

Mining Classical Greek Gender

Helma Dik

helmadik@mac.com

University of Chicago, USA

Richard Whaling

rwhaling@uchicago.edu

University of Chicago, USA

This paper examines gendered language in classical Greek drama. Recent years have seen the emergence of data mining in the Humanities, and questions of gender have been asked from the start. Work by Argamon and others has studied gender in English, both for the gender of authors of texts (Koppel et al. 2002, Argamon et al. 2003) and for that of characters in texts (Hota et al. 2006, 2007 on Shakespeare); Argamon et al. (in prep. a) study gender as a variable in Alexander Street Press's Black Drama corpus (gender of authors and characters) and (in prep. b) in French literature (gender of authors).

Surprisingly, some of the main findings of these studies show significant overlap: Female authors *and* characters use more personal pronouns and negations than males; male authors *and* characters use more determiners and quantifiers. Not only, then, do dramatic characters in Shakespeare and modern drama, like male and female authors, prove susceptible to automated analysis, but the feature sets that separate male from female characters show remarkable continuity with those that separate male from female authors.

For a scholar of Greek, this overlap between actual people and dramatic characters holds great promise. Since there are barely any extant Greek texts written by women, these results for English give us some hope that the corpus of Greek drama may serve as evidence for women's language in classical Greek.

After all, if the results for English-language literature showed significant differences between male and female characters, but no parallels with the differences found between male and female authors, we would be left with a study of gender characterization by dramatists with no known relation to the language of actual women. This is certainly of interest from a literary and stylistic point of view, but from a linguistic point of view, the English data hold out the promise that what we learn from Greek tragedy will tell us about gendered use of Greek language more generally, which is arguably a question of larger import, and one we have practically no other means of learning about.

There is some contemporary evidence to suggest that Greek males considered the portrayal of women on stage to be true to life. Plato's Socrates advises in the *Republic* against actors portraying women. Imitation must lead to some aspects of these inferior beings rubbing off on the (exclusively male) actors

(*Republic* 3.394-395). Aristophanes, the comic playwright, has Euripides boast (*Frogs* 949f.) that he made tragedy democratic, allowing a voice to women and slaves alongside men.

Gender, of course, also continues to fascinate modern readers of these plays. For instance, Griffith (1999: 51) writes, on Antigone: "Gender lies at the root of the problems of *Antigone*. (...) Sophocles has created one of the most impressive female figures ever to walk the stage." Yet there are no full-scale studies of the linguistic characteristics of female speech on the Greek stage (pace McClure 1999).

In this paper, we report our results on data mining for gender in Greek drama. We started with the speakers in four plays of Sophocles (*Ajax*, *Antigone*, *Electra*, and *Trachiniae*), for a total of thirty characters, in order to test, first of all, whether a small, non-lemmatized Greek drama corpus would yield any results at all. We amalgamated the text of all characters by hand into individual files per speaker and analyzed the resulting corpus with PhiloMine, the data mining extension to PhiloLogic (<http://philologic.uchicago.edu/philomine>).

In this initial experiment, words were not lemmatized, and only occurrences of individual words, not bigrams or trigrams, were used. In spite of the modest size, results have been positive. The small corpus typically resulted in results of "100% correct" classification on different tasks, which is to be expected as a result of overfitting to the small amount of data. More significantly, results on cross-validation were in the 80% range, whereas results on random falsification stayed near 50%. We were aware of other work on small corpora (Yu 2007 on Dickinson), but were heartened by these positive results with PhiloMine, which had so far been used on much larger collections.

In our presentation, we will examine two questions in more depth, and on the basis of a larger corpus.

First, there is the overlap found between work on English and French. Argamon et al. (2007) laid down the gauntlet:

"The strong agreement between the analyses is all the more remarkable for the very different texts involved in these two studies. Argamon et al. (2003) analyzed 604 documents from the BNC spanning an array of fiction and non-fiction categories from a variety of types of works, all in Modern British English (post-1960), whereas the current study looks at longer, predominantly fictional French works from the 12th - 20th centuries. This cross-linguistic similarity could be supported with further research in additional languages."

So do we find the same tendencies in Greek, and if so, are we dealing with human, or 'Western' cultural, universals? Our initial results were mixed. When we ran a multinomial Bayes (MNB) analysis on a balanced sample, we did indeed see some negations show up as markers for female characters (3 negations in a top 50 of 'female' features; none in the male top

50), but pronouns and determiners show up in feature sets for both the female and male corpus. An emphatic form of the pronoun 'you' turned up as the most strongly male feature in this same analysis, and two more personal pronouns showed up in the male top ten, as against only one in the female top ten. Lexical items, on the other hand, were more intuitively distributed. Words such as 'army', 'man', 'corpse' and 'weapons' show up prominently on the male list; two past tense forms of 'die' show up in the female top ten. A larger corpus will allow us to report more fully on the distribution of function words and content words, and on how selections for frequency influence classifier results.

Secondly, after expanding our corpus, regardless of whether we find similar general results for Greek as for English and French, we will also be able to report on variation among the three tragedians, and give more fine-grained analysis. For instance, in our initial sample, we categorized gods and choruses as male or female along with the other characters (there are usually indications in the text as to the gender of the chorus in a given play, say 'sailors', 'male elders', 'women of Trachis'). Given the formal requirements of the genre, we expect that it will be trivial to classify characters as 'chorus' vs. 'non-chorus', but it will be interesting to see whether gender distinctions hold up within the group of choruses, and to what extent divinities conform to gender roles. The goddess Athena was the character most often mis-classified in our initial sample; perhaps this phenomenon will be more widespread in the full corpus. Such a finding would suggest (if not for the first time, of course) that authority and gender intersect in important ways, even as early as the ancient Greeks' conceptions of their gods.

In conclusion, we hope to demonstrate that data mining Greek drama brings new insights, despite the small size of the corpus and the intense scrutiny that it has already seen over the centuries. A quantitative study of this sort has value in its own right, but can also be a springboard for close readings of individual passages and form the foundation for a fuller linguistic and literary analysis.

References

- Argamon, S., M. Koppel, J. Fine, A. Shimoni 2003. "Gender, Genre, and Writing Style in Formal Written Texts", *Text* 23(3).
- Argamon, S., C. Cooney, R. Horton, M. Olsen, S. Stein (in prep. a). "Gender, Race and Nationality in Black Drama."
- Argamon, S., J.-B. Goulain, R. Horton, M. Olsen (in prep. b). "Vive la Différence! Text Mining Gender Difference in French Literature."
- Griffith, M. (ed.), 1999. *Sophocles: Antigone*.
- Hota, S., S. Argamon, M. Koppel, and I. Zigdon, 2006. "Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters." *Digital Humanities Abstracts* 2006.
- Hota, S., S. Argamon, R. Chung 2007. "Understanding the Linguistic Construction of Gender in Shakespeare via Text Mining." *Digital Humanities Abstract* 2007.
- Koppel, M., S. Argamon, A. Shimoni 2002. "Automatically Categorizing Written Texts by Author Gender", *Literary and Linguistic Computing* 17:4 (2002): 401-12.
- McClure, L., 1999. *Spoken Like a Woman: Speech and Gender in Athenian Drama*.
- Yu, B. 2007. *An Evaluation of Text-Classification Methods for Literary Study*. Diss. UIUC.