Gender in Shakespeare: Automatic Stylistics Gender Character Classification Using Syntactic, Lexical and Lemma Features

Sobhan Raj Hota, Shlomo Argamon
Laboratory of Linguistic Cognition,
Computer Science Department,
Illinois Institute of Technology, Chicago, IL, 60616
hotasob@iit.edu, argamon@iit.edu

Rebecca Chung
Lewis Department of Humanities
Illinois Institute of Technology, Chicago, IL, 60616
chung@iit.edu

Abstract

For a variety of text types, methods for automatically determining the gender of a document's author can now reliably achieve accuracy of at least 70-80%. Our aim here is to extend this research, to examine determining the gender of literary characters from the author's differing word use between characters of different genders. Here we describe results showing how Shakespeare used language differently for his male and female characters, and we have studied the top discriminating features from characters of both genders. We used Sequential Minimal Optimization (SMO) to classify of gender character, based on various lexical and syntactic features to analyze the language Shakespeare used for gendering characters. Our methods achieve classification accuracy as high as 82% for classifying character gender. We further observe several interesting patterns in the most distinguishing features, including the fact that some constellations of features match well to previous reports of features that distinguish between male and female authors.

1 Introduction

A recent development in the study of language and gender is the use of automated text classification methods to examine how men and women might use language differently. Such work on classifying texts by gender has achieved accuracy rates of 70-80% for texts of different types (e-mail, novels, non-fiction articles), indicating that noticeable differences exist (de Vel et al. 2002; Argamon et al. 2003). More to the point, though, is the fact that the distinguishing language features that emerge from these studies are consistent, both with each other, as well as with other studies on language and gender. Work on more formal texts from the British National Corpus (Argamon et al. 03) similarly shows that the male indicators are mainly noun specifiers (determiners, numbers, adjectives, prepositions, and post-modifiers) indicating an 'informational style', while female indicators are a variety of features indicating an 'involved style' (explicit negation, first- and second-person pronouns, present tense verbs, and the prepositions "for" and "with").

To the best of our knowledge, there has been little work on understanding how novelists and playwrights portray (if they do) differential language use by literary characters of different genders. To apply automated analysis techniques, we need a clean separation of the speech of different characters in a literary work. In novels, such speech is integrated into the text and difficult to extract automatically. To carry out such research, we prefer source texts which give easy access to such structural information; hence, we focus on analyzing characters in plays. We have been extending this research on gender in text to analyze the relation of language use and gender for literary characters from a playwright's word use. In our recent work in understanding gender character using Moby Shakespeare (Hota et al. 06), we studied only two lexical feature sets (Function Words - FWs and Bag of Words - BoWs) we found stylistic differences exists how Shakespeare

presents character gender. Now we are interested in observing if syntactic, lexical and lemma features play a significant role in Shakespeare's written distinctions between character gender and in analyzing the features from both genders.

We thus ask the following questions. Can the gender of Shakespeare's characters be determined from their word use? Are the differences (if any) between male and female language in Shakespeare's characters similar to those found in modern texts by male and female authors? Keep in mind that here we examine text written by one individual (Shakespeare) meant to express words of different individuals with differing genders, as opposed to texts actually written by individuals of different genders. Which feature sets provide language use information for discriminating gender character and what are the top features for both genders? To address these questions, we applied text classification methods using machine learning. High classification accuracy, if achieved, will show that Shakespeare used different language for his male and for his female characters. If this is the case, then examining the most important discriminating features should give some insight into these differences and their possible relation to previous work on male/female language. The general approach of our work is to achieve a reasonable accuracy using syntactic, lexical and lemma based various feature sets collected from the characters' speeches as input to machine learning, and then to study those features that are most important for discriminating character gender.

2 Corpus Construction

We constructed a corpus of characters' speeches from 35 of Shakespearean plays, collected from the Nameless Shakespeare[1]. The Nameless Shakespeare is fully tagged for parts of speech (PoS) and lemma for each lexical occurrence, and while we do not expect its editorial procedures to unduly affect our differential analysis. The plays are basically in XML (Extensible Markup Language) format and to import them into our system, we extracted the speeches of each character automatically by cleaning the stage directions. The gender of each character was imported automatically; it was readily available in the XML file. A text file for each character in each play was constructed by concatenating all of that character's speeches in the play. We only considered characters with 200 or more words. From that collection, all female characters were chosen. Then we took the same number of male characters as female characters from a play, restricted to those not longer than the longest female character from that particular play. In this way, we balanced the corpus for gender, giving a total of 101 female characters and 101 male characters, with equal numbers of males and females from each play (Table - 1). We also split each corpus (somewhat arbitrarily) into 'early' and 'late' characters. We used the term 'early' to those plays which were written in sixteenth century and 'late' to those plays in seventeenth century. The chronology in plays as captured from Wikipedia[2]. We have 58 male characters and 58 female characters from early Shakespeare. Like wise we have 43 male and 43 female characters were chosen from late Shakespeare. The overall corpus statistics is given in Table – 2.

3 Feature Extraction

We processed the text using the ATMan system, a text processing system in Java that we have developed[3]. In ATMan text is tokenized and the system produces a sequence of tokens; each corresponds to a word in the input text file. We use syntactic (PoS), lexical and lemma features with n = 1, 2 and 3 gram combinations, in order to understand the gender a bit deeper as a stylistic approach to solve this classification approach(Tables 3, 4, 5). There are 31 various unigram PoS are collected from the corpus (Table - 6). We also collected 2259 bi grams and 1634 tri grams from the corpus which occurs more than ten times. The lexical features also contain the stylistic feature set (FWs) is a list (645) of more-or-less content-independent words comprising mainly function words, numbers, prepositions, and some common contractions (e.g., "you'll", "he'll"). We collected content-based feature set comprises all

[1] http://www.library.northwestern.edu/shakespeare/

[2] http://en.wikipedia.org/wiki/Chronology_of_Shakespeare_plays

[3] http://lingcog.iit.edu/download.xml

words that occur more than ten times in a corpus, termed Bag of Words (BoWs), which serve as unigrams (2426) in lexical category and then we collected bi grams (2670) and tri grams (356). We also collected most frequent 500 lexical features in a separate feature set. In the same way 2001 unigrams, 2860 bi grams, and 571 tri grams from lemma were collected. We calculate the frequencies of these various features and turn them into numeric values by computing their relative frequencies. The list of various feature sets with their counts is given in Table (3-5).

4 Text Classification

The classification learning phase of this task is carried out by Weka's (Frank & Witten 1999) implementation of Sequential Minimal Optimization (Platt 1998) (SMO) using a linear kernel and default parameters. The output of SMO is a model that linearly weights the various features. Testing was done via 10 fold cross validation. For better error estimation, we repeated the cross validation process 10 times and then average the results. This solution was chosen to get a reliable error estimate. We used two different thresholds (default and 0.0) for filtering features for information gain (thus removing features likely to be meaningless and thus possibly distracting to the learner); this was done with Weka's (Frank & Witten 1999) Information gain attribute evaluator. Information gain is defined as the expected reduction in entropy caused by partitioning the training set according to the attribute.

Many feature combinations give classification accuracies near or above 70%, which is quite good (random would be 50%, since the corpus is balanced). The highest accuracy of all **(82.66%)** was attained using the 500 most frequent word lemmas as features. Lexical (surface tokens) also worked well, with unigrams+bigrams giving the highest accuracy (79%). PoS features did not work as well, with unigrams giving the highest accuracy, at 67%.

In both the Early and Late sub corpora, we found that the most frequent word lemmas gave the highest accuracies in each category (72% for Early, 70% for Late). Note that in these cases there is less training data to learn from, so lower accuracies than on the entire corpus are to be expected. In the Early corpus, interestingly, an even higher accuracy (73%) is attained by using unigram PoS features, although PoS features do not work at all well for the Late corpus (the best accuracy is 65% for PoS trigrams). This would seem to indicate that over Shakespeare's he shifted from characterizing gender via syntax together with lexis towards a more subtle lexical approach, which kept syntactic features more constant between the sexes.

5 Feature Analysis

The feature analysis phase is carried out by taking the results obtained from Weka's implementation of SMO (Sequential Minimal Optimization). SMO provides weights to the features corresponding to both class labels. To discriminate binary class labels, SMO uses positive and negative weight values in a linear model. After sorting the features based on their weights, we collected the top ten features indicative of each gender.

Feature Analysis:
We collected top-10 features[4] (Features Analysis section) from both character gender and analyzed them on over all and with chronological impact on gender based on the early and late Shakespeare. For reasons of space we consider here just those feature sets giving the most insight (as well as good classification accuracy).

PoS unigrams:
We observed that Shakespeare's female characters more often use adverbs, interjections, adjectives, personal pronouns, negations, and question words, while male characters use more determiners, articles, prepositions, subordinating conjunctions, modal verbs, adjectival particles, infinitives, and question pronouns. These observations overall are largely in line with previous work (Argamon et al. 2003) on discriminating author gender in modern texts, supporting the idea that the playwright projects characters' gender in a manner consistent with authorial gender projection. While parts-of-speech are less distinguishing in Late Shakespeare than in Early, the same pattern of significant features obtains for both.

[4] http://lingcog.iit.edu/~hotasob/Nameless_Results/

Lemma unigrams:
Here emerges several meaningful clusters of words that give shape to 'typical' female and male concerns in Shakespeare. Female lemmas indicate concern with family relationships ('husband', 'mother', 'court') and feelings ('sick', 'merry'), as well as expression of personal feelings ('alas', 'o', 'prithee') and integrating personal context into the discourse ('he', 'you'). On the other hand, male features indicate concern with quantification ('three'), social status ('noble', 'solemn', 'savage'), as well as some more obscure verb forms ('begin', 'alight', 'beat'). In Early Shakespeare, we see similar patterns in different words, with the addition for females of negation (as 'never') and for males of Wh-words ('whence', 'wherein', 'who'); in Late Shakespeare the picture is less clear, with females preferring exclamations ('prithee', 'o', 'alas'), motion words ('hie', 'messenger'), and some others ('I', 'dear', 'sharp', 'such', 'false'), and males preferring determiners ('the') and prepositions ('of', 'to'), together with certain verbs ('begin', 'beat', 'embrace') and nouns ('motion', 'loss', 'description').

Lemma trigrams:
More specific meaning patterns can be seen in lemma triples. Female trigrams mostly indicate construal of self and others ('I/see/you', 'for/I/to', 'I/know/I', 'be/he/not', 'say/I/be')[5], with trigrams of politeness ('thank/you/for'), conditionals ('if/he/have'), and questions ('who/be/that'). Male trigrams, on the other hand, focus on assertions ('I/say/to') mainly about personal/social status ('but/I/be', 'be/a/very', 'be/a/ass', 'I/be/he'), possessions ('have/no/more', 'I/have/lose'), and manner ('the/manner/of'). In Early Shakespeare, we see female trigrams expressing epistemic stances ('do/not/know', 'I/can/tell', 'I/can/speak'), assertions ('when/they/be', 'be/not/to', 'be/not/yet'), questions ('who/be/that'), and telling expressions of friendship and love ('I/love/you', 'fare/you/well'), while Early male trigrams show mainly assertions of personal status and behavior ('I/go/to', 'but/I/be', 'I/be/in', 'I/lord/of', 'when/I/have', 'when/I/be', 'be/a/ass', 'I/have/see'), as well as requests ('I/beseech/you') By contrast, in Late Shakespeare, we see female trigrams showing requests ('do/beseech/you', 'I/do/beseech'), while male trigrams show more factual and personal assertions ('there/be/no', 'but/it/be', 'this/be/a', 'be/not/yet', 'thou/be/a', 'but/I/be').

[5] Note that these are lemmas, hence the actual realizations in the text may be different; e.g., 'be/he/not' may be realized as 'is he not'.

6 Discussion

These findings capture word patterning in Shakespeare inaccessible to non-computational methods of literary analysis, because of the scale of data processing involved. Literary scholars work almost exclusively with well-elaborated methods of semantic analysis (New Criticism, structuralism, and post-structuralism), developed with all the strengths and limitations posed by a book-only, eye-centered, subjectivity-dependent research context. In contrast, these findings encourage comparisons between non-computational and computational approaches. It is remarkable that these findings support aspects of non-computational methodology (words linked together in meaningful patterns like informational discourse/male and involved discourse/female), while also bringing to light new structural features of Shakespeare's gender-marking through language: parts of speech use, tri-gram combinations of words. These findings may, in addition, capture patterning below any literary author's conscious awareness, although still part of his or her creative process.

Limitations

The editorial procedures for *The Nameless Shakespeare* are sound and practical for the project's purposes and for the work of this paper, but they need to be read and understood fully: both by literary scholars wanting to apply these findings to particular words in particular plays, and by computational scholars thinking through the problem of establishing textual accuracy prior to inviting a wider community to conduct searches. Also, with respect to Shakespeare's literary art, the findings here do not at this stage account for the impact of blank verse dialogue (for high or elite characters) versus prose dialogue (for low or common-born characters) on word choices and the numbers of words. It may be that blank verse fosters se-

mantically significant tri-gram constructions because Shakespeare needed short words to complete plays mostly written in ten-syllable lines. But at least the question can be asked, and the answer will tell us something about Shakespeare both as dramatist and as poet.

7 Conclusions

This is the first work, to our knowledge, in analyzing various features (syntactic, lexical and lemma) collected from a single source in understanding literary character's gender. We see, as in our earlier work (Hota et al. 2003) that the male and female language in Shakespeare's characters is similar to that found in modern texts by male and female authors (Argamon et.al 2003). Here we also observed the importance of trigrams in PoS, lexical and lemma features, which are few in numbers, so became computationally effective and contain more information. We also see some evolution of Shakespeare's gendering of his characters, based on different patterns between Early and Late. The true import of the features identified by this analysis need to be confirmed by more traditional digital humanities methods such as examining concordance lines, to allow a more properly contextual interpretation. In any case, we believe that this study shows how classification learning can be used as a tool in developing new 'statistical' interpretative methodologies for bodies of literary works.

Acknowledgements:
Many thanks to Dr. Martin Mueller for providing us the Nameless Shakespeare corpus and many helpful comments in gender characterization in Shakespeare.

References:

Hota S., Argamon S., Koppel M., Zigdon I. (2006). Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters: Digital Humanities (DH), 100-106

Koppel M., Argamon S., Shimoni A. (2004). Automatically Categorizing Written Texts by Author Gender: *Literary and Linguistic Computing* 17(4).

Argamon S., **Koppel M.**, **Fine J.**, **Shimoni A**. (2003). Gender, Genre and Writing Style in Formal Written Texts: *Text* 23(3), pp. 321–346

Argamon S., **Koppel M., Avneri G.** (1998). Routing documents according to style. In proceedings of the First International Workshop on Innovative Internet Information Systems (IIIS-98)

Corney M., Vel O., Anderson A., Mohay G. (2002). *Gender Preferential Text Mining of E-mail Discourse:* In *Proceedings of 18th Annual Computer Security Applications Conference ACSAC*

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with many relevant features. *ECML-98, Tenth European Conference on Machine Learning.*

Mueller, Martin. (2005). *The Nameless Shakespeare*. TEXT Technology 14(1), pp. 61-70. http://texttechnology.mcmaster.ca/pdf/vol14_1_06.pdf

Platt, J. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Microsoft Research Technical Report MSR-TR-98-14,

Witten I., Frank E. (1999). *Weka3: Data Mining Software in Java* http://www.cs.waikato.ac.nz/ml/weka/

Tables:

Play Name	Gender Count
All's Well That Ends Well	8
Antony and Cleopatra	6
As You Like It	6
Cymbeline	4
King Lear	6
Loves Labours Lost	8
Measure for Measure	6
Midsummer Nights Dream	8
Much Ado About Nothing	8
Othello The Moore of Venice	6
Pericles Prince of Tyre	6
Romeo and Juliet	6
The Comedy of Errors	8
The First part of King Henry The Fourth	4
The First part of King Henry The Sixth	6
The Life and Death of Julies Caesar	4
The Life and Death of Richard The Second	6
The Life and Death of Richard The Third	8
The Life of King Henry The Eighth	6
The Life of King Henry The Fifth	4
The Merchant of Venice	6
The Merry Wives of Windsor	6
The Second part of King Henry The Sixth	4
The Taming of the Shrew	4
The Tempest	2
The Third part of King Henry The Sixth	4
The Tragedy of Hamlet	6
Titus Andronicus	4
Troilus and Cressida	4
Twelfth Night	6
Two Gentlemen of Verona	6
Winter's Tale	6
The Tragedy of Coriolanus	6
King John	6
Macbeth	8

Table 1: Shakespeare Corpus

	Male	Female
All	101	101
Early	58	58
Late	43	43

Table 2: Overall Corpus Statistics

Features	Count
Uni Gram	31
Bi Gram	2259
Tri Gram	1634
Uni plus Bi Gram	2290
Bi plus Tri Gram	3893
Uni plus Tri Gram	1665
Uni plus Bi plus Tri Gram	3924

Table 3: PoS Feature Sets

Features	Count
Function Words	645
500 Most Frequent Words	500
Bag of Words	2426
Bi Gram	2620
Tri Gram	356
Uni plus Bi Gram	5096
Bi plus Tri Gram	2976
Uni plus Tri Gram	2782
Uni plus Bi plus Tri Gram	5452

Table 4: Lexical Feature Sets

Features	Count
500 Most Frequent Words	500
Uni Gram	2001
Bi Gram	2860
Tri Gram	571
Uni plus Bi Gram	4861
Bi plus Tri Gram	3431
Uni plus Tri Gram	2572
Uni plus Bi plus Tri Gram	5432

Table 5: Lemma Feature Sets

Table 6: Parts of Speech Tag List

PoS-Tag	Part of Speech
pnx	Pronoun-Reflexive
v	Verb
dtq	Determiner-Question
cjs	Conjunction-Subordinating
fr	French Words
prp	Preposition
itj	Interjection
avq	Adverb-Question
chr	Character - A B C
neg	Negation
la	Latin Words
ajp	Adjectival Particle(French, Italian)
it	Italian Words
dt	Determiner
av	Adverb
aj	Adjective-Comparative
np	Proper Noun
ge	German Words
pni	Indefinite Pronoun
pnr	Pronoun (which, that, who)
vm	Verb-Modal
nu	Number(Ord-Cardinal)
pnp	Pronoun-Personal
pcl	Used when to+Verb
cjc	Conjunction
pnq	Pronoun-Question
fo	Foreign Words
cjq	Conjunction-Question
at	Article
vp	Verb Phrase
n	Noun