversation. Men seemed to be more influenced by the effects of the pronoun THOU while speaking, and used it to express contempt, as well as inclusiveness with higher ranked characters, thus “forcing an intimacy, or assert a dominance which he does not have” (Walker, 2003, p. 340), more often than female speakers. Moreover, YOU is the overall preferred pronoun in her data, and thus likely to be the more common form in both time periods.

In her later publication on an extension of the same research, Walker (2007) expanded her time periods to 1560-1780, and excluded the handbook dialogues from her text types. Besides looking at the effects of status, gender and age of the characters involved, Walker also describes her reasoning for not looking at several other factors, such as intimacy and emotion based on the subjectivity and the simple impossibility to discern between certain classes (e.g. when would one character change from a simply acquaintance to a friend?) (Walker, 2007). In the results of this study, she found that THOU is mainly used to convey negative attitudes such as an insult, and that there are massive differences between the genres, which she accredits to the differences in formality between them (Walker, 2007). The other results agreed with her earlier findings, thus again concluding that the pronoun choice depends on gender, emotion and genre primarily, with some influences of age as well. Over time, Walker noticed that the use of THOU became less frequent, which was also expected due to the disappearance of THOU overall from the language.

In addition, Walker found no support for the hypothesis that pronouns differ based on syntactic function, or in combination with certain verbs. Her hypothesis that THOU is more likely to co-occur with some closed-class verbs such as modals and private verbs than YOU, has thus been rejected after her study, though she mentions that more research should be done in order to confirm this. A similar comparison was made by Mulholland, who did not get any significance for concordances between the pronouns and any of the closed-class verbs (Mulholland, 1987). However, Mulholland did find that THOU occurred less often in statements and questions than YOU, which does suggest some relation between pronoun choice and syntactic function, even though it might not be very strong (Mulholland, 1987; Kielkiewicz-Janowiak, 1994).

Similar findings came out of both Walker's and Mulholland's research on the different forms of the two pronouns (*thee*/*thy*/*thine*/*ye*), which did not occur in concordance with certain verbs, or even suggested fixed expressions that were hinted at in prior research (Walker, 2007; Mulholland, 1987). Thus, Walker's overall hypothesis that the syntactic and linguistic factors would influence the pronoun choice less than the social and conversational factors was supported by her and Mulholland's research. She emphasises the need for further research, mainly on the pronoun use of lower class characters in drama, as this is the least investigated group.

2.3.3 YOU and THOU in Shakespeare and their relation with nominal address terms

Ulrich Busse (2003) looks at nominal and pronominal address terms in Shakespeare's plays and their co-occurrence. Similarly to an idea in Burnley's work, Busse tries to find a pattern in which the nominal used can be an indicator of the pronoun that should be used. He divided the nominal address terms into six categories; titles of courtesy, terms of address indicating occupation, terms of family relationship, generic terms of address, terms of endearment, and terms of abuse (Busse, 2003).

Surprisingly, in order to compare the co-occurrence of these nominal address terms to the pronouns, Busse uses the raw frequency count as his statistical measurement. Through using raw frequencies, it is difficult to see any patterns or correlations, and even more complicated to make reliable comparisons between frequencies in different categories, plays, or genres. This is problematic due to the difficulty of comparing numbers and basing conclusions on these frequencies. The absolute difference in occurrence between two categories, figuratively speaking, could be 120 occurrences of YOU and 35 of THOU. This makes it seem as if there are bigger differences between the categories for YOU than for THOU. However, if not using raw frequencies but instead something as simple as percentages, they might actually show that both YOU and THOU occur approximately 10% more in one category than the other, since THOU simply occurs less often in both categories. This is not measurable when only using raw frequency counts.

He additionally splits the counts based on genre, so that there is a count for the co-occurrence of every nominal address term with either YOU or THOU, divided per genre. In some cases, that means that the number of co-occurrences is very low, such as is the case for *wife*, which only has 17 co-occurrences with pronouns. Even with such low counts, Busse tries to come up with an explanation for the co-occurrence pattern, which he cannot find and therefore describes as "variable and not strictly correspond[ing] to a fixed pattern" (Busse, 2003, p. 201). Busse further applies the pronoun choice as a way of classifying nominal address forms based on the usage of the pronouns as well as the social status of the characters they were used for. In other words, a higher number of co-occurrences with YOU predicts a higher status for the character identified with the nominal address term, a higher number of THOU co-occurrences portrays a lower status, and a somewhat equal distribution shows that the nominal address term in question was used for people from different statuses. An example of this would be the nominal address term *sir*, which appears to have had a far higher occurrence with YOU, meaning that characters identified with this nominal address term were generally of a higher status (Busse, 2003). Again, these interpretations are based on the frequency counts of YOU and THOU in the co-occurrence.

Gabriella Mazzon (2003) looked at a similar pattern between nominal address terms and pronouns in Shakespeare, but looked more closely at the relationship between the characters. She looked at three of Shakespeare's plays; *King Lear*, *Othello*, and *Hamlet*, which are all Tragedies. What is particularly interesting about her study, is that Mazzon focuses on the fact that pronoun use in this time period was in a transitional phase, which can cause ambiguities and more specifically, means that exceptions are important and should be investigated rather than ignored.

2.3.4 The addition of (situational) markedness

Mazzon notes that in previous research the pronoun choice is often depicted as based on social status differences between the characters and the emotional aspect of conversation, such as anger or love (2003). As mentioned by Quirk (1974), YOU is seen as the statistically unmarked form. "It is not so much 'polite' as 'not impolite'; it is not so much 'formal' as 'not informal' " (p. 50). This means that THOU is the form that should be most interesting as it conveys information about the pragmatics and social aspect of interaction. This is supported by Wales, who describes language as a strategic tool, through which "characters on stage manipulate appropriate forms and conventions of language according to whether they are proposing marriage, or presiding over a legal case" (Wales, 2001, p. 205). In order to adjust to any given situation, language allows for changes in tone, formality, choice of lexis, rhetoric, etc.

However, markedness is something that cannot be used as black and white as is shown by Quirk (1974) and Wales (2001). As Mazzon correctly observes, "statements about what is 'normal' usage for any period of time for which we only have indirect and partial evidence should always be very cautious" (2003, p. 226). Most of the studies done on this topic have looked at small datasets of only a few plays, or the occurrence of these pronouns in combination with a limited set of factors that could influence it. There have been contradictory results because of this, but Mazzon's work is not really different from the works she critiqued. While Mazzon does acknowledge the importance of exceptions and notes the limitations of looking at a subset of the available resources, she also only analyses the pronoun usage in typical character relationships and thus does not solve the problem.

In this, she does find some suggestions towards the factors that influence this pronoun choice. Some examples are social class, intimacy between the characters, and age and gender differences. Age and gender are, according to Mazzon, placed within the factor of intimacy, as she found evidence for a different effect of intimacy based on age and gender (2003). Additionally, she found a different use of pronouns in situations where one character was trying to court another, thus a level of intimacy in which the characters are evolving towards a closer relationship. The role of social attitudes is most present in this study, as Mazzon links sudden pronoun changes to "voice, syntax, vocabulary and appearance" (2003, p. 138). Her conclusion is that we cannot view pronoun choice as a simple choice based on emotions, respect, and intimacy only, but that it is a more nuanced choice that is also highly influenced by the strategic reasons through which the speaker tries to appear more/less emotional, respectful or intimate with the addressee.

Stein analyses the "subtle linguistic indicators of social relations and the management of emotional states at a time" (2003, pp.251-252). He attempts to provide a reason for the changes in pronoun choice based on a study of two plays, *King Lear* and *As You Like It*, that addresses the pronouns as a marked or unmarked choice. This follows Quirk's perspective of pronominal use. However, contradictory to previous works, Stein shows that the 'unmarked case' is dependent on the situation, and therefore not by default YOU. Stein notes that the social relationship between two characters is often the main source of information to base a choice for the preferred, unmarked pronoun on (2003; Calvo, 1992).

In addition, his study focuses on the idea that the choice of pronoun depends on what would be the marked pronoun in every situation separately, rather than one of the pronoun forms being marked in all situations equally. A speaking character could use such a situationally marked form whenever he/she wants to deviate from the expected, unmarked form. "The marked form of address predominantly conveys a change of emotional state, which, on social level, may either create a feeling of like-mindedness or, exactly the opposite, a feeling of alienation towards the addressee" (Stein, 2003, p. 253-254). A change from the unmarked form to a marked pronoun would thus signal a semantic, temporary, break from the conventions of interaction, but the notion of markedness itself should be approached as a dynamic concept that is context-dependent.

In this study, Stein tries to find a statistical pattern that shows the social conventions that affected the pronoun usage. This approach is different from the other studies so far in that it is corpus-based rather than text-based. In other words, it treats every occurrence of the pronoun equally, and rather than selecting individual examples the focus of the study is more globally at every occurrence as a whole. Whereas other studies focus mainly on specific cases and examples to point out markedness, his study is "the study of the historical context of the socio-pragmatic facts and the demonstration, in an interpretive case study, of their exploitation in the hands of Shakespeare for the purpose of the management of emotions in emotionally charged drama" (Stein, 2003, p. 254).

Stein additionally notes that there are some differences between Shakespeare's plays and spoken language in Early Modern English as it was probably spoken. He mentions Salmon's (1967) remark on the quantity of hypotactic constructions in Shakespeare's work, which is much higher than in regular spoken conversation as there is a "strong tendency towards paratactic constructions" (Stein, 2003, p. 258) in spoken language. Moreover, and more important for this study, is the heightened use of marked pronouns, which is explainable through the characteristics of drama as a genre. The dense occurrence of events, often emotional, call for more changes in pronoun forms and thus also makes the use of marked or unmarked forms more relevant.

As social status is a big influencer of pronoun choice, changes in social status should be reflected in an analysis of pronominal usage. This is exactly what Stein did for several characters that drastically change status, such as Edgar in *King Lear*, who goes from 'aristocrat', to 'insane', to a 'lower class' character[2]. In each of these circumstances, he attempts to find the unmarked pronoun in interaction with different characters and bases his identification of markedness on this. For example, he finds that aristocrats among each other use YOU as their unmarked form as this is the most often used pronoun. However, when there is an emotional relationship between aristocrats, this influences their choice of pronoun and results in a more even distribution of YOU and THOU. In general, characters of lower status use YOU to address higher status characters, and higher status characters use THOU towards lower status characters. But between two lower class characters, THOU is most often used, which is in direct contrast with the usage of equal higher status characters. Any deviations from these found patterns are explained based on the emotional relationship between the characters, whether these are signs of affection or of anger. Overall, most marked forms occur between higher status characters or from lower status characters towards higher status characters, whereas the lower status characters were often addressed as expected. Stein also found some small differences between the patterns of the two plays he looked at, which he attributes to the roles of the jester, fool, and the insane (Stein, 2003).

Beatrix Busse (2006), who looks at the role of the vocative in Shakespeare like Ulrich Busse (2003) and Mazzon (2003), does so in more detail. Her corpus consists of 17 of Shakespeare's plays[3], which are selected based on their inclusion in the First Folio, as well as the differences between them based on genre and publication date (Busse, 2006). She contrasts her work with that of previous researches, as being a more elaborate work based on the fact that her corpus is much bigger than the more common choice to look at two or three plays at a time. The co-occurrence of vocatives and pronouns has been researched multiple times, and the more general result is that some titles do occur almost exclusively with one pronoun or the other, such as *lord* with YOU (Busse, 2006; Barber, 1987). However, the majority of the vocatives only show a small preference or just an overall preference

[2] Three of the classes chosen by Stein (2003).

[3] *A Midsummer Night's Dream*, *The Merchant of Venice*, *Much Ado About Nothing*, *Measure for Measure*, *The Winter's Tale, The Tempest, Richard III*, *Henry IV Part I*, *Henry VIII*, *Titus Andronicus*, *Romeo and Juliet*, *Hamlet, Othello, King Lear*, *Macbeth*, *Antony and Cleopatra*, *Cymbeline*

based on the comparison of types of vocatives (e.g. types of endearment against personal names). This, again, also depends on the genre they are used in, as the trends are not alike for all plays (Barber, 1987).

Busse stresses the fact that an in-depth study on Shakespeare's language is an interdisciplinary work that combines linguistic, pragmatic, sociological and cognitive research into one (2006). Stein (2003) emphasises the role of sociolinguistics and conversation analysis in a similar manner. Each of the fields capture a small part of interpretation, whether that's based on social rules and restrictions or human interactional behaviour. Conversation analysis is used here rather unexpectedly at first sight, as Shakespeare's work is all written. However, the genre of drama and the fact that they are play scripts designed to be read out loud and to copy natural speech, make conversation analysis a valid strategy of interpreting these texts (Busse, 2006). Moreover, conversation analysis and discourse analysis go hand in hand, which makes both suitable for the analysis of plays.

Busse's analysis on the use of pronouns and the vocatives in Shakespeare is from a mainly literary perspective in "historical and contextual dimensions" (2006, p. 50). As she mentions, "ultimately, it is a context that governs language usage, function, and meaning […] and it is a context that is construed by language" (Busse, 2006, p. 10). Thus, the choices made in language use are based on context. Additionally, she expresses the influence of researchers themselves on any description of language and language use, as it will be a subjective perspective of an individual user of said language (Busse, 2006). An example of this is mentioned in Mulholland's study: When looking at THOU as the marked form, Celia in *As You Like It* can be seen to use this marker to express her outgoingness and emotional side, whereas if YOU is viewed as the marked form, this would show Rosalind as markedly reserved and distant in the same play (McIntosh, 1963, in Mulholland, 1987). In other words, in a literary or even a sociolinguistic approach, the analysis of language is never context free, nor objective enough to fully capture the constant flux in the norm and its deviations.

2.3.5 Genre and Shakespeare as a representation of Early Modern English

When discussing research on second person pronoun usage in Shakespeare, Busse's study is probably the most elaborate and well-supported one (2002). His study focuses on the morpho-syntactic variability of the pronouns itself, rather than the occurrence of the pronouns in combination with another linguistic type. He argues that the reasons of solely looking at YOU and THOU rather than other combinations of pronouns/determiners is that these have already been resolved and explained in enough depth[4] (Busse, 2002). While the other combinations could be explained based on, for example, syntactic structure or phonological constraint, the usage of YOU and THOU appears to be more complicated.

[4] This refers to the combinations *you*/*ye, thy*/*thine,* and *thou*/*thee.*

Moreover, from multiple studies, including Busse's (2002), it has been found that there is a difference in pronominal usage over different genres, with the highest number of THOU forms in the Histories and the least in the Comedies. It should be noted though, that not everyone classifies the plays in a similar manner. Whereas the most common classes, the Tragedies, Comedies and Histories, are quite clear, there have been introductions of classes for Romance, Roman, or even Problem plays. Obviously, a difference in classification of the plays can result in a different distribution of pronominal use. Busse himself sticks to the traditional division between Tragedies, Comedies and Histories, unless discussing his results in comparison with another study that uses a different distribution.

He also raises the fact that we should keep in mind that although Shakespeare is likely to use language that is common for his time, his work makes up a literary corpus and is thus not necessarily truly authentic to the language used in interaction (Busse, 2002). Some choices in language use, and therefore also in pronoun usage, may have been made to match the rhyme or the metre, or even for the dramatic effect in a scene. We cannot simply assume that:

> … Shakespeare's usage of the pronouns of address is an exact mirror-image of that of society around him, [even though] it seems on the other hand indispensable to have a knowledge of communicative value of the "language-coins", i.e. the pronominal system and its social grammar in the historical context at large (Busse, 2002, 27).

The reason for still using Shakespeare's work as the corpus for this study is to gain a better understanding of his work in particular, but also to indirectly gain insight into EModE. As mentioned previously, there is not a lot of work that survived from the English Modern English period on common conversation of that time, and the speech used in plays is supposed to represent common, daily speech. Similarly to the development of Digital Humanities and the applications thereof, one needs to be aware of the issues that come with it. We cannot apply computational models to research in the humanities without knowledge of the model, and likewise, we cannot take Shakespeare's words to represent the entire EModE period. However, when creating the right balance between computer science and humanities, or Shakespeare and EModE research, we might be able to explore more than when viewing the two as completely separate.

Busse found that the number of THOU instances in Shakespeare's non-dramatic texts was significantly higher than that in the plays, as THOU was almost three times as likely to be used in non-dramatic work in comparison to drama. Partially, this might be due to the consistency of pronoun use in sonnets, as there were almost no changes from one pronoun to the other form within the sonnets (Busse, 2002). This can be explained through genre, as the poetic genre is more formal, and meant to be more private and thus read rather than spoken. Additionally, it was often intended for a loved one, and therefore the more personal and intimate pronoun THOU would have been preferred.

Overall, the results from Busse's research are broadly in accordance to that of previous studies. In terms of the distribution of the YOU/THOU occurrences, Histories use the most THOU and Comedies the least, with later plays containing less occurrences of THOU than earlier plays (Busse, 2002). The main exception here is *Henry VIII*, a historical play which only had 111 instances of THOU (*thou*: 86; *thee*: 25) in the entire play (YOU: 738)[5], which Busse attributes as possibly due to the collaboration with another writer on this play[6]. He also concludes that YOU was indeed the 'normal' pronoun overall, with the higher classes using THOU more than lower classes, though this is levelled out later in time (Busse, 2002).

2.4 Chapter conclusion

To sum up, there has been a lot of research on the use of singular second person pronouns in Shakespeare, both from a literary and a linguistic perspective. The findings overall support one another in that the pronouns appear to be used as linguistic markers that identify respect, social status, and familiarity. For example Gilman and Brown's notion of power and solidarity (1960), in one of the first big studies on YOU and THOU, gave a starting point for further research in the nuances of the pronoun as its use proved not to be as black and white as suggested in this first study. Mazzon (2003) and Quirk (1974) introduce the role of the pronoun as a linguistic marker, in which its usage can be seen as either marked or unmarked, though not as much in direct contrast as in previous studies. "It is not so much 'polite' as 'not impolite'; it is not so much 'formal' as 'not informal' " (Quirk, 1974, p. 50). However, even this does not fully explain the ways in which YOU and THOU are used. Stein (2003) and Calvo (1992) expand on this by concluding that the unmarked case is dependent on the context and the situation rather than just social status and familiarity between the characters. The emotive effect of the utterances is of importance as well, through which feelings such as anger and love for the other character can be expressed. Here respect also ties in, as it can be a disrespectful pronoun usage based on the addressee's social status that is the result of an angry remark (Stein, 2003).

A study on YOU and THOU in Shakespeare cannot and should not be limited to a single field of study, as was already mentioned by Busse (2006) and Stein (2003). It is a combination of literature, sociolinguistics, pragmatics, and conversation analysis, which are all useful in their own way to capture the complexity of the pronominal address terms and through them the social constrictions that came with these pronouns. Overall, the previous studies have shown a relatively reliable interpretation in the role of pronouns in that their conclusions are alike, though they do not fully cover all possible instances over all genres and all plays, nor do they discuss the strength of the different features that appear to affect the choice of pronoun.

The research laid out in this study is again a combination of multiple fields, to try and find a pattern or model that best explains the use of second person singular pronouns in Shakespeare. It is a strictly empirical study and will attempt to verify the findings of previous research through a computational

[5] These numbers are not identical to those used in the current study, but do show the same striking distribution of pronouns in comparison to other plays.
[6] John Fletcher.

approach. As mentioned prior by Busse (2006) and Mulholland (1987), the analyst will give a subjective perspective when attempting to interpret the role of language use. The use of a computational, statistical method is to create a more accurate and objective representation of Shakespeare's use of YOU and THOU in his plays. This will, hopefully, not only be more reliable in its findings, but also give an idea of the size of the effect of each feature that influences the pronominal address term choice. The application of a computational model to a linguistic study is to benefit from the combination of two fields. The enhancement of the analysis of YOU and THOU will allow for a more accurate comparison and a better supported prediction than has been possible so far, which is often the goal in Digital Humanities.

Additionally, it should be noted that Busse (2002) addresses the topic of variation vs. choice. As Schulze puts it, there is a difference between variation and choice, but also "a considerable overlap between the [two] concepts" (Schulze, 1998: in Busse, 2002, p. 8). In this study, pronoun usage is treated as a linguistic choice that is made, but the choice will be analysed as dependent on the variation of other linguistic features in the text, the co-text and the context. This ties in with Stein's approach: "The notion of variation employed here is not a statistical notion, but rather semiotically-based, relational and social-deixis-based" (Stein, 2003, 254).

The next chapter will go into more detail on the different computational methods that will be used as well as the reasoning behind the choice for these methods. A full description of the dataset and the features that will be tested for their influence on the pronoun choice are given in this next chapter as well.

3 Methodology

This chapter will go further into the data used in this study as well as the models that are used to analyse the data. The chapter will start with the hypotheses for this study. Next, a full description of the linguistic and extra-linguistic features is given and how these features were extracted from the data.

3.1 Hypotheses

The following hypotheses were made based on the literature:

> 1. It will be too complex to build a model that fully predicts the pronominal address term solely based on linguistic and extra-linguistic features.

Even though there is a lot of research that claims to know exactly what pronoun choice is based on, there are many differences between them. Thus, it is not likely that a single model will be able to fully predict which pronoun should be used based on linguistic and extra-linguistic features. This is because these features are likely not the only factors that potentially influence the choice of pronoun, as other studies have shown that sociological or situational and contextual knowledge might change the decision as well. However, this, being a null-hypothesis, is exactly what we wish to falsify through this study. Additionally, it is difficult to say what would be a "good" model in this case, as there is no prior study that created a model such as will be attempted in this study. There is no set ceiling for this experiment, but it is unlikely to be 100% due to the complexity and the nuances of the pronoun choice that were found in previous studies. Because of this, an estimate of the baseline will be made through the relative proportional division of the pronouns in the entire dataset, which is 37.4% percent THOU and 62.6% percent YOU. The aim of this study, when assessing this hypothesis, is thus to create a model that performs better than a baseline that works solely on the distribution of the pronouns in the dataset by adding the knowledge of the features to the model.

> 2. Some features will be a better predictor of pronoun choice than other features.
> 2A. Social status, age and sentiment will show a greater influence on the pronoun choice.
> 2B. There will be no measurable influence of genre, production date or gender on the pronoun choice.

The model will show a hierarchy in the linguistic and extra-linguistic features that are kept in the final model. As suggested by the literature, social status, age and sentiment will be part of the most influential features, as these features have been showing up most reliably in previous research. On the other hand, the effect of genre, production date and gender are not likely to be included in the final models, even though the literature has shown some support for these features. My reasoning for this assumption is that, firstly, the role of genre is dependent on the division of the plays, which is not a clear division that all research agrees on. Any differences in pronoun choice might therefore not be noticeable between genres. Production date would, over the entire EModE period, definitely show an effect, but the time span of 22 years in my study will not show enough difference in the pronoun occurrence. Lastly, gender could be an interesting feature, but is not as strong as some of the other features and will therefore be lost in the model. This would relate back to Walker's study (2003), where she found that the pronoun choice between genders did seem to differ slightly, but not as much as some of the other features.

> 3. The algorithm to make the most accurate prediction and show the influence of each linguistic feature best will be the support vector machine.

The complexity of the support vector machine and the option to alter the decision boundaries it creates between the classes will prove to be the most reliable. Additionally, support vector machine models keep an as large as possible margin on their decision boundaries, which allows for new data to fit the model as well. This gives a support vector machine model a certain robustness that might not be as strong as in the other models. It will also work best with the size of this dataset as the number of linguistic features is relatively small and the number of test items is of medium size.

3.2 Data

The data that is used for this study comes from the Encyclopaedia of Shakespeare's Language project[7], which will be described below. Then a more detailed description of the linguistic features already present in this dataset as well as the additions I made will follow.

3.2.1 Encyclopaedia of Shakespeare's Language project

The Encyclopaedia of Shakespeare's Language project is a research project of Lancaster University in the United Kingdom. Their corpus consists of 38 of Shakespeare's plays, which are all fully annotated for speakers, act and scene, as well as other features which will be discussed below. They also provided spelling normalisation throughout the corpus, making it both easier to read and to process the text. An example of the spelling normalisation and the XML format that contained all their annotations can be seen below in Figure 2. Spelling standardisation was only used to get rid of variants of the same word, such as *burie* and *bury*, which were both *bury* after normalisation. This makes the corpus more regular throughout, allowing for more reliable frequency counts and grouping of the data. It is

[7] More information on the project can be found on http://wp.lancs.ac.uk/shakespearelang/.

important to note here that the difference between THOU and YOU was kept as it was originally used in the text. All of these annotations were made for the ease of using corpus tools, which is what this corpus was created for.

```
<scene n="2" type="scene" title="" id="RJ_2_2">
    <lb/>
    <u who="RJ_Romeo" label="Rom.">
        He
        <normalised orig="ieats" auto="false">jests</normalised>
        at
        <normalised orig="Scarres" auto="false">Scars</normalised>
        that
        <normalised orig="neuer" auto="false">never</normalised>
        felt a wound,
        <lb/>
        But soft, what light through yonder window breaks?
        <lb/>
        It is the East, and
        <normalised orig="Iuliet" auto="false">Juliet</normalised>
        is the
        <normalised orig="Sunne" auto="false">Sun</normalised>
        ,
        <lb/>
```

Figure 2: Snippet of annotated text from *Romeo and Juliet*, Act 2, Scene 2

The 38 plays include the 36 plays that are in the First Folio, with the addition of *The Two Noble Kinsmen* and *Pericles: Prince of Tyre*. These last two plays were added based on the fact that they were included in earlier publications of the First Quartos. A list of all plays in the corpus can be found in Appendix A, and a broadly annotated version of the full Shakespeare corpus can be found on CQPweb[8] (Hardie, 2012). The corpus data for the Shakespeare project was based on the version provided by Internet Shakespeare Editions (ISE)[9]. Some of the annotation and all of the abbreviations used for the titles of the plays follows *The Arden Shakespeare*.

The reason for using the First Folio is that it is seen as an important edition. It was published seven years after Shakespeare's death and included 20 plays that were not printed elsewhere before. In addition, almost none of the original manuscripts survived, which makes the First Folio the main resource of Shakespeare's work we have today (Culpeper, in preparation). It is also a more reliable version than the First Quartos, which were often imprecise reconstructions of the plays, whereas the First Folio was constructed out of multiple manuscripts and good quartos to get the best and most likely version that was used to act out the plays. The division of the plays in three different genres, comedies, tragedies and histories, was not made by Shakespeare himself, but at a later stage by the people who composed the First Folio. This again also differs from edition to edition. The people at the Shakespeare project came up with the most likely division for their corpus.

[8] http://cqpweb.lancs.ac.uk

[9] http://isebeta.uvic.ca/Foyer/quality/

3.2.2 Linguistic and extra-linguistic features

Feature	Acronym	Annotation	Example
Genre	Genre	Pre-annotated	Romeo and Juliet is a Tragedy (Can be Comedy/Tragedy/History/Roman)
Play name	Play	Pre-annotated	"RJ" identifies *Romeo and Juliet*
Play, act and scene	Scene	Pre-annotated	"RJ_2_3" identifies *Romeo and Juliet*, Act 2, Scene 3
Speaker ID	S_ID	Pre-annotated	"RJ_Romeo" identifies Romeo from *Romeo and Juliet*
Speaker gender	S_Gender	Pre-annotated	RJ_Romeo is labelled as "male" (Can be male/female)
Speaker status	S_Status	Pre-annotated	RJ_Romeo is scored "1" (Score from 0 to 7, 0 being highest status)
Production date	Prod_Date	Pre-annotated	*Romeo and Juliet* is labelled as "1595"
N-gram	LW1-3, RW1-3	Automatic	Identifies the context/co-text, has three words left and three words right of the pronoun, each word treated as separate
Positive sentiment	Pos_Sent	Automatic	Score from 1 to 5, 5 being the strongest
Negative sentiment	Neg_Sent	Automatic	Score from -1 to -5, -5 being the strongest
Speaker age	S_Age	Manual	RJ_Romeo is labelled as "younger" (can be younger/adult/older)
Location	Location	Manual	Can be "private" or "public"
Addressee ID	A_ID	Automatic	"RJ_Juliet" identifies Juliet from *Romeo and Juliet*
Addressee gender	A_Gender	Pre-annotated	RJ_Juliet is labelled as "female" (can be male/female)
Addressee status	A_Status	Pre-annotated	RJ_Juliet is scored "1" (Score from 0 to 7, 0 being the highest status)
Addressee age	A_Age	Manual	RJ_Juliet is labelled as "younger" (can be younger/adult/older)
Status differential	Stat_Diff	Automatic	Difference in status between speaker and addressee, RJ_Romeo's status – RJ_Juliet's status = "0"
No. of people addressed	A_Number	Pre-annotated	RJ_Romeo is "singular", R3_Lords is "plural" (can be singular/plural)

Table 1: List of all features used in this study

As mentioned above, the corpus that the people of the Encyclopaedia of Shakespeare's Language project have created is richly annotated. However, this did not include all features that I want to include in this study. Some of the features that were already annotated are identification for all speakers with a personal code, POS-tagged sentences, and spelling standardisation. There are also annotations to separate between plays, and every act and scene within the plays.

Other features that were already annotated were the social status and the gender of all characters. The social status incorporates the social hierarchy of Shakespeare's period that we know of. The scale runs from 0 to 7, with 0 being the highest status and 7 being the lowest. An example for every category can be seen in Table 2 below. Most of the examples are from *The Merchant of Venice*, except for the last class, which is from *A Midsummer Night's Dream*, as there are no supernatural characters in *The Merchant of Venice*. For the Roman plays[10], some changes in the basic format were made for the specific characters that do not occur outside of the Roman plays, such as a Senator, who is labelled with a similar status as the Gentry (status 2, here MV_Lorenzo). These characters range on the same scale but are mentioned separately as they do not entirely fit the social classes from the non-Roman plays.

Social status	Explanation	Character example
0	Monarchy	MV_Duke
1	Nobility	MV_Portia
2	Gentry	MV_Lorenzo
3	Professional	MV_Shylock
4	Middling	MV_Tubal
5	Commoners	MV_Leonardo
6	Lowest groups	MV_Giobbe
7	Supernatural beings	MND_Titania

Table 2: Social status annotations

Another feature that might need some explanation is the use of the production date. Since the use of THOU became less common and disappeared over time, I want to see if there is any relation between the pronominal address choice and the date the play was performed. There is also a publication date available for the plays, but these are usually from after the performances were already completed. In other words, the more reliable date to pinpoint when the play was written, is the production date. It should be noted that the span of the production dates is only 22 years in this study, which might be too small to be a predictor as it is too narrow a period to actually evince language change, even more so when looking at just one writer as for this study. So far, there has not been any conclusive research

[10] Roman plays are the plays set in ancient Rome. An example of one of those plays is *Anthony and Cleopatra*.

that showed a difference over time, but this study will reveal if perhaps in combination with other features it does.

In order to do a full analysis of what extra-linguistic features could be predictors of the pronominal address term, some additional annotation was necessary. As a referent such as a pronoun is dependent on both context and co-text (Burnley, 2003), I decided to use the utterance itself that the pronoun occurred in as a feature to test the role of co-text in the pronoun choice. For this, I took the 7-gram the pronoun occurred in, which gave me six co-textual words. An example of this can be seen below in Table 3, with “LW” describing 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, thus LW3 is the third word on the left of the pronoun. The reason for using the n-gram rather than a full speech act, is the high occurrence of multiple pronouns within a speech act. This would mean that I would have to either use the same co-text for several pronouns, or exclude all utterances that have more than one pronoun in it. The use of a n-gram avoids this problem, as even pronouns that occur close together will have a small difference between their n-grams, possibly with some words overlapping. If a pronoun occurred in a position that would have the 7-gram extend into the previous or next speech act, the n-gram was cut off at the speech act edge to ensure it only includes words that are from the same speaker. An example for this can also be seen in Table 3, in which the position of RW3 is empty as the utterance ends with the word in RW2.

The reasons for using a 7-gram over other sizes is that it is relatively normal in corpus linguistic studies. In corpus linguistics, looking for collocations is often done with a three-word-window, meaning there are three words on either side of the word of interest, which creates a n-gram. While I am not necessarily looking at the specific collocations, the 7-gram will be used to look at similarities and differences in contextual words surrounding the pronouns to see if they can predict the pronoun choice.

Context	LW3	LW2	LW1	Pronoun	RW1	RW2	RW3	Context
With more of thine	this	love	that	thou	hast	shown	doth	add more grief to
That feel no love	in	this	dost	thou	not	laugh	‘ ’	</u>

Table 3: Example of two n-grams, surrounded by context (from *Romeo and Juliet*, Act 1, Scene 1)

Another feature that has been mentioned as important in the literature is the role of sentiment. With this I mean the use of the pronoun in some context to convey positivity or negativity, such as a compliment or an insult. Sentiment has been annotated with the use of the same n-gram as above. SentiStrength is a lexicon-based sentiment analysis program that scores phrases with a score for positivity and negativity (Thelwall, Buckley, Paltoglou, Cai & Kappas, 2010). These scores are based on the scores of every individual word in the phrase, and overall give a score to represent whether the phrase is negative, positive or neutral. Thus, a sentence containing three words with expressed sentiment will have a higher score overall than a sentence where only one word expresses sentiment. Originally, SentiStrength was created for Modern English, and would thus not accurately represent sentiment in Early Modern English. However, for a study by Culpeper, Findlay, Cortese and Thelwall (under review), this algorithm was altered for Early Modern English. They changed the scores of positivity and negativity for several words to convey the strength and the sentiment as it was used in EModE, as well as added words from this time period that were not in the Modern English version. Since SentiStrength was developed to work with online comments rather than perfect sentences, it works just as well with n-grams, which are not always the most sensible phrases (Thelwall et al., 2010). The final scores that will be used to determine the level of sentiment in the n-grams are the scores for positivity and negativity, which are incorporated as separate variables. For example, if a n-gram is scored a -3 for negativity and a 4 for positivity, the overall score would have been 1, which shows that the positivity in the n-gram is stronger than the negativity, but neglects the fact that both are quite high scores and thus represent high emotional value. This is why these scores are kept apart and treated as separate features.

The corpus already included quite a lot of data on the speakers, such as gender and status, as well as a unique code to differentiate all the characters. However, I wanted more details to accommodate some of the other features that were described as predicting the pronoun, notably, age. This was one of the most complicated features, as there is often no specification of age for the characters, except for when it is an important attribute of the character. Besides the main text from the plays itself, Quennell and Johnson's (2002) description of all characters were used in the classification. 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. The distinction here is made based on how important said character is in the play as well as the importance of the child's character. For example, 'Romeo' and 'Juliet' from *Romeo and Juliet* were classified as 'younger', as their youth, especially for 'Juliet', is important for their characters. Thus, their parents were classified as 'adult'. However, in *The Merchant of Venice*, I classified both 'Jessica' and her father 'Shylock' as 'adult', since they both act as adults, and there are no mentions of either of them being of a different age than the other characters they interact with. Since 'Jessica' is also not a major character in the play, the distinction between them is not as important even if she is technically younger than him as she is his daughter. Characters that returned in multiple plays, such as 'Falstaff', only changed categories if they were identified as having aged between the plays.

A more global feature to include is the location where the scene is set. This was another difficult feature to annotate, due to the stage directions that are not always reliable. Therefore I used a binary annotation of 'public' and 'private'. Besides the text itself to determine the location based on what characters said about their location, Bate and Rasmussen's (2007) annotation was contrasted with

Greenblatt, Cohen, Howard and Maus' (1997) annotations. The use of these three resources allowed for manual annotation of location for every scene. Some of the annotations were easier than others, for example someone's bedroom is a private place, whereas a tavern is public. Overall, the decision for either was made based on how likely an unrelated person would be to walk in. For example, in a residence, one could expect a relative to walk in or perhaps a messenger, but not a complete outsider. Whereas in a forest, anyone would be able to walk onto the scene. Therefore, residential locations would be classified as private and more open locations such as forests would be public.

In order to keep a conversation analysis perspective in mind, we need the information not only of the speaker and the scene, but also of the addressee. Conversation analysis theory is important here as the speech in plays is supposed to mimic natural speaker-addressee interactions (Busse, 2006). While the corpus is already annotated for every speaker, this still had to be done for addressee. As a manual annotation would be incredibly time consuming, I decided to use an automatic method to treat the previous speaker as the addressee. 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. In other words, it is often the case that the current speaker will address the last speaker under the expectation that they will be the most likely to respond. In addition, there is the idea that conversations happen in sequence and are based on responses in adjacency pairs (Drew & Curl, 2008). While annotating this, if the utterances were interrupted by the start of a new scene or a stage direction (e.g. someone walking into the scene), the addressee would be the next speaker rather than the previous speaker for the first utterance after the interruption. This might not always be accurate for every instance of YOU or THOU, but will overall be relatively accurate.

Based on the speaker and the addressee and their social status, we can also create a status differential, which will show the difference in status between them more clearly. For example, a king (status = 0) and a servant (status = 6) are socially far apart in terms of status, 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 feature was automatically generated from the already annotated features.

Additionally, the annotated speaker ID's will be used to establish whether the addressee is a singular or plural entity. This has been pre-annotated for every speaker ID, and is especially useful for cases in which YOU is used to address multiple characters rather than as a linguistic choice of pronoun, since THOU cannot be used in such a situation according to the literature.

A feature that had to be excluded from this study based on the time frame is familiarity between characters. 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 to represent low familiarity. Obviously, this will not cover all aspects of familiarity, but does show that excluding familiarity as a distinct feature does not limit this study as much as leaving out any of the other features might have.

Another feature that was excluded was the number of speakers in a scene, thus identifying whether an utterance was spoken in a monologue, dialogue or a conversation with three or more speakers. This feature was excluded based on the time constraints of this study.

3.3 Classification based on three algorithms

In order to attempt to classify each n-gram and hopefully correctly predict the pronoun based on the features described above, three different algorithms will be used. It should be noted that whereas it would be great to get a high precision and recall score, the main goal of this research is to see whether we can even predict the pronoun choice *at all*. If this is indeed the case, what features contribute to this prediction? It is thus more important to find out which features influence the choice and to what extent they do so. Therefore the main scores we will look at are precision and F-measure, as we are interested in the probability that a prediction is correct rather than just the probability that a correct prediction is retrieved.

The reason for using multiple algorithms is their difference in assumptions and learning biases to see if this affects the prediction. The features will be tested for the influence through feature ablation, thus by starting with a model that contains all features and trying to improve this model by taking features out. Through this, we want to show which features truly help the prediction of the pronoun by helping it beat the baseline. The baseline is based on the distribution of the pronouns, thus 37.4% THOU and 62.6% YOU.

On top of the three algorithms, we will also look at the difference between keeping *thou* and *thee* separated and combining them into the one category THOU. Because of this, we will run both a binary (YOU and THOU) and a trinary (*you*, *thou* and *thee*) classification, to see if this might influence the scores or the features that will be left in the best models.

Below follows a description of the relevance of the three algorithms that will be used.

3.3.1 Naive Bayes

Naive Bayes assumes that all the features that are put in the model are independent of one another, and also hold equal value in their weight affecting the prediction of the pronoun. So even if one feature might be a better predictor than another, every feature contributes equally to the overall probability of a given prediction outcome. The learning of a Naive Bayes model works through a simple multiplication of probabilities for all the features. The formula for a Naive Bayes model can be seen below in Figure 3. Even in cases where the assumptions do not hold, it usually comes up with a relatively decent model. However, it is unable to learn any interactions between the features, which might make it difficult for the model to deal with contradicting features or even features that are strengthened through interaction with others (Chen, 2011).

$$\hat{s} = \underset{s \epsilon S}{\operatorname{argmax}} P\,(s) \prod_{j=1}^{n} P(f_j|s)$$

Figure 3: Formula for a Naive Bayes model (Jurafsky & Martin, 2009)

3.3.2 Decision trees

The main contrast with the Naive Bayes is that in decision trees, the features are assumed to not be independent of each other. The algorithm will make a decision to divide the data according to a feature's values. This splitting criterion will give us more information about the class to which the instances belong, and also provides the tree structure of this algorithm. For example, in a case where a features X, Y and Z all make split '*a*', it could result in a pronoun prediction for THOU, whereas a different split than '*a*' for those features does not necessarily end up with the same prediction. This example is shown below in Figure 4. In addition, decision trees make no assumptions on the distribution of the data, which means they easily deal with any outliers in the data. A big plus of decision trees is that they clearly show a hierarchical structure of the features. A possible disadvantage is that a decision tree can quickly become overfitted, so some more work is needed to avoid this, perhaps by using ensemble methods such as random forests, which are fast and do not need much parameter tuning (Chen, 2011).

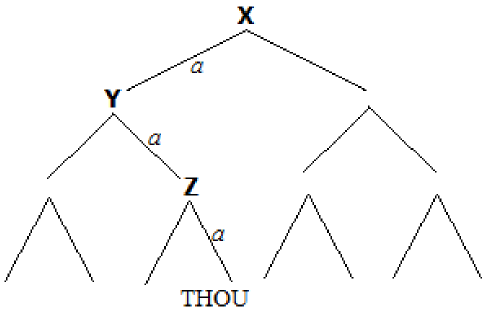

Figure 4: Example of a decision tree where features X, Y, and Z predict THOU when making split '*a*'

3.3.3 Support vector machines

Of all the algorithms mentioned here, support vector machines (SVMs) are the most complex. It is a similar approach to logistic regression, but unlike logistic regression, SVMs do not output a probability score. Instead, a support vector machine will make a direct prediction of the pronoun by minimising the cost function (Ng, n.d.). The easiest way to describe a SVM is that it tries to discover a line to divide the two classes based on the features. While this is what logistic regression does too, support vector machines works with several linear *p*-dimensional lines (which creates so-called hyperplanes) to replace the logistic cost function of logistic regression. In addition, it often has a high accuracy score and it is powerful enough to be unlikely to overfit to the training data. It can give a cleaner and more powerful model than logistic regression (Ng, n.d.). However, as it is more complicated, it can be hard to interpret the results and is rather memory-intensive to run in comparison to the other algorithms (Chen, 2011).

3.4 Miscellaneous

The three algorithms were run using the Waikato Environment for Knowledge Analysis (Weka) software[11]. This is a freely available program that runs on Java and is able to support several classification models, including the three described above[12]. The algorithms were run using a 10-fold cross-validation to ensure the best model based on training and testing of all folds combined. This makes my training set 90% of the total dataset for each fold, and 10% my test set.

As mentioned before, the distribution of the pronouns in the dataset is that there is 37.4% THOU (*thou* 23.9%; *thee* 13.4%) and 62.6% YOU. In addition, the number of pronouns extracted from the dataset is 22,932, which makes up 99.5% of the total number of 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.

3.5 Chapter summary

In summary, this chapter explains the dataset, where it came from, and how and why each linguistic feature was extracted. It also went into more detail on the three algorithms used in this study: Naive Bayes, decision trees and support vector machines. In short, these three algorithms are chosen for their differences in learning biases and their approach to the features as individual or related influences on the classification.

Additionally, the hypotheses of this study were laid out, which predict that none of the models will be completely successful in building an accurate model, but that they will show which features are most accurate and therefore most salient when predicting the pronominal address term. It is most likely that the best algorithm for this will be the support vector machine due to its power and learning biases.

In the next chapter, the results of this study will be laid out. Each algorithm will be discussed separately and some first interpretations will be made of the results based on a comparison among the models.

[11] Download is available from: http://www.cs.waikato.ac.nz/ml/weka/.

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

4 Results

In this chapter I will present the results of the study and briefly discuss some of the first findings. A more thorough discussion of the results will be reserved for the next chapter. Firstly, a short look at the distribution of the pronouns will be taken. Next, there will be an overall presentation of all the scores for recall, precision, F-measure and accuracy obtained by the various classifiers, with an explanation of the feature ablation approach used to reach these results. The confusion matrices for the algorithms will also be shown and discussed here. Afterwards, a more detailed comparison of the three algorithms will follow in which we will look at the features that are in each best model.

4.1 Pronoun distribution

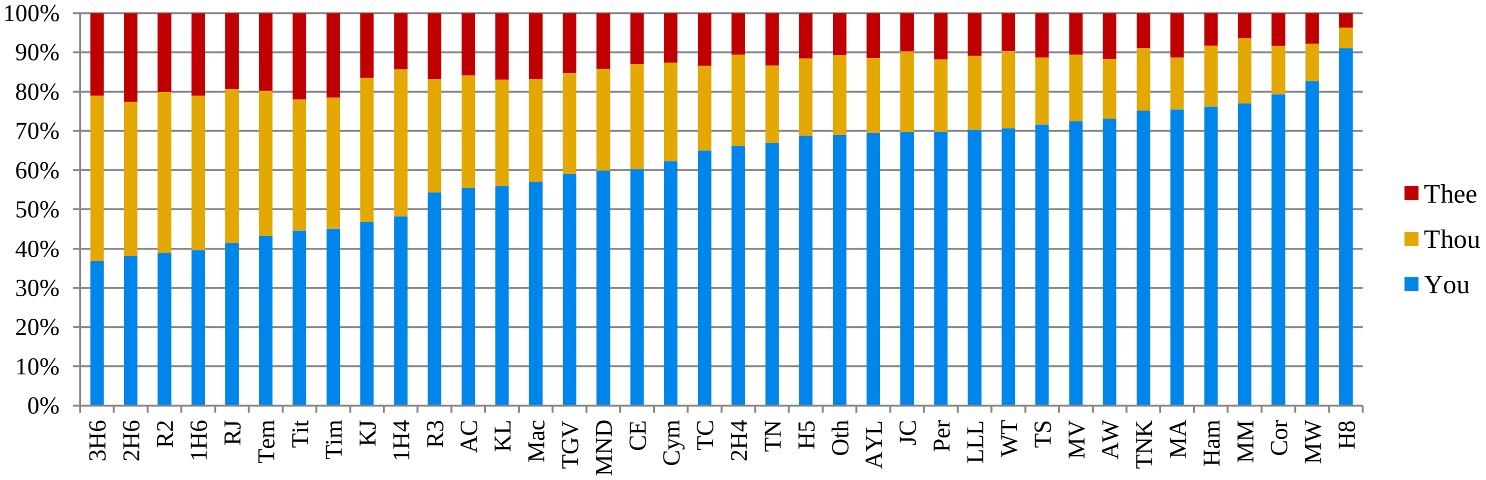

Figure 5: Relative pronoun distribution per play, ordered by percentage of YOU

A relative distribution of the pronouns in every play can be seen above in Figure 5, which shows clear differences between them. In *Henry VI, Part 3*, the most THOU pronouns occur in comparison to the amount of YOU pronouns in that play, while in *Henry VIII*, there are almost no THOU pronouns with an astonishing preference for YOU. A more precise list of the counts for every pronoun in each play can be found below in Table 4. The number of pronouns that were extracted from each play range from 363 (in *Macbeth*) to 811 (in *Coriolanus*).

Play	***You***	***Thou***	***Thee***	**Total**
1H4	327	254	97	678
1H6	177	176	94	447
2H4	483	170	77	730
2H6	197	203	117	517
3H6	191	218	109	518
AC	357	184	102	643
AW	481	100	77	658
AYL	520	143	86	749
CE	301	134	65	500
Cor	643	100	68	811
Cym	424	171	86	681
H5	370	106	62	538
H8	441	25	18	484
Ham	532	108	58	698
JC	393	116	55	564
KJ	224	175	79	478
KL	429	209	130	768
LLL	348	93	54	495
Mac	207	95	61	363
MA	495	87	74	656
MM	533	115	44	692
MND	273	118	65	456
MV	445	104	65	614
MW	552	64	52	668
Oth	490	145	76	711
Per	356	94	60	510
R2	160	169	83	412
R3	385	204	119	708
RJ	295	280	138	713
TC	445	147	92	684
Tem	216	185	99	500
TGV	328	143	85	556
Tim	284	210	135	629
Tit	231	173	114	518
TN	396	84	47	527
TNK	451	133	90	674
TS	515	123	81	719
WT	470	131	64	665
Total	14,365	5,489	3,078	22,932

Table 4: Complete distribution of pronouns over the plays in the dataset.

4.2 Overview of the results

What follows below is an overview of the results of the best models that were achieved for three different algorithms: Naive Bayes, decision trees and support vector machine. For this, the algorithms were implemented in Weka, using the default settings. For the evaluation of the models, a 10-fold cross-validation was used for all three algorithms. The baseline was also performed in this way, which in Weka is the ZeroR algorithm.

The three algorithms were run using two different approaches. The first one treated *thou* and *thee* as separate, thus keeping the separation in the THOU pronouns and creating a trinary classification problem. The second approach combined the THOU pronouns into one class to see if this would change the scores and the features in the final models. The second one is therefore a binary classification problem.

As mentioned in the previous chapter, I started with a model containing the full set of features and attempt to improve or maintain the scores for precision, recall, F-measure and accuracy while making the model simpler by excluding features. When there were conflicting changes in the scores, the scores of precision and F-measure were prioritised. Below in Table 5 is a short overview of the results of the first models of Naive Bayes and the decision tree algorithm. These are models created based on the full set of features, to see if through this first attempt we can already see an effect of the features on the pronoun prediction. The best performing model here is the binary Naive Bayes model, of which I have highlighted the F-measure score.

Out of the scores of the two algorithms shown below in Table 5, we can already see that the difference between the algorithms influences the prediction model. The Naive Bayes models, which assume that the features are all independent of one another, scores better than the decision tree models, which assume more of a dependence between the features. However, all four models shown here do perform better than the baseline. For this reason, I expanded the experiment to include a more complex algorithm, the support vector machine, which can handle more complex relations of features that seem to be present from these four initial models. Due to the good scores of the two algorithms shown here as well, I decided to compare all three to be able to interpret and compare the results from multiple perspectives and different assumptions.

Algorithm		**Precision**	**Recall**	**F-measure**	**Accuracy**
Baseline	Trinary 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	
	Binary 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	Trinary weighted avg.	0.820	0.806	0.812	80.6035%
	you	0.906	0.853	0.879	
	thou	0.794	0.821	0.807	
	thee	0.462	0.562	0.507	
	Binary weighted avg.	0.862	0.861	**0.861**	86.0893%
	YOU	0.895	0.882	0.888	
	THOU	0.806	0.826	0.816	
Decision tree	Trinary weighted avg.	0.726	0.749	0.710	74.9433%
	you	0.739	0.954	0.833	
	thou	0.833	0.578	0.699	
	thee	0.387	0.099	0.158	
	Binary weighted avg.	0.778	0.761	0.740	76.1425%
	YOU	0.742	0.949	0.833	
	THOU	0.838	0.448	0.740	

Table 5: Scores of Naive Bayes and decision tree models with the full set of features

In order to improve these models, I used feature ablation. I first started by taking out groups of features that were related to see what the effect would be rather than one feature at a time. With the 23 features, I created six different sets, which are listed here:

1. Play name, production date, genre, scene and location
2. N-gram
3. Speaker information: name, status, gender, age
4. Addressee information: name, status, gender, age, number of people
5. Status differential
6. Positive and negative sentiment

The first group portrays the wider linguistic and social context, while the second group is for the closer linguistic context. The information of the speaker and the addressee were separated based on the person, since the characters are the grouping factor of those features. I decided to keep status differential on its own, because it relates to multiple groups. Finally, the last group is for the positive and negative sentiment. Secondly, I went back over the features to see if individual exclusions would improve the model as well. This way, I ended up with the simplest and best model for every algorithm of which we can see the scores below in Table 6, Table 7 and Table 8.

4.2.1 Trinary classification scores

Shown below in Table 6 are the results of the trinary classification where *you*, *thou* and *thee* were all treated as separate categories. The scores of the best model of each algorithm for precision, recall, F-measure and accuracy are shown, as well as the overall weighted average for each algorithm. 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.

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	

Table 6: Scores for precision, recall, F-measure and accuracy for trinary pronoun prediction

4.2.2 Binary classification scores

Similarly to the section above, below in Table 7 are the results of the best models for the binary classification, thus where *thou* and *thee* are condensed into a single category THOU. Here, again, the three algorithms all perform better than the baseline, with even higher scores than for the trinary 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 shown in this thesis.

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	

Table 7: Scores for precision, recall, F-measure and accuracy for binary pronoun prediction

4.2.3 Confusion matrices

Below, in Table 8, are the confusion matrices of the best models. As can be seen here as well, the models from the three algorithms perform relatively well in comparison to the baseline, of which the confusion matrices are also shown in the table below.

Baseline	*you*	*thou*	*thee*		YOU	THOU	
	14,365	0	0	*you*	14,365	0	YOU
	5489	0	0	*thou*	8567	0	THOU
	3078	0	0	*thee*			
Naive Bayes	*you*	*thou*	*thee*		YOU	THOU	
	12,716	451	1198	*you*	13,210	1155	YOU
	510	4664	315	*thou*	1865	6702	THOU
	1232	275	1571	*thee*			
Decision Tree	*you*	*thou*	*thee*		YOU	THOU	
	13,797	240	328	*you*	12,397	1968	YOU
	2233	3152	104	*thou*	2197	6370	THOU
	2655	125	298	*thee*			
Support Vector Machine	*you*	*thou*	*thee*		YOU	THOU	
	13,317	292	756	*you*	13,132	1233	YOU
	755	4588	146	*thou*	1684	6883	THOU
	1226	110	1742	*thee*			

Table 8: Confusion matrices for all three algorithms, both in binary and trinary classification

4.3 Feature comparison of the models

Overall, the final models contain similar sets of features. The exact compositions can be seen below in Table 9. What is surprising, is that the binary classification model for the decision tree is very different from the other models as it does not contain any of the words from the n-gram as a predictor, whereas the others did. Instead, it contains some of the other features that were predictors in the other models, as well as some of the prior non-predicting features, such as gender, status and age. One of the attempts to get to the best model did contain RW1[13], which is the best predictor in all the other models, but this model scored significantly worse than the best model containing the features shown below.

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

Table 9: Features included in the best model of each algorithm[14]

4.4 Chapter summary

Overall, it is fair to say that the algorithms beat the baseline, with the best model being the support vector machine model for the binary classification. For both the binary and the trinary classification, support vector machine scored best, but additionally, all binary scores were higher than their trinary counterparts. It also became clear that not all features had the same effect on the pronoun choice, but in particular the features from the n-gram as well as the names of the speaker and addressee did have a positive effect on the accuracy of the choice. Only one of the best performing models stood out, which is the binary decision tree, which does not contain any of the n-gram features in its final model, whereas all other models do.

In the next chapter, we will go further into the comparison of the models, their scores and the features. We will also link back to the literature review to try and explain the outcomes of this study.

[13] Acronym used as laid out in Table 1.
[14] Acronyms used as laid out in Table 1.

5 Discussion

In this chapter we will continue discussing the results presented in the previous chapter. We will take a closer look at the results and relate them back to the literature to see if there is an explanation for them. At the end of the chapter, the results will be linked back to the hypotheses that were made for this study to see how the study performed. As in the previous chapter, the abbreviations used to identify the features in this chapter are specified in Table 1, in the Methodology.

5.1 Trinary versus binary classification

As can be seen in the previous chapter, both a binary and a trinary classification were run. The reason for this is that *thou* and *thee* are related in that they both serve as an alternative to *you*, but they occur in different positions. Whereas *thou* is the nominative case, *thee* functions as the accusative and dative form. The argument for keeping them separate is therefore mainly interesting from a linguistic perspective. Another reason for also performing the binary classification task, besides the fact that it is a more simplified classification task and is in line with the perspectives of previous studies, is to see whether the distinction between *thou* and *thee* creates a different model containing more or other features that predict the pronoun based on grammatical rather than pragmatic aspects.

When looking at the differences between the binary and trinary models, the first thing that stands out is that all the binary models' scores are higher than those of the trinary models, as can be seen when comparing Table 6 and Table 7. An easy explanation for this could be that, probabilistically, it is easier to predict a pronoun when there are only two options rather than three. Additionally, from a computational or statistic perspective, the higher scores, the better the model. However, there is more difference between the models than just their scores. In particular, the difference for the decision tree models is interesting. Though the decision tree models perform least well in comparison to the Naive Bayes and the support vector machine models, they do still perform much better than the baseline. When looking at the features in the best binary and trinary models of the decision tree algorithm, we see huge differences. While all other models include some n-gram features, the binary decision tree model does not. As mentioned above as well as in prior chapters, the main difference between the two THOU pronouns is their case, thus influencing the context in which they can occur. This is clearly visible in the binary decision tree model, where the n-gram, which represents the pronoun's linguistic context, does not show up as an important feature for the prediction. Instead, we have other features such as gender, status and positive sentiment come through. These are features that are partially also included in the other models, but always as secondary features where the most prominent features are the n-gram features.

When looking at the confusion matrices in Table 8, it is also visible that the algorithms are almost always mainly predicting the correct category, thus the pronoun that is indeed used by

Shakespeare in that instance, in both trinary and binary classification. However, when looking more closely at the wrongly predicted pronouns, it stands out that in the trinary classification, *thou* and *thee* are the least often confused with each other and more likely to be misclassified as *you*. This also shows that the models depend heavily on the context to predict the pronoun, which is the main difference between *thou* and *thee*.

So even though the highest scores were achieved in the binary classification, this might not be the best way of addressing the problem. As we see from a comparison of the models, this classification problem appears to be more nuanced due to the underlying extra-linguistic features that are poking through in all the models. Looking back at the previous studies done on pronoun interpretation and comparing them to the features in the models of this study, we can definitely conclude that there is the need for a balanced approach that does not assume that *thee* and *thou* are completely equal in their opposition to *you*. Additionally, the trinary models contain less or an equal number of features as the binary models, which makes them less complex in terms of input, even if the classification task itself is a more complicated one.

5.2 Comparison of the algorithms

When looking at the performance of the algorithms in comparison to the baseline of this study, we can confidently say that they outperform this baseline. Even the decision trees, which have the lowest score of the three algorithms, scores almost 13% (for trinary) and 19% (for binary) better than the baseline on accuracy, as can be seen in Table 6 and Table 7. The best performing algorithm is the support vector machine one, which has the highest accuracy scores for both binary and trinary classification. However, the Naive Bayes models are not far behind the support vector machine models, with only a few percent difference in the accuracy scores (support vector machine 85.7-87.3%;.Naive Bayes 82.6-86.8%).

As mentioned in the Methodology chapter, the models would be assessed based on multiple scores, not just the overall score for accuracy. In particular, the score for precision is important, as it shows how often the model predicted a pronoun to be YOU/THOU and it was right. For the trinary classification, the highest scores for precision were for *thou* in the decision tree and support vector machine models, but in the Naive Bayes model, the highest score was for *you*. In all three algorithms, the precision scores for *thee* were the lowest. In particular the decision tree and Naive Bayes model performed badly here, with the Naive Bayes only being just above 0.5, and the decision tree scoring 0.4. In comparison, the support vector machine scored 0.7 for *thee* for precision, which was still the lowest score for precision in that model.

This is different in the binary classification models, where the precision scores of the three algorithms were all relatively similar. Still, the highest scores were for the support vector machine model, closely followed by the Naive Bayes model, with the lowest scores for the decision tree model. The best scores were for YOU, this is the same for all three algorithms.

The binary classification precision scores are almost mimicked in the scores for recall, where the same order is kept for best to worst performing model, as well as the scores for YOU being higher than for THOU. This is for the recall scores of both the binary and the trinary classification. This means that in the trinary classification, the scores for recall and precision are different from one another. Specifically the recall score of YOU in the decision tree model is interesting, as it is the highest recall score out of all six models, even though the other scores are all among the lowest. Therefore, when looking at the F-measure for this model, this high recall score is completely levelled out by the precision score into what is again one of the least good scores overall.

5.3 Features

In this section we will go further into the feature evaluation and look at which features are included in the best models and which ones are not. There will also be a comparison to earlier studies as well as a more critical discussion on what it means to have these features included or excluded.

5.3.1 Included features

As mentioned previously, the features that are in the best models are not the same for each model. However, some of them do occur in (almost) all of them. An example is the n-gram, in particular LW1 and RW1. I already briefly discussed the role of linguistic context on the pronoun interpretation for the difference between *thou* and *thee*, but clearly there is some effect as well between YOU and THOU overall, as these features do still contribute as predictors in both the Naive Bayes and support vector machine binary models. They do not only come up as an important feature in these two models and the three trinary models, they are the most prominent features.

In particular RW1 shows through as containing valuable information about the pronoun. I ran an additional decision stump model, which chooses one value out of a single feature that appears to be the best predictor for the pronoun. This model came up with ‘RW1 = *art*’ as its final outcome. This means that the word *art* has a strong preference for one pronoun. In this case, the decision stump predicted that if RW1 = *art*, then the pronoun is THOU. If RW1 ≠ *art*, then the predicted pronoun is YOU. Several studies have looked into the role of verbs in relation to pronoun choice (Walker, 2007; Mulholland, 1987). These studies mainly looked at the role of different types of verbs, namely closed and lexical verbs. According to their studies, THOU is more likely to occur with closed-class verbs, of which *to be* is one (Walker, 2007; Busse, 2002). This does, however, not take into account the morphological inflection that is used in *art*. Mazzon (1995, in Busse, 2002) does note the use of verbal inflection in connection to THOU, where a null-inflection would occur with YOU. Similarly, Görlach (1999, in Busse, 2002) mentions that these verb endings were lost around the same time as the THOU pronouns, which does suggest some relation between them. In contrast, Stein (1974, in Stein, 2003 and Busse 2002) comes to the different conclusion that the combination of verb with a particular inflection is not related to the pronominal address choice, and is more likely to have a phonological reason. So even though there might not be a conclusive answer on the co-occurrence of verbs with specific morphological inflections and YOU/THOU pronouns yet, RW1 definitely has an important role in predicting the pronoun. It should be noted that the finding of *art* as a reliable predictor is not strange, as there are no cases in the data where *art* occurs as RW1 for YOU. In addition, the fact that in particular RW1 and LW1 are significant, also in comparison to RW2 and LW2, shows that perhaps the importance is found in verbs. Especially for the *thou*/*thee* distinction, this would explain why both RW1 and LW1 are good predictors, as in both positions a verb is likely to occur. One could also argue that the verbal inflections are used to accommodate to the pronoun and thus as an effect to the chosen pronoun. Thus, from a linguistic perspective, the presence and the importance of RW1 and LW1 in the models could be interpreted as a drawback of these models as RW1 and LW1 are features that are being influenced themselves by the pronoun. For this reason specifically, it is good to see that in the trinary classification models, the importance of LW1 and RW1 is equally present as in the binary models. This shows that the predictive value of RW1 and LW1 goes beyond the differences in inflections between YOU and THOU.

Another feature that is included in all models is S_ID or A_ID, which are the names of the speaker and the addressee. In both the Naive Bayes models, one of them occurs, as well as in the trinary decision tree model. In the other three models, which are both support vector machine models and the binary decision tree model, both S_ID and A_ID are found. This shows that the specific characters themselves were important, rather than just their characteristics such as gender, social status or age. Obviously, S_ID/A_ID do carry this specific information with them as well in that they do not differ from one utterance to the other, but it is not an individual characteristic that occurs in all models. An explanation for the positive influence on the pronoun prediction by S_ID and A_ID might be that certain characters act completely different than should be expected for their gender, social status or age. An example could be KL_Fool, the Fool in *King Lear*. This character repeatedly uses THOU when addressing Lear, even though the Fool is much lower in social status and younger than Lear. According to the literature, based on social status and age, the Fool is supposed to use YOU, as it is a sign of respect (Adamson, 2001; Quirk, 1974). However, this is clearly not the case, which is more in line with the studies of Stein (2003) and Calvo (1992). In particular in the case of the Fool, this might also relate to the comedic role of the Fool as a character to act differently from typical behaviour.

Additionally, status differential (Stat_Diff) seems to influence the pronoun choice as well. While this feature does not occur in either of the Naive Bayes models, it is found in the other four models. This is an interesting feature, as it looks at the differences between the speaker and the addressee, in terms of social status. However, from the literature, we know that between two higher class characters, YOU is normally used, and among lower-class characters, the more common pronoun is THOU (Stein, 2003). With a higher Stat_Diff, there is still no clear answer, as a higher class toward lower-class character would use THOU, but the other way around YOU would be used. This is not visible with Stat_Diff, or at least not by itself. A possible explanation for this could be that it is related to S_ID and A_ID, which means that the predictive strength of the combination of the status differential and the speaker/addressee names bypasses the influence of the status of the speaker or addressee by themselves.

Lastly, some of the other features that return in multiple models are Neg_Sent (negative sentiment) and Pos_Sent (positive sentiment). The importance of these two features is not strange when looking back at the literature, which numerously mentions the use of THOU/YOU as an emotive tool, or a choice driven by emotion (Stein, 2003; Calvo, 1992, Hickey, 2003; Walker, 2003). Perhaps more interesting is that the trinary decision tree model contains Neg_Sent, whereas the binary decision tree model contains Pos_Sent. This shows that these scores might be connected more closely than thought, and that any high or low score on emotive language use is sufficient in predicting the pronominal address term. Moreover, the difference between positive and negative emotion is thus not as important as simply the presence or lack of any type of emotion to differentiate between the possible pronouns.

The other features that occur in the models, but to a lesser degree, are related to the speaker and addressee, with the exception of Scene, in the binary decision tree model. As Scene is the only feature in the binary decision tree model that gives some contextual information, which is what the n-gram features did in the other five models, this could mean that Scene for that particular model functions as the contextual predictor. However, it is of course not a similar contextual feature, in that it does not convey any information about the linguistic context of the pronoun specifically. The other features, S_Gender (speaker's gender), A_Status (addressee's status), A_Age (addressee's age), S_Age (speaker's age) and A_Number (number of people addressed), occur only in a few of the models. Clearly they all do have some effect, but they are not as prominent in their influence as the other features that were just discussed. What is interesting about these is that while Age is present for both the speaker and the addressee as a predictor, this is not the case with S_Gender and A_Status. The absence of the speaker's status (S_Status) could be explained by the influence of Stat_Diff, which has shown an important influence. However, for Gender, there is no similar explanation for the lack of the addressee's gender (A_Gender) when S_Gender does occur in one of the models. A possible reason could be that the effect of gender is mainly affecting the speaker. As Walker (2003) explained, men appear to be more influenced by the use of the pronoun THOU, through which they express contempt or inclusiveness. Women, on the other hand, are not as influenced by the pronoun, which means that overall, men use THOU more than women. Since Walker explains the preference as expressed by the speaker, this might explain why S_Gender is included and A_Gender is not included in the models.

5.3.2 Excluded features

In contradiction to the previous section, there is of course also a number of features that have not shown any positive influence on predicting the pronoun. There were only six features that were excluded from the final models, most of which were from the first group as specified in the Results chapter (chapter 4). These features are Genre, Play, Prod_Date (production date) and Location. Including them in the models had a negative effect on the prediction scores, and were not part of any of the best models.

The first one, Genre, was relatively well-supported by the literature as having an effect on the prediction of the pronoun, as THOU was more likely to occur in Histories and least in Comedies (Busse, 2002). However, the genre distinction used in this study did not show up as a predictor in any of the models. There might still, of course, be an overall difference between the number of THOU pronouns between genres, but this difference is not prominent enough to show up in a prediction model like the ones in this study.

Similarly, the difference between plays did also not improve the model. While there were quite big differences between the plays, such as with *Henry VIII*, which contained almost no THOU pronouns, while *Henry VI, Part 3* had the most THOU pronouns in comparison to YOU pronouns. The absence of this feature in the models is therefore surprising, but might be explained by the majority of the plays in which the distribution is not as skewed as in the two I just mentioned. So on the larger scale of this study, the effect of this feature is lost in the size of the data and the more significant effect of other features.

Prod_Date did have some support from the literature, in particular Busse (2006). However, as already mentioned in previous chapters, the span of this study is a mere 22 years, which would make it difficult to find any effect. It is therefore not surprising that this feature does indeed not help predict the pronoun.

The last feature from group 1 that does not shown up in any of the models is Location. A reason for this might be that the influence of this feature is too subtle in its effect and therefore overshadowed by the other features. On the other hand, it might simply not influence the choice of pronoun. The idea behind this feature was that the difference between “private” and “public” might affect how open one character would be to another, thus perhaps altering the choice of pronoun. However, a likely reason why this did not show up would be that this is probably captured in the emotive language already, which did get included in the models through Pos_Sent and Neg_Sent. An example of where emotive language might have included information about the Location already, is between secret lovers. Towards each other, they are likely to use emotive language to express their love, but only in cases where they know no one will see or hear them, thus in a private location.

The final two features that were not included in any of the models are S_Status and A_Gender. I already shortly discussed them in the previous section, but then mainly in comparison to their speaker/addressee counterparts that were included. As individual features, there might be a slightly different reasoning. It might be that social status and gender might not be as influential as pronoun predictors as suggested by the literature (Adamson, 2001; Brown and Gilman, 1960; Walker, 2003). This would be particularly interesting when looking back at Brown and Gilman's 'power and solidarity' theory (1960): perhaps the influence of power that comes through social status and gender is not important. Over the years, the idea of 'power and solidarity' as a constant value has been rejected already, but the results of this study might suggest that the role of specifically power is not as strong as still assumed. Obviously, the other features that convey power through status and gender (A_Status, S_Gender, Stat_Diff) are included in the model, so I am not suggesting there is no effect of power, but it does show that the role is perhaps not as big as suggested by previous studies.

5.4 Hypotheses

We will now move on to an evaluation of the hypotheses that were presented in the Methodology chapter (chapter 3). They will each be discussed with the results of this study in mind, as well as link back to the literature again.

5.4.1 First hypothesis

The first hypothesis was as follows:

> 1. It will be too complex to build a model that fully predicts the pronominal address term solely based on linguistic and extra-linguistic features.

As a reminder, this is a null-hypothesis that we tried to falsify through this study. Since we do not know what a "good" model is exactly, the only comparison of our models is to the baseline. Based on that, we can definitely build a model to predict the pronominal address term based on linguistic and extra-linguistic features. All three algorithms have yielded better models than the baseline, with the smallest improvement being an approximately 13% higher accuracy score than the baseline. Our best accuracy score is from the support vector machine binary model, at approximately 87%. This is 24% better than the baseline of 63% accuracy.

If we are being picky, and look at the hypothesis again, does this model predict the pronoun "fully" and based "solely" on the features incorporated in this study? Then no, we have obviously not achieved a 100% accuracy score, nor were any of the other scores 100%. Additionally, as mentioned before, there are likely other features that might predict the pronoun as well that were not included in this study. One of the most often named features in the other studies that was not in this study, is familiarity (social distance). Unfortunately this feature had to be excluded because of the scope of this study, but therefore we cannot conclude that we have tried every feature that might influence the pronoun.

However, this is also not what we set out to do in this study. The aim was to create a model that performed better than a baseline that looked just at the distribution of the pronouns by adding different features to the model that help predict the pronoun. Through creating a ‘good’ model, we would know which features actually influence the pronoun choice. This goal has definitely been achieved through this study.

When looking at the literature, it was to be expected that we could predict the pronoun based on the set of features we started the experiment with. Almost all of them had results of other studies to back up their likely influence on the pronoun choice.

5.4.2 Second hypothesis

The second hypothesis consists of one main hypothesis and two sub-hypotheses as shown below:

> 2. Some features will be a better predictor of pronoun choice than other features.
> 2A. Social status, age and sentiment will show a greater influence on the pronoun choice.
> 2B. There will be no measurable influence of genre, production date or gender on the pronoun choice.

To start off with the main hypothesis: Yes, there are indeed some features that are better predictors of the pronoun choice than other features. We have a couple of features that have shown great influence on the pronoun choice, as they are in all models. Other features, however, did not positively affect the pronoun prediction and are thus not included in the final models.

While the main hypothesis was correct, the first sub-hypothesis was incorrect. While social status, age and sentiment are all included in the final models, they are not the best predictors and are not included in all six models. The absolute best feature, included as best predictor in five of the six models, was RW1, followed by LW1 and the other n-gram features. This shows that linguistic context is most important when predicting the feature. As mentioned earlier, this could be due to the occurrence of verbs with specific morphological inflections related to the pronouns in the RW1 and LW1 positions. This contradicts the literature, which assumed that social status, age and sentiment would be among the best predictors. From the results, we do still see an influence of these features, in particular of sentiment and age. Status is mainly represented in the status differential, which shows that it is mainly the difference in status or the lack thereof that is important when predicting the pronoun.

The second sub-hypothesis was partly correct. There was indeed no effect of Genre, even if some of the literature showed some differences in pronoun distribution between genres. It is likely that the difference overall is there, but this does not affect the pronoun prediction of a specific example noticeably. Secondly, the production date (Prod_Date) did also not have any effect, as expected. As explained previously, there is likely to be a visible effect over time in a study with a broader time span, but since this study only spanded 22 years, the overall loss of THOU during the EModE period is not noticeable. Finally, I was not entirely correct with my prediction on Gender, as there did happen to be an effect on the pronoun prediction. The speaker's gender (S_Gender) showed up as significant in just one of the six models, but this does show that it is not one of the least effective predictors. Some of the features that performed worse as predictors than Gender were Location and Play.

To link back to the literature, it was to be expected that Gender would actually show an effect, as there are a few studies that support the role of gender (Walker, 2007; Mazzon, 2003). Contrariwise, there is no immediate effect found in previous studies for Location and Play. The reason for this is, though, that there has been no direct research on the effect of these two features.

5.4.3 Third hypothesis

The third hypothesis of the study is as follows:

> 3. The algorithm to make the most accurate prediction and show the influence of each linguistic feature best will be the support vector machine.

As I hypothesised, it was indeed the support vector machine that performed best out of the three algorithms. It scored highest on the accuracy and precision score, but also the F-measure and recall score were overall the highest of the three. However, it was not the only well-performing algorithm. Whereas the decision tree models were not nearly as good, the Naive Bayes models scored surprisingly similar to the support vector machine models. The prediction was that the support vector machine models would be the best due to the complexity of the algorithm and its robustness. In comparison to Naive Bayes, which is a relatively simple model that assumes all features are independent, it is actually the Naive Bayes that scored surprisingly well. It also shows that most of the features in the models are indeed independent of one another.

The biggest difference between the Naive Bayes models and the support vector machine models is not in their scores, but in their features. Firstly, the Naive Bayes models only contain five (trinary classification model) and seven (binary classification model) features, while both the support vector machine models contain ten features. This makes the support vector machine models a lot more complex for only a slightly better score. For that reason, one could prefer the Naive Bayes models over the support vector machine models. However, the Naive Bayes models solely rely on the n-gram features, S_ID and A_ID. They do not show any of the smaller nuances of all the other features that did show up as influential in all the other four models. It is for that reason that I do believe that the complexity of a support vector machine trumps the Naive Bayes models, as it provides more linguistic information on top of its high scores.

5.5 Chapter summary

To summarise, this study has given some new insights into the analysis of pronominal address terms. Looking at the pronoun choice can be as both a binary and a trinary classification problem, which will result in somewhat different results. From the model comparison, we have seen, in particular for the decision tree models, that the context is most important when predicting the pronoun. This is evidence of the difference between *thou* and *thee*, and supports the idea of keeping them separate instead of condensing them into one THOU category. Therefore, linguistically, it makes more sense to perform a trinary classification, whereas a binary classification gives a computationally better result.

The differences between the algorithms were visible, but all three algorithms outperformed the baseline. The best models were from the support vector machine algorithm, but the scores of the Naive Bayes models were quite similar. A choice for either 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 best predicting features were from the n-gram, which supports the importance of the direct linguistic context of the pronoun. In particular RW1 showed up as the most important feature in predicting the pronoun. Some of the other important features were the speaker's name (S_ID), addressee's name (A_ID), status differential (Stat_Diff), positive sentiment (Pos_Sent) and negative sentiment (Neg_Sent), with additional support from the speaker's gender (S_Gender), addressee's status (A_Status), addressee's age (A_Age), speaker's age (S_Age) and number of people addressed (A_Number). Only six features were not included in any of the models: Genre, Play, production date (Prod_Date), Location, speaker's status (S_Status) and addressee's gender (A_Gender).

When looking at the hypotheses, we were able to falsify the null-hypothesis that it is not possible to build a prediction model based on linguistic and extra-linguistic features. The second hypothesis, about the features, was partially correct in which features would be good predictors and which features would not. The final hypothesis, about the best algorithm for this prediction problem, was correct as well.

The next and final chapter will conclude this thesis, summarise the full study and its results., and will also give some suggestions for future research.

6 Conclusion

In this chapter we will summarise the thesis as a whole, and mention some of its limitations as well. The chapter will finish off with some suggestions for future research in pronoun prediction and interpretation.

6.1 Summary of the thesis

To sum up, there has been a lot of research done on Shakespeare's use of pronominal address terms and the nuances between these pronouns, but they did not cover all plays and all aspects to see which features do truly affect the pronoun choice overall. As suggested in the literature (Busse, 2006; Stein, 2003), the study of pronoun use should be a combination of multiple fields and perspectives as they all work together in this decision. Therefore this study looked at all 38 plays written by Shakespeare, and used the perspectives and applications from linguistic, literary and computational fields to attempt to create a pronoun prediction model based on linguistic and extra-linguistic features. The use of all of Shakespeare's plays in a single study and the use of a data-driven perspective in the analysis of this topic are what make this study stand out from previous research on pronoun use.

Through this study, we found out that we can indeed build such a prediction model that is based on several linguistic and extra-linguistic features. However, the features in the models of this study do not exactly align with features mentioned as important in previous studies. This study revealed that, in particular, the direct linguistic context of the pronoun is important, with RW1 being the best predicting feature of all. Other important features include the speaker's and addressee's names, status differential and both positive and negative sentiment[15]. These all showed up in most of the models, and were supported by several studies, such as Mazzon (2003) and Calvo (1992).

There was also a difference between the three algorithms that were used in this study, with support vector machine scoring best with 87.3% accuracy with its binary classification model. Naive Bayes performed almost as well as support vector machine, with simpler models, but did not show as many influences of the different features and almost completely relied on the n-gram features alone. Some of the features that did not show any positive influence on the pronoun predicting were genre, play name, production date and location of the scene[16]. The lack of influence from genre directly contradicts Walker (2003), who did find a difference in pronoun use between genres. Also the speaker's status and addressee's gender (S_Status and A_Gender) did not have any influence, while addressee's status and speaker's gender did show up as predictors in the models. Both status and gender were also supported as influencing factors by Walker (2003), of which just a small effect was found.

[15] In the same order: S_ID and A_ID, Stat_Diff, Pos_Sent and Neg_Sent.

[16] In the same order: Genre, Play, Prod_Date and Location.

Overall, it is fair to say that, based on this study, the pronoun choice is nuanced and influenced by several features, though some surprising differences with previous literature were found. The main important finding of this study is that the pronoun choice is most dependent on the linguistic context it is used in.

6.2 Limitations of the study

Unfortunately, there were some limitations to this study that should be mentioned. First of all, the exhaustive annotation of all possible features was not feasible for a study of this scale. The main example of this is familiarity (social distance), which is a specific feature that is mentioned several times in the literature, but due to its complexity, would have been hard to annotate reliably in this study on top of some of the other more complicated features.

Secondly, we used automatic tagging for the addressee. While this method was based on the theory of the last-as-next bias, it might not be entirely accurate. Since we did find some positive influences from the addressee features, we can say that overall it worked, but it could be improved with manual checking to see if it indeed tagged the correct addressee. Since this did not only affect one of the features but the whole group of addressee features (ID, Status, Gender, Age, Number), it might have altered the outcome of the models slightly in cases of incorrect addressee tags.

6.3 Future research

The results of this study create opportunities for several future research topics within the different fields connected to pronoun interpretation. An easy one to see is the implementation of other algorithms to see if there is one that can improve the score achieved by the support vector machine models. Some interesting options could be a maximum entropy model, such as logistic regression, or a neural network. Another option is to add additional features to the dataset. I already mentioned familiarity as a well-supported addition, but also the role of spelling and morphology might be interesting to gain more linguistic information directly surrounding the pronoun beyond the words themselves. Similarly, the n-gram could be changed to contain part-of-speech tags as well, since I suggested that the significance of RW1 and LW1 might be related to their likelihood to be a verb.

Another way to expand would be to do a more comparative study and compare the results of this study to the pronoun usage of other writers. This could be both from Shakespeare's contemporaries or from writers before and after his time. Another possible comparison could be between languages, which might reveal some information about the loss of THOU in English in comparison to the T/V distinction kept in other languages.

Lastly, the study revealed some differences between binary and the trinary classification, and concluded that we cannot assume that *thou* and *thee* are completely equal in their opposition to *you*.

Therefore, it might be interesting to run another binary classification on *thou* and *thee*, to see if this highlights additional information about the differences between the two pronouns.

7 Bibliography

Adamson, Sylvia. (2001). "Understanding Shakespeare's grammar: Studies in small words", in Sylvia Adamson (ed.), *Reading Shakespeare's Dramatic Language: A Guide*, pp. 210-236. London: Bloomsbury Arden Shakespeare.

Barber, Charles. (1987). "'You' and 'thou' in Shakespeare's *Richard III*", in Vivian Salmon and Edwina Burness (eds.), *A reader in the language of Shakespearean drama*, pp 163-180. Amsterdam/Philadelphia: John Benjamins.

Bate, Jonathan & Rasmussen, Eric (eds.). (2007). *William Shakespeare: Complete works*. London: The Royal Shakespeare Company.

Baugh, Albert C. & Cable, Thomas. (2013). *A history of the English language*, 6th ed. London: Routledge.

Brown, Roger W. & Gilman, Albert. (1960). "The pronouns of power and solidarity", in T.A. Sebeok (ed.), *Style in language*, pp. 253-276. Cambridge: MIT Press.

Burnley, David. (2003). "The T/V pronouns in later Middle English literature", in Irma Taavitsainen and Andreas H. Jucker (eds.), *Diachronic perspectives on address term systems* [Pragmatics & Beyond New Series 107], pp. 27-45. Amsterdam/Philadelphia: John Benjamins.

Busse, Beatrix. (2006). *Vocative constructions in the language of Shakespeare* [Pragmatics & Beyond 150]. Amsterdam/Philadelphia: 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 Irma Taavitsainen and Andreas H. Jucker (eds.), *Diachronic perspectives on address term systems* [Pragmatics & Beyond New Series 107], pp. 193-221. Amsterdam/Philadephia: 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* [Pragmatics & Beyond New Series 106]. Amsterdam/Philadelphia: John Benjamins.

Calvo, Clara. (1992). "Pronouns of address and social negotiation in As You Like It", in *Language and Literature*, Vol. 1(1), pp. 5-27. London: Longman Group UK Ltd.

Chen, Edwin. (2011). "Choosing a Machine Learning Classifier". (30-05-2017). Retrieved from http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/.

Culpeper, Jonathan. (in preparation). "Introduction to the Encyclopaedia of Shakespeare's Language", in *Encyclopaedia of Shakespeare's Language*, pp. 1-14.

Culpeper, Jonathan, Findlay, Alison, Cortese, Beth & Thelwall, Mike. (under review). "'What means this passionate discourse?' (Henry VI, Part II, 1.1.101): Measuring emotional temperatures in Shakespeare's drama", *English Text Construction*. John Benjamins.

Drew, Paul & Curl, Traci. (2008). “Conversation analysis: Overview and new directions”, in Vijay K. Bhatia, John Flowerdew and Rodney H. Jones (eds.), *Advances in discourse studies*, pp. 22 -35. London/New York: Routledge.

Gillings, Mathew. (2016). A *topic modelling analysis of the Lancaster Corpus on Climate Change (British) and a methodological evaluation of topic model interpretation* (Master’s dissertation). Retrieved from Lancaster University (code 66592895).

Greenblatt, Stephen, Cohen, Walter, Howard, Jean E. & Maus, Katharine E. (1997). *The Norton Shakespeare: Based on the Oxford edition*. New York: W.W. Norton & Company, Inc.

Hardie, Andrew. (2017, July). “Exploratory analysis of word frequencies across corpus texts: Towards a critical contrast of approaches”, presented at the 9th International Corpus Linguistics Conference, University of Birmingham. 25 July 2017. Retrieved from: https://www.youtube.com/watch?v=ka4yDJLtSSc.

Hardie, Andrew. (2012). “CQPweb – combining power, flexibility and usability in a corpus analysis tool”. *International Journal of Corpus Linguistics* 17(3), pp. 380-409.

Hickey, Raymond. (2003). “Rectifying a standard deficiency: Second-person pronominal distinctions in varieties of English”, in Irma Taavitsainen and Andreas H. Jucker (eds.), *Diachronic perspectives on address term systems* [Pragmatics & Beyond New Series 107], pp. 343-374. Amsterdam/Philadelphia: John Benjamins.

Jones, Steven E. (2016). “The emergence of Digital Humanities (as the network is everting)”, in Matthew K. Gold and Lauren F. Klein (eds.), *Debates in the Digital Humanities 2016*, pp. 3-15. Minneapolis/London: University of Minnesota Press.

Jurafsky, Daniel & Martin, James H. (2009). *Speech and language processing*. 2nd Ed. New Jersey: Pearson Education Inc.

Kielkiewicz-Janowiak, Agnieszka. (1994). “Sociolinguistics and the computer: Pronominal address in Shakespeare”, in *Studia Anglica Posnaniensia XXIX*, pp. 49-56.

Klein, Lauren F. & Gold, Matthew K. (2016). “Introduction”, in Matthew K. Gold and Lauren F. Klein (eds.), *Debates in the Digital Humanities 2016*, pp. ix-xv. Minneapolis/London: University of Minnesota Press.

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 Irma Taavitsainen and Andreas H. Jucker (eds.), *Diachronic perspectives on address term systems* [Pragmatics & Beyond New Series 107], pp. 223-249. Amsterdam/Philadelphia: John Benjamins.

Mulholland, Joan. (1987). “‘Thou’ and ‘you’ in Shakespeare: A study in the second person pronoun”, in Vivian Salmon and Edwina Burness (eds.), *A reader in the language of Shakespearean drama*, pp. 153-162. Amsterdam/Philadelphia: John Benjamins.

Ng, Andrew. (n.d.). *Machine Learning* [Free online course on Coursera.org]. Stanford University.

Posner, Miriam. (2016). “What’s next: The radical, unrealized potential of digital humanities”, in Matthew K. Gold and Lauren F. Klein (eds.), *Debates in the Digital Humanities 2016,* pp. 32-41. Minneapolis/London: University of Minnesota Press.

Quennell, Peter & Johnson, Hamish. (2002). *Who’s who in Shakespeare*. London/New York: Routledge.

Quirk, Randolph. (1974). “Shakespeare and the English language”, in Randolph Quirk, *The linguist and the English language,* pp. 46-64. London: Edward Arnold.

Robertson, Stephen. (2016). “The differences between Digital Humanities and Digital History”, in Matthew K. Gold and Lauren F. Klein (eds.), *Debates in the Digital Humanities 2016,* pp. 289-307. Minneapolis/London: University of Minnesota Press.

Salmon, Vivian. (1967). “Elizabethan colloquial English in the Falstaff plays”, in *Leeds studies in English and kindred languages,* pp. 37-70. Leeds: Leeds University.

Stein, Dieter. (2003). “Pronomial usage in Shakespeare: Between sociolinguistics and conversation analysis”, in Irma Taavitsainen and Andreas H. Jucker (eds), *Diachronic perspectives on address term systems* [Pragmatics & Beyond New Series 107], pp. 251-307. Amsterdam/Philadelphia: John Benjamins.

Taavitsainen, Irma & Jucker, Andreas H. (2003). “Introduction”, in Irma Taavitsainen and Andreas H. Jucker (eds.) , *Diachronic perspectives on address term systems* [Pragmatics & Beyond New Series 107], pp. 1-25. Amsterdam/Philadephia: John Benjamins.

Thelwall, Mike, Buckley, Kevan, Paltoglou, Georgious, Cai, Di & Kappas, Arvid. (2010). “Sentiment strength detection in short informal text”. *Journal of the American Society for Information Science and Technology,* 61(12), pp. 2544-2558.

Wales, Katie. (2001). “Varieties and variation”, in Sylvia Adamson (ed.), *Reading Shakespeare’s Dramatic Language: A Guide,* pp. 192-209. London: Bloomsbury Arden Shakespeare.

Walker, Terry. (2003). “*You* and *thou* in Early Modern English dialogues: Patterns of usage”, in Irma Taavitsainen and Andreas H. Jucker (eds.), *Diachronic perspectives on address term systems* [Pragmatics & Beyond New Series 107], pp 309-342. Amsterdam/Philadelphia: John Benjamins.

Walker, Terry. (2007). *Thou and you in Early Modern English dialogues: Trials, depositions, and drama comedy* [Pragmatics & Beyond New Series 158]. Amsterdam/Philadelphia: John Benjamins.

8 Appendix A

List of plays and abbreviations

Number	Full name	Abbreviation
1	Henry IV, Part I	1H4
2	Henry VI, Part I	1H6
3	Henry IV, Part II	2H4
4	Henry VI, Part II	2H6
5	Henry VI, Part III	3H6
6	Anthony and Cleopatra	AC
7	All’s Well That Ends Well	AW
8	As You Like It	AYL
9	The Comedy of Errors	CE
10	Coriolanus	Cor
11	Cymbeline	Cym
12	Henry V	H5
13	Henry VIII	H8
14	Hamlet	Ham
15	Julius Caesar	JC
16	King John	KJ
17	King Lear	KL
18	Love’s Labour’s Lost	LLL
19	Macbeth	Mac
20	Much Ado About Nothing	MA
21	Measure For Measure	MM
22	A Midsummer Night’s Dream	MND
23	The Merchant of Venice	MV
24	The Merry Wives of Windsor	MW
25	Othello	Oth
26	Pericles: Prince of Tyre	Per
27	Richard II	R2
28	Richard III	R3
29	Romeo and Juliet	RJ
30	Troilus and Cressida	TC
31	The Tempest	Tem
32	The Two Gentlemen of Verona	TGV
33	Timon of Athens	Tim
34	Titus Andronicus	Tit
35	Twelfth Night	TN
36	The Two Noble Kinsmen	TNK
37	The Taming of the Shrew	TS
38	The Winter’s Tale	WT

Table A1: List of plays in the Shakespeare Corpus