Literary and Linguistic Computing Advance Access published November 10, 2011

A Naïve Bayes classifier for Shakespeare's second-person pronoun

Kyle Mahowald
Department of Brain and Cognitive Sciences, Cambridge MA 02139, USA

Abstract

In order to investigate in explicit detail the way that *y*- and *th*- pronouns alternate in the Shakespearean corpus, I have undertaken a collocational analysis of the full corpus of Shakespeare's 37 plays and found that (1) second-person pronouns can be disambiguated based on context alone, (2) *y*- pronouns seem to be used in more formal situations or when an inferior is addressing a social better, and (3) the *th*- pronoun is reserved for addressing peers, servants, or other familiar personages. Through the Python Natural Language Toolkit (Bird *et al.*, 2009, *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media), I implemented a Naïve Bayes classifier that in effect treats each occurrence of a second-person pronoun as a black box that must be resolved into either a *y*- pronoun or a *th*- pronoun based only on the surrounding words. Using tenfold cross-validation, the classifier achieves an accuracy of 78.3% when fellow *th*- and *y*- pronouns are excluded from the context and 88.0% when we allow fellow *th*- and *y*- pronouns to assist in classification. Most interesting, however, are the context words that prove most informative in categorizing the pronouns. Significantly, the words most useful in classifying a pronoun as a *y*- pronoun include high-register words such as *lordship*, *madam*, *lords*, and *sir*. After a group of conjugated second-person verbs like *art* and *wert*, the words most associated with *th*- pronouns are words such as *torment*, *nuncle*, *lesser*, and *villain*. The ability to discriminate between forms based only on context confirms the hypothesis that the two classes of second-person pronoun are indeed used distinctly in the Shakespearean corpus. The list of words most helpful in making that distinction strongly suggests a difference in formality. We can also gain additional insight into the plays by examining some of the unexpected words that collocate with either one form or the other.

Correspondence:
Department of Brain and Cognitive Sciences, Building 46, MIT, 43 Vassar Street, Cambridge MA 02139, USA.
E-mail:
kylemaho@mit.edu

1 Introduction

Unlike Present Day English (PDE), Old English maintained a strict distinction between the second-person singular pronoun and the second-person plural pronoun (Mitchell, 1985). Old English had nominative *þu*, accusative/dative *þe*, and genitive *þin* for the singular alongside the plural *ge*, *eow*, and *eower*. By Shakespeare's time, perhaps following languages such as French and Spanish, the number distinction had given way to a formality distinction among second-person

Literary and Linguistic Computing © The Author 2011. Published by Oxford University Press on behalf of ALLC.
All rights reserved. For Permissions, please email: journals.permissions@oup.com
doi:10.1093/llc/fqr045

pronouns, whereby the once singular *th*- pronouns are used informally while the plural y-pronouns are used formally. OE nominative *þu*, accusative/dative *þe*, and genitive *þin* had developed into *thou*, *thee*, and *thine*. *Thy* was also available in the genitive and typically preceded a consonant-initial word, whereas *thine* was used before a vowel. OE *ge*, *eow*, and *eower* gave way to *you* (sometimes *ye* from nominative *ge*), *your*, and *yours*.

Hope (2003) cites a number of examples showing the formality distinction. For instance, consider Brutus's words from *Julius Caesar*: 'Farewell to you, and you, and you Volumnius,/ Strato thou hast bin all this while asleep:/ Farewell to *thee*, to Strato' (5.5.31-3). He uses *you* for his friends, like Volumnius. But he switches to the more familiar *thou* and *thee* for addressing his servant Strato. This is one of many examples in Shakespeare of this type of register shift. It is not unlike the distinction in French between *tu* and *vous*.

While the early modern distinction between the *th*- and *y*- pronouns has often been discussed as in Barber (1976), Wales (1983), Busse (2002), and Walker (2007) among others, I do not believe that it has ever been adequately investigated computationally. In order to investigate in explicit detail the way that *y*- and *th*- pronouns alternate in the Shakespearean corpus, I have undertaken a collocational analysis of the full corpus of Shakespeare's 37 plays and found that (1) second-person pronouns can be disambiguated based on context alone, (2) *y*- pronouns seem to be used in more formal situations or when an inferior is addressing a social better, and (3) the *th*- pronoun is reserved for addressing peers, servants, or other familiar personages. Using the Python Natural Language Toolkit (Bird *et al*., 2009), I implemented a Naïve Bayes Classifier that in effect treats each occurrence of a second-person pronoun as a black box that must be resolved into either a *y*- pronoun or a *th*- pronoun based only on the surrounding context words. Using tenfold cross-validation, the classifier achieves an accuracy of 78.3% when fellow *th*- and *y*- pronouns are excluded from the context and 88.0% when we allow fellow *th*- and *y*- pronouns to assist in classification.

Most interesting for my purposes, however, are the context words that prove most informative in categorizing the pronoun. Significantly, the words most useful in classifying a pronoun as a *y*- pronoun include high-register words like *lordship*, *madam*, *lords*, and *sir*. After a group of conjugated second-person verbs like *art* and *wert*, the words most associated with *th*- pronouns are words like *torment*, *nuncle*, *lesser*, and *villain*. The ability to discriminate between forms based only on context confirms the hypothesis that the two classes of second-person pronoun are indeed used distinctly in the Shakespearean corpus. The list of words most helpful in making that distinction strongly suggests a difference in formality. We can also gain additional insight into the plays by examining some of the unexpected words that collocate with either one form or the other.

2 Methodology

The corpus used is Jon Bosak's 1999 XML markup Shakespeare 2.00 (Bosak 1999), which consists of 37 plays. The classifier is a Naïve Bayes Classifier, as described in Bird *et al*. (2009) and Jurafsky and Martin (2008), and implemented in Python using the Natural Language Toolkit.[1] In effect, we seek to maximize the probability of the pronoun class given the data available. We will refer to the hypothesized pronoun class (*th*- or *y*-) as our hypothesis and the context words as our data. We want to maximize $P(H|D)$ where H is *hypothesis* and D is *data*. Bayes' law allows us to estimate this by

$$\frac{P(D|H)*P(H)}{P(D)} \qquad (1)$$

Because the $P(D)$ does not vary within a given instance, it acts as a constant and lets us assume that

$$P(H/D)\,\alpha\,P(H/D)*P(H) \qquad (2)$$

$P(H)$ is referred to as the prior and, for us, is the prior probability that the given pronoun is a *th*- pronoun or a *y*- pronoun. This depends on the total proportion of *th*- pronouns to *y*- pronouns in the corpus. In the extreme case where every possible second-person pronoun was a *th*- pronoun, the prior that $P(H=th)$ would be 1. $P(D|H)$ is the likelihood of the context given that H is either *y*- or *th*-. For instance, if every time a *y*- pronoun appeared in

Table 1 Pronouns used by classifier

y- pronouns	th- pronouns
you, your, yours, yourself, yourselves, yours, ye	thou, thee, thy, thine, thyself, thyselves

the test set, the same set of words, *x*, appeared in the context, the likelihood $P(D=x|H=\mathrm{y})$ would be 1. Of course, this is quite unrealistic as the likelihoods in question will almost always be very small. The classifier is learned by training on the data to obtain a set of features that determine which context words are most strongly collocated with *th-* or with *y-*.

The classifier seeks to maximize the posterior distribution, which is $P(H|D)$ by maximizing the product of the likelihood and the prior. Based on that maximization, it makes a binary choice to assign either a th-pronoun or a *y-* pronoun to the test case. The use of the prior makes the classifier robust to overfitting to context words that occur near target words in the training set just by coincidence. Furthermore, only context words that occurred more than 25 times in the corpus were considered. Experimentation with different parameters also proved that the best number of context words to take into account on each side of the target words was 8.

In one version of the classifier, all words were available as context words, including neighboring fellow *th-* and *y-* pronouns. This version achieved an accuracy of 88.0%, in large part, because a *th-* pronoun in a given sentence is a very good indicator that another second-person pronoun in that sentence will also be a *th-* pronoun. The same goes for *y-* pronouns. A classifier that includes these features is interesting insofar as it definitively shows that, in a given context, a pronoun can be disambiguated quite accurately. That said, it tells us less about things like formality since it relies so much on just the other pronouns in the neighborhood.

It could also be argued that if we are going to exclude neighboring *th-* and *y-* pronouns from the context, we should also exclude conjugated second-person verbs since they tell us nothing about issues like register or social status. That, however, leads to a slippery slope: one could then claim that we should exclude plural words from the *y-* category since *y-* pronouns were the only ones available for the plural. We could also exclude words often used with *your*, as in the phrase *your lordship*, since those sometimes seem fossilized. Were we to exclude all of these factors, we would be undermining our fundamental goal: discovering if and how second-person pronouns can be automatically disambiguated based on context. That said, a table that simply shows how *thyself* is more likely to be seen near *thy* than *your* is simply not that interesting. Thus, I have limited myself to two versions of the classifier: the one that includes a full context and the one that excludes neighboring fellow *th-* and *y-* pronouns.

The set of pronouns used as *th-* pronouns and *y-* pronouns are listed in Table 1. Capitalization was ignored for both target word and context words. To avoid using context words across different speakers, only context words within the same utterance as the target word were considered viable. For instance, if a target pronoun occurred as the first word of a speaker's utterance, only the eight words after it were considered as context. Similarly, if a target word occurred in an utterance-final position, only the preceding eight words were considered.

The order of utterances was scrambled in order to avoid any unwanted biases. A version was tried in which the speaker of the utterance was taken into account as a feature. But that caused a decrease in accuracy, likely due to overfitting as a result of the relatively large number of speakers compared to the total number of utterances.

As an example, take the following utterance from *Julius Caesar*: 'Truly sir in respect **of a fine workman, I am but, as** you **would say, a Cobler'** (1.1.7–12). The target word is the underlined *you*, and the context words are the words in bold. Each context word is associated with likelihoods $P(D|H=y)$ and $P(D|H=th)$ that says how likely it is to occur given that it is near a *y-* or given that it is near a *th-*. These likelihoods are determined during training by tabulating frequencies for each possible context word in each environment. Of course, since we do not know which pronoun the target word is during testing, both possibilities must be considered. The probabilities are multiplied together to find an overall likelihood for the context given either a *th-* pronoun or a *y-* pronoun. The maximum posterior, which allows us to make a decision on the test case, is then found by multiplying the likelihood by the prior.

In order to account for previously unseen context words, the NLTK default smoothing was used. NLTK uses the ELEProbDist package to calculate the expected likelihood estimate for each occurrence as follows: $(c + 0.5)/(N + B/2)$, where c is the count, N is the number of outcomes, and B is the number of bins. In effect, we can think of this as '+0.5 smoothing' where 0.5 is added to each count. The frequencies are then renormalized to ensure that the probabilities add up to 1. I did not find any other smoothing methods that significantly improved the model.

One of the strengths of this methodology is that the classifier itself does not rely on any annotated judgments from a modern English speaker to make formality judgements about earlier English. That said, one reviewer of this article astutely noted that the methodology used here would ideally be used not for testing an existing hypothesis but for making a new hypothesis. This reviewer suggests that in order to test a hypothesis like 'y- pronouns are used formally and th- pronouns informally', one would need a human to annotate a set of contexts as either 'formal' or 'informal' and then test whether there is a significant correlation. The problem with this approach is one that plagues many historical English studies: modern speakers of the language cannot be trusted to have accurate intuitions about historical English.

Thus, what the present classifier does, in effect, is not test a hypothesis but (i) show that context alone can indeed be used to automatically disambiguate the type of pronoun used and (ii) produce a list of words that are indicative of *th-* or *y-* pronouns. While the subsequent discussion of that list of words will rely on my own coarse judgments, it is crucially important that the classifier itself remain 'pure' of such modern English annotation, which is best reserved for the discussion of the results and not the results themselves.

3 Results

Tenfold cross-validation was used to train and test the classifier. That is, each 10% chunk of the corpus was given a turn as the test set, while the other 90% was used for training. This was repeated for each 10% chunk, and the results were then averaged to determine the overall accuracy, which was 88.0% for the full-context version (inclusive version) and 78.3% for the version in which neighboring *th-* and y- pronouns were excluded from the context (exclusive version).

Overall, in each classifier, there were 21,048 *y-* pronouns and 12,192 *th-* pronouns classified across all tenfolds. In the inclusive version, 92.8% of *y-* pronouns were correctly classified, compared to 82.1% of *th-* pronouns. Of the errors made, 54.7% misclassified a *th-* as a *y-*, whereas the other 45.3% did the opposite.

In the exclusive version, 84.6% of *y-* pronouns were correctly classified, compared to 67.4% of *th-* pronouns. 55.1% of the errors involved misclassifying a *th-* pronoun as a *y-* pronoun, whereas 44.9% misclassified a *y-* pronoun as *th-*. Both versions of the classifier performed significantly better (paired *t*-test, $P < 0.0001$) than a trivial majority vote classifier, which would achieve 63.3% accuracy by marking all instances as *y-*. See Tables 2 and 3 for the full results.

Table 2 Results for inclusive version of the classifier

	Actually a *y-* form	**Actually a *th-* form**
Classified as a *y-* form	19,236	2,183
Classifed as a *th-* form	1,808	10,013

Table 3 Results for exclusive version of the classifier

	Actually a *y-* form	**Actually a *th-* form**
Classified as a *y-* form	17,808	3,980
Classifed as a *th-* form	3,240	8,212

Table 4 Most informative features by pronoun type

th- words		*y*- words	
shalt	229.5	lordship	49.6
didst	137.3	voices	18
hast	137.1	ladyship	12.2
wert	132.9	worship	11.9
wilt	111.5	madam	11.8
canst	102.6	intent	10.2
wouldst	101.3	sir	9.6
dost	96.4	slaves	9.1
know'st	74.2	provided	9.1
seest	69.6	humbly	8.4
wast	64.8	private	7.4
think'st	58.1	please	7.2
couldst	57	beholding	7.1
speak'st	55.8	breach	6.8
art	51.4	gentlemen	6.7
sayest	51.2	wives	6.7
liest	50.7	received	6.7
shouldst	49.3	beseech	6.6
lovest	23.4	heartily	6.4
fish	15.5	needful	6.4
thunder	14.4	acquainted	6.2
cursed	12.6	masters	6.2
snow	12.1	troubled	6
prithee	11.8	notice	6
hateful	10.6	pleasures	5.7
fiend	10.5	meeting	5.6
prey	9.8	lords	5.4
damned	8.4	weather	5.2
villain	7.7	eight	5.2
chair	7.6	mistake	5.2
doom	7.4	displeasure	5
tear	7.3	faces	5
tyranny	7.1	puts	4.8
damn	6.8	city	4.8
begins	6.3	majesty	4.8
food	6.2	quality	4.7
wolf	6.2	special	4.5
torture	6.2	faithful	4.5
coward	6	wishes	4.5
ago	5.9	liege	4.5
fury	5.7	waters	4.4
wretch	5.7	price	4.4
gait	5.7	maids	4.4
peevish	5.7	gate	4.3
giddy	5.7	are	4.3
banished	5.6	four	4.3
seat	5.5	ladies	4.2
curse	5.4	guest	4.2
act	5.2	knew	4.2
unhappy	5.2	leaves	4.1
reign	5.2	minds	4.1
buried	5.2	whereof	4.1
hill	5.2	ships	4.1
slew	5.2	door	4
triumph	5	conduct	3.9
rail	4.8	lieutenant	3.9
tomb	4.7	satisfaction	3.9
bones	4.7	expect	3.9
monstrous	4.6	punish	3.9
knee	4.6	advise	3.9
despite	4.6	certain	3.9
lovely	4.5	audience	3.8
divine	4.5	amongst	3.8
cheque	4.5	knock	3.7
slow	4.5	study	3.7
guilt	4.5	several	3.7
poison	4.5	came	3.7
wound	4.4	wisely	3.7
babe	4.4	felt	3.7
mourn	4.4	takes	3.7
beauteous	4.4	senate	3.6

While the numerical results are intriguing, the most interesting results in this paper are probably the so-called most informative features that let us know which words were most useful in making a classification. Note that these tables are not merely likelihoods but reflect the relative likelihood of a *th*- pronoun *vis-a-vis* a *y*- pronoun and vice versa. That is, based on the training set, the word *lordship* is 49.6 times more likely to occur with a *y*- pronoun than with a th-pronoun. The classifier used to produce these tables is trained over the entire corpus, and the results for the 78.3% accurate neighboring pronoun exclusive version appear in Table 4. The full-context tables would look very similar with the exception that the tops of each list would be peppered with the pronouns themselves.

3.1 *th*- pronouns

The most obvious trend is for certain conjugated second-person verbs to be strongly associated with *th*- pronouns. This reveals that, even though *you* could be used in Shakespeare's time as a singular subject, it could not be used with the second-person singular conjugation. This explains *shalt*, *wilt*, *hast*, *canst*, *dost*, and the like. Considering that the number distinction was largely lost and on its way to being entirely lost, it is not immediately obvious that this would be the case.

Aside from those conjugated verbs (in bold), the *th-* list is populated largely with words that one might use to address someone they dislike. These include *coward*, *villain*, and *fiend* as well as *cursed*. A typical usage for a word like that might be as in *Henry VI*, Part 1: 'This be Damascus, be thou **cursed** Cain,/ To slay thy brother Abel, if thou wilt.' (1.3.42–43). It is difficult to imagine any of those words being used in, say, a servant's respectful speech to his master.

The list also reveals some surprising facts, such as the low connotation of the word *fish*. Trinculo addresses Caliban, the sea monster frequently associated with *fish*, as 'thou deboshed **fish**, thou' (3.2.24) in *The Tempest*. The word is frequently used among the lowly fisherman in *Pericles*. And the servant Gregory, in the midst of a bawdy rant at the beginning of *Romeo and Juliet*, says, 'Tis well thou art no **fish**' (1.128). Interestingly, Hamlet's famous retort to Polonius's asking whether Hamlet knows who he is—'Excellent well; you are a **fishmonger**'—uses a *y-* pronoun despite displaying a distinct lack of respect.

Another interesting case is *thunder*. Why is *thunder* collocated with *th-* words? It seems that when the word *thunder* occurs in the neighborhood of a second-person pronoun, it is frequently used when addressing the heavens, as in Lear's famous 'And thou, all-shaking **thunder**,/ Smite flat the thick rotundity o' the world!' (3.2.6–7). Or in *Cymbeline*: No more, thou **thunder**-master, show/ Thy spite on mortal flies (5.4.30–31) and *Troilus and Cressida*: 'O thou great **thunder**-darter of Olympus' (2.3.9). This is actually quote consistent with what we should expect, as it has long been noted that the *th-* pronoun is used for apostrophe to address a ghost or god or other absent presence. Indeed, Hamlet consistently refers to his father's ghost with *thee* and *thou*—'Remember thee!' (1.5.95)—rather than the more formal pronoun one might expect for addressing a parent.

It is often claimed that lovers use *th-* pronouns to address each other, but there is very little in the classifier data to suggest that this is the case. That said, a spot check of a few unequivocally romantic dialogs in the corpus reveals that lovers do tend to use *th-* pronouns. The balcony scene in *Romeo and Juliet*, for instance, uses exclusively *th-* pronouns. One possible explanation for the lack of evidence in our table of informative features is that the words typically surrounding the second-person pronouns in lovers' speech may not necessarily be indicative of a romantic context.

3.2 *y-* pronouns

Among the list of most informative features for *y-* pronouns, the highest scoring words include words of deference like *lordship, ladyship, worship, madam, sir, humbly,* and *please*. The appearance of the word *humbly* on this list is slightly interesting in that it at first might seem inconsistent with lofty words like *lordship* and *ladyship*. But, of course, it is exactly the type of word one would expect to be used when addressing a superior, as in Iago's address to Othello: 'I **humbly** do beseech you of your pardon/ For too much loving you' (3.3.212–213).

A perhaps unexpected result turned up by this analysis is that *please* occurs predominantly with *you* words, whereas *prithee* occurs much more often with *th-* pronouns. This suggests that, while the two words have similar uses, their level of formality and politeness is quite different. Consider Prince Henry's address to his father King Henry IV in *Henry IV*, Part I: 'So **please** your majesty, I would I could' (3.2.18). Compare that to Falstaff's familiar address to Hal in the same play: 'And, I **prithee**, sweet wag, when thou art king...' (1.2.13–14). Perhaps this betrays the etymological origins of *prithee*, which comes from *pray thee* (*OED*, s.v. *prithee*).

The prevalence of plural nouns on the *y-* list reveals that there is still a tendency for *you* to be used for the plural, even though that had become a less strict constraint. Especially telling is the prevalence of *slaves* on the *y-* list—a word that one would expect to be more strongly associated with the *th-* pronouns. But it finds company among *voices, wives*, and other such plural nouns. The high score for *y-* pronouns for *voices* is largely a byproduct of one speech in Act 2, Scene 3 of *Coriolanus* in which Coriolanus uses the phrase *your voices* repeatedly when addressing a group of citizens.

4 Conclusion

The analysis presented in this article demonstrates that a quantifiable difference exists between Shakespeare's use of *th*- pronouns and *y*- pronouns for the second person as evidenced by the fact that they can be automatically disambiguated. In general terms, *y*- pronouns consistently collocate with high-register words, including titles for nobility and words of supplication. *Th*- pronouns co-occur with second-person conjugated verbs and with insult words and other base vocabulary. They are also used for apostrophe.

Because our object of study was exclusively plays consisting of representations of spoken dialog, it is reasonable to assume that this trend reflects a wider pattern of usage in Early Modern English. More broadly, this article joins a growing body of work that uses statistical and computational techniques to provide empirical evidence on the language of Shakespeare and other literary figures. This type of work can not only illuminate usage patterns but can also aid in literary analysis. That is not to say that such methodology could or should replace traditional literary analysis. For instance, I have noted that, based on the data, it is unusual that Hamlet uses *you* rather than *thou* to call Polonius a fishmonger. The task left to literary critics is to say what that means. Is it a small clue that Hamlet is not actually mad since he still has at least some sense of decorum? Would the incongruous use of a formal pronoun with the word *fishmonger* been funny to Shakespeare's audience? Answering such questions requires work outside the scope of a computational analysis of pronouns. But the availability of exciting new techniques in computational linguistics can and should be used as a springboard for asking them.

Acknowledgement

This work was completed while the author was funded by a Marshall Scholarship. The author also gratefully acknowledges Simon Horobin, Daniel Donoghue, Aditi Lahiri, and two very helpful anonymous reviewers.

References

Barber, C.L. (1976). *Early Modern English*. Edinburgh: Edinburgh University Press.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media.

Busse, U. (2002). *Linguistic Variation in the Shakespeare Corpus: Morpho-syntactic Variability of Second Person Pronouns*. Philadelphia: John Benjamins Publihsing Company.

Bosak, J. (1999). Shakespeare 2.0 <http://www.cs.wisc.edu/niagara/data/shakes/shaksper.htm> (accessed 30 January 2011).

Jurafsky, D. and Martin, J.H. (2008). *Speech and Language Processing*. NJ: Prentice Hall.

Hope, J. (2003). *Shakespeare's Grammar*. Bristol: Arden Shakespeare.

Mitchell, B. (1985). *Old English Syntax*. Oxford: Clarendon Press.

Wales, K.M. (1983). Thou and You in Early Modern English: Brown and Gilman Re-Appraised. *Studia Linguistics*, **37**: 107–125.

Walker, T. (2007). *Thou and You in Early Modern English Dialogues: Trials, Depositions, and Drama Comedy*. Philadelphia: John Benjamins Publishing Company.

Note

[1] A Maximum Entropy Classifier was also implemented, but it achieved an only marginally better result.