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

Isolde van Dorst

Lancaster University (UK)
University of Malta (MT)
University of Groningen (NL)

Background: Early Modern English

- Early Modern English (EModE): 1500-1700
  - William Shakespeare: 1564-1616
- T/V distinction
  - Still occurs in other European languages (German *du*/*Sie*, French *tu*/*vous*, Spanish *tú*/*vos*)
  - In EModE:

	Nominative	Accusative/Dative	Possessive
YOU	You	You	Your
THOU	Thou	Thee	Thy/Thine

- YOU/THOU; *you*/*thou*/*thee*

Background: Research on pronoun use

- Power and solidarity, gender, age, status, genre, emotion, role of (situational) markedness
- “It is not so much ‘polite’ as not ‘impolite’; it is not so much ‘formal’ as ‘not informal’ ” (Quirk, 1974, p. 50)
  - It is not a static choice, but a situational marker
- One big issue: Use of raw frequency counts
- Another issue: Most studies were done on a small dataset
- Results so far have been contradictory

Hypotheses

- Null-hypothesis: No single model will be able to predict the pronominal address term solely based on linguistic and extra-linguistic features.
- Hypothesis 2: The features of social status, age and sentiment will be better prodictors of the pronoun choice than other features.
- Hypothesis 3: The best performing algorithm will combine features both dependent and independently.

Encyclopaedia of Shakespeare's Language

http://wp.lancs.ac.uk/shakespearelang/

@ShakespeareLang

- AHRC-funded research project at Lancaster University
- 38 plays: 36 from the First Folio, plus *Two Noble Kinsmen* and *Pericles: Prince of Tyre*
- Approx. 1 million words
- Richly annotated: Speaker ID, gender, genre, play name, scene
  - Social status:

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

Data & Features

- 22,932 instances
  - 14,365 *you*; 5,489 *thou*; 3,078 *thee*

- 23 linguistic and extra-linguistic features
- 10 pre-annotated: Genre, play name, play/act/scene, speaker ID, speaker gender, speaker status, production date, addressee gender, addressee status, no. people addressed
- 10 automatic: N-gram (LW1-3, RW1-3), positive sentiment, negative sentiment, addressee ID, status differential
- 3 manual: Speaker age, addressee age, location

Methodology

- 3 algorithms: Naive Bayes, decision tree, support vector machine
- Implemented through Weka
  - Feature ablation
  - Evaluated through 10-fold cross-validation
- Two types of classification
  - Trinary classification: *you*/*thou*/*thee*
  - Binary classification: YOU/THOU
- Baseline based on the distribution of the pronouns
  - 62.6% YOU; 37.4% THOU

Results: Binary classification

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	

Results: Feature comparison

Algorithm	Type	Features included
Naive Bayes	Trinary	LW1, LW2, RW1, RW2, speaker ID
	Binary	LW1, LW2, LW3 RW1, RW2, RW3, addressee ID
Decision tree	Trinary	LW1, LW2, RW1, RW2, speaker ID, status differential, negative sentiment
	Binary	Scene, speaker ID, speaker gender, addressee ID, addressee status, addressee age, status differential, positive sentiment
Support vector machine	Trinary	LW1, RW1, speaker ID, speaker age, addressee ID, addressee age, no. of people addressed, status differential, positive sentiment, negative sentiment
	Binary	LW1, RW1, speaker ID, speaker age, addressee ID, addressee age, no. of people addressed, status differential, positive sentiment, negative sentiment

- Most surprising model: Binary decision tree
- Most prominent features: N-gram, speaker ID
- Features in none of the models: genre, play name, production date, location

Hypotheses

- Null-hypothesis: No single model will be able to predict the pronominal address term solely based on linguistic and extra-linguistic features.
  - Best model (binary support vector machine) scores 24% higher on accuracy than the baseline (with 87%)
- Hypothesis 2: The features of social status, age and sentiment will be better prodictors of the pronoun choice than other features.
  - Partly true as they were indeed good predictors, but the actual best predictors were the N-gram (LW1 and RW1) and speaker ID
- Hypothesis 3: The best performing algorithm will combine features both dependent and independently.
  - On all scores, support vector machine scored best
  - However, Naive Bayes scored surprisingly well
  - Depends on preference: simplicity or complexity?

Conclusion

- Overall, it is possible to predict the pronoun based on the linguistic and extra-linguistic features
- Some features are definitely influencing the pronoun choice more than others
- Features are mostly independent of one another
- Linguistic context appears to be the key

- Some limitations
  - Familiarity (social distance)
  - Automatic tagging of the addressee

Thanks for your attention.

Questions?

References

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

Busse, Beatrix. (2006). *Vocative constructions in the language of Shakespeare* [Pragmatics & Beyond 150]. Amsterdam/Philadelphia: 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.

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.

Stein, Dieter. (2003). “Pronomial usage in SHakespeare: Beteween 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.

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.

Feature examples

Feature	Value	Explanation
Genre	T	Tragedy/Comedy/History
Production date	1595	
Play name	RJ	*Romeo and Juliet*
Play/act/scene	RJ_2_2	*Romeo and Juliet*, Act 2, Scene 2
Location	Private	Private/Public
N-gram LW3	'Romeo'	Word occurring third place on the left of the pronoun
N-gram LW2	'wherefore'	Word occurring second place on the left of the pronoun
N-gram LW1	'art'	Word occurring first place on the left of the pronoun
N-gram RW1	'Romeo'	Word occurring first place on the right of the pronoun
N-gram RW2	'deny'	Word occurring second place on the right of the pronoun
N-gram RW3	'thy'	Word occurring third place on the right of the pronoun
Speaker ID	RJ_Juliet	Character called 'Juliet' from *Romeo and Juliet*
Speaker gender	Female	Male/Female/Dressed as male/Dressed as female
Speaker status	1	Value from 0-7
Speaker age	Younger	Younger/Adult/Older
Addressee ID	RJ_Romeo	Character called 'Romeo' from *Romeo and Juliet*
Addressee gender	Male	Male/Female/Dressed as male/Dressed as female
Addressee status	1	Value from 0-7
Addressee age	Younger	Younger/Adult/Older
No. of people addressed	Singular	Singular/Plural
Status differential	0	Speaker status – Addressee status
Positive sentiment	1	Value from 1-5
Negative sentiment	-2	Value from -1 – -5
PRONOUN	thou	

Data distribution

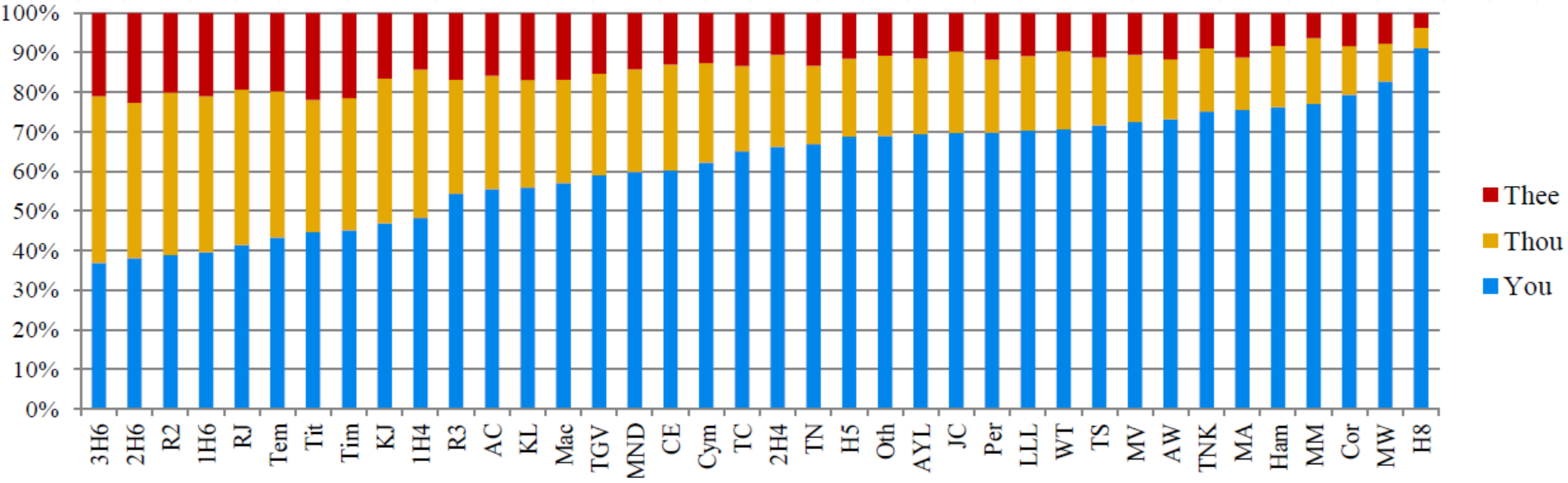

Figure: Relative pronoun distribution per play, ordered by percentage of YOU

- No. of pronouns extracted from each play range from 363 (in *Macbeth*) to 811 (in *Coriolanus*)
- In *Henry VIII*, almost no THOU pronouns occur

Results: Trinary classification

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	