digital humanities
Lausanne – Switzerland '14

Unil
UNIL | Université de Lausanne

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

PROGRESS THROUGH REGRESSION. MODELING STYLE ACROSS GENRE IN FRENCH CLASSICAL THEATER

Schöch, Christof
University of Würzburg, Germany

Riddell, Allen
Dartmouth College, USA

Category: Long Paper
Session: 4
Date: 2014-07-10
Time: 09:00:00
Room: 413 - Amphimax

1. Introduction

Considerable scholarship in stylometry has focused on authorship attribution. Such work is based on the assumption that rates of high frequency "function" words (in contrast to "content" words) are reliable clues to authorship and are largely independent of factors like theme or genre[1]. More recently, focus seems to have moved beyond the most frequent words to involve all vocabulary appearing in a corpus ([2], [3], [4]). As many of these words vary strongly by context, factors like theme, genre, literary period or literary form have received greater attention.

This paper makes two contributions. First, we test the hypothesis that authorial style depends on genre and find that this is indeed the case, even when only considering the most frequent words. Second, in light of this result, we argue that adding additional features such as genre to a familiar model of authorship attribution offers a useful and novel way to investigate how authors' writing varies depending on context. We demonstrate how stylistic analysis making use of more articulate probabilistic models might move beyond established but limited models such as principal component analysis and distance-based clustering and achieve a better fit between model and hypothesis.

2. Data

In French literary studies, there is longstanding interest in analyzing the formal and stylistic constraints associated with classical theater ([5], [6]). Playwrights from this period, such as Pierre Corneille and Jean Racine, figure prominently in early quantitative work in French literary studies, predating the use of digital computers ([7], [8]). Whereas this pioneering research focused on single texts or a single author's works, today's availability of a wide range of digital texts, of flexible tools, and of vastly increased computing power permits more complex methods of analysis.

We have chosen to work on a corpus of 108 plays in three genres written by eight authors. The plays were produced over a period of roughly five decades (1630-1678) and the authors were selected because they wrote several plays in more than one genre. Table 1 illustrates the distribution of the plays across authors and genres.

	comedy	tragi-comedy	tragedy
Corneille, Pierre	9	1	20

Corneille, Thomas	8	0	15
Du Ryer	1	7	6
Molière	7	1	0
Quinault	1	1	3
Racine	1	0	9
Rotrou	1	4	3
Scarron	8	2	0
Totals	**36**	**16**	**56**

All texts are taken from the "théâtre classique" collection ([9]) and have been preprocessed to include only character speeches.[10] In order to better explore the variability of writing found among the authors and genres in the corpus, each play has been split into approximately 1,000 word sections. After processing, the corpus used for analysis contains 1,605 sections. Only the most frequent 100 function words in the corpus are retained.[11]

3. Hypothesis and Method

Our hypothesis is that authorial style varies depending on genre. In order to test this hypothesis, we compare three models that predict the author of a section based on word frequencies and the genre of the section. The first model predicts the author based on word frequencies alone, ignoring information on genre. The second model adds to the first rudimentary information about how likely authors are to appear in each genre. The third model differs from the second in that it predicts the author of a section based on word frequencies for each genre separately. If authorial style varies by genre, then the third model should perform significantly better than the first model.

All three models are multinomial logistic regressions.[12] Multinomial logistic regression has been used for authorship attribution before[13], but our approach expands on this by examining the use of a non-traditional covariate such as genre. Our aim is to encourage the building of interpretable models in order to understand how variables such as genre influence authorial style.

In statistical terms, the first model includes a global intercept parameter and word frequencies as predictors. The second model adds a genre-specific intercept parameter. The third model differs from the second model by allowing the regression coefficients associated with word frequencies to be different depending on the genre. These models may be expressed symbolically as shown in Fig. 1 (where the text section is indexed by *i* and *softmax_k(a)* is the extension of the inverse logistic function to multiple categories).

$$\begin{aligned}
&\text{Model 1: } \Pr(y_i = \text{author}_k) = \text{softmax}_k(\alpha + x_i\beta) \\
&\text{Model 2: } \Pr(y_i = \text{author}_k) = \text{softmax}_k(\alpha_{\text{genre}[i]} + x_i\beta) \\
&\text{Model 3: } \Pr(y_i = \text{author}_k) = \text{softmax}_k(\alpha_{\text{genre}[i]} + x_i\beta_{\text{genre}[i]})
\end{aligned}$$

$$\text{softmax}_k(a) \;=\; \frac{\exp(a_k)}{\sum_{j=1}^{K} \exp(a_j)}$$

Fig. 1: Three models of logistic regression

A point estimate for the parameters is obtained by maximizing the likelihood function using numerical methods. Models are fitted using randomly selected sections corresponding to four-fifths of the corpus. Models are then compared by measuring their out-of-sample predictions: an error rate for each model is calculated on the remaining fifth of the play sections, asking the model to predict the sections' authors based on word frequencies and, where applicable, genre information. This procedure is repeated fifty times, each time randomly partitioning the corpus.[14]

4. Results

The error rates associated with each model are shown in Fig. 2.

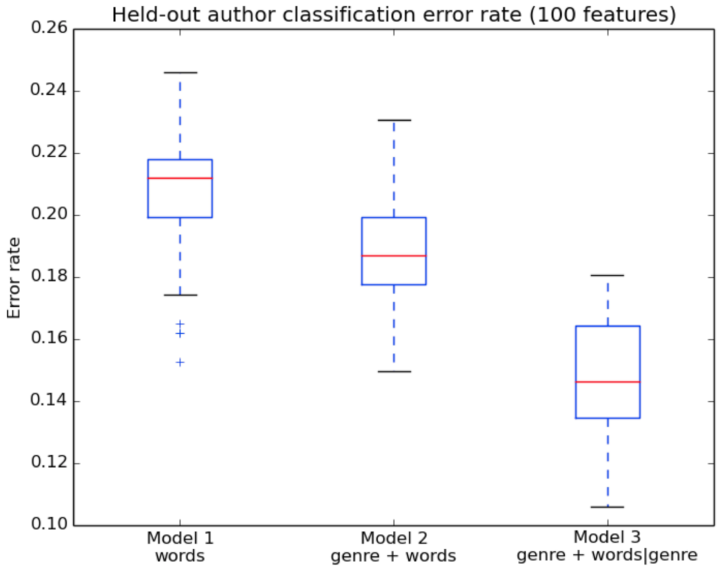

Fig. 2: Held-out author classification error rate (100 features)

In 49 out of the 50 trials, model 3 had the lowest error rate. In this corpus and for these authors, there is therefore little doubt that authorial style varies by genre. Table 2 shows the average error rates by model and genre.

	Model 1	**Model 2**	**Model 3**
Comedy	0.24	0.23	0.19
Tragi-comedy	0.25	0.22	0.10
Tragedy	0.17	0.15	0.13

5. Discussion

The variation of authorial style by genre underlying these results is best illustrated by looking at the frequencies of selected words that depend on both author and genre. For example, a few words are used with consistency across genres by one author but in another author vary considerably depending on genre. Table 3 indicates relative frequencies for three such cases.

	Pierre Corneille: comedy	Pierre Corneille: tragedy	Thomas Corneille: comedy	Thomas Corneille: tragedy
"est"	22.0	20.9	31.6	24.7
"par"	6.4	6.4	6.3	9.4
"au"	5.1	5.9	6.1	6.3

The auxiliary "est" and the preposition "par" are both used consistently across genres by Pierre Corneille but with a widely varying frequency between comedy and tragedy by Thomas Corneille, while the opposite behavior is true of "au". The preposition "par" is associated very frequently, in Thomas Corneilles plays, with causality (reason or effect) linked to emotions or moral principles (par bonté, par la gloire, par le respect). While the auxiliary "est" (third person singular present tense of "être") has an even more elusive semantic charge, it is mostly associated, in Thomas Corneille's plays, with statements of fact. Both phenomena seem to indicate a greater reliance, by Thomas Corneille, on causal relations and factuality in the tragedies than in the comedies, whereas the same contrasting treatment cannot be observed in Pierre Corneille.

The existence of such variation points to two notable facts. First, and contrary to common understanding, some very frequent function words other than personal pronouns do vary with genre within the work of a given author. Second, whether this is the case does not depend on the word in itself, but may differ from author to author. Therefore, such words are not exclusively or inherently markers of genre. Even when using only the very most frequent function words and even when excluding personal pronouns, then, authorship attribution cannot rule out that some influence from genre also comes into play.

On a different level, an explanation for the better performance of model 3 over model 1 brings in contextual information from literary history. Tragedies are usually described as being more closely bound to conventions of the "doctrine classique" than comedies or tragi-comedies ([15], [16]). Therefore, the range of vocabulary and the pattern of usage would be expected to be more predictable in tragedy than in other genres. Were this indeed the case, a model might achieve a lower variance in its predictions by considering tragedy separately. This hypothesis is difficult to evaluate as it is difficult to "hold constant" authorship; authors tend not to write in equal amounts in different genres.

A critical explanation of model 3's superior predictive performance would point out that the task of predicting an author on the basis of word frequencies might change dramatically depending on the authors being compared. It might therefore be suggested that the better performance obtained by model 3 reflects this fact more than it reflects within-author variation across genre. In response to this criticism, it should be observed that model 3 performs better even when the same authors are being compared; Pierre Corneille and Thomas Corneille dominate numerically the samples from comedies and tragedies. Furthermore, the words shown in table 3 demonstrate that there is variation within an author's style across genres. Model 3 is designed to use this variation to attribute authorship.

6. Conclusion

We offer the following conclusions from this experiment. First, authorial style does appear to vary with genre even when considering only the 100 most frequent words. This suggests that factors such as genre should be systematically taken into account for authorship attribution. Second, logistic regression is a useful method in this context and should be part of the stylometric toolbox as it permits a range of information to be modeled jointly with authorship. Logistic regression could also be used to test for further relevant factors beyond genre, such as form (e.g. verse and prose) or theme (e.g. historical plays vs. religious plays).

References

Hoover, D. (2004). *Testing Burrows's Delta*. LLC 19. 453-475.

Hoover, D. (2004). *Testing Burrows's Delta*. LLC 19. 453-475.

Burrows, J. (2007). *All the Way Through: Testing for Authorship in Different Frequency Strata*, LLC, 22. 27-47.

Rybicki, J. and Eder, M. (2011). *Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?* LLC 26: 315-21.

Bray, R. (1927).*La formation de la doctrine classique en France*. Paris.

Scherer, J. (1951). *La Dramaturgie classique en France*. Paris: Nizet.

Muller, Ch. (1967). *Étude de statistique lexicale: le vocabulaire du théâtre de Pierre Corneille*. Paris: Larousse.

Bernet, Ch. (1983). *Le Vocabulaire des tragédies de Racine*. Analyse statistique. Geneva / Paris: Slatkine / Champion.

Fièvre, P., ed. (2007-2013). *Théâtre classique*, www.theatre-classique.fr/.

Speaker names, stage directions, dramatis personae, prefaces, metadata and other paratextual elements have been excluded from the analysis. Trailing sections having fewer than 500 words were discarded. Trailing sections having between 500 and 1,000 words were normalized and put in terms of rates per 1,000 words.

Because of the relatively small sample size, three content words that may have an association with a specific genre ("coeur", "amour", and "yeux") appeared in the initial list of the top 100 most frequent words. These words were removed from the vocabulary so that the corpus contained only function words.The 100 most frequent graphical words used are: a, ai, au, autre, aux, avec, bien, c, ce, ces, cet, cette, comme, d, dans, de, des, donc, dont, du, elle, en, enfin, est, et, faire, fait, faut, grand, ici, il, j, jamais, je, l, la, le, les, lui, m, ma, mais, me, mes, moi, moins, mon, même, n, ne, non, nous, on, ont, ou, où, par, pas, peu, peut, plus, point, pour, puis, qu, quand, que, quel, quelque, qui, quoi, rien, s, sa, sans, se, ses, si, son, sont, suis, sur, t, tant, toujours, tous, tout, trop, tu, un, une, veux, voir, vois, vos, votre, vous, y, à, être.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2003) *Bayesian Data Analysis*. 2nd ed. Chapman and Hall/CRC, 2003, pp. 430-33.

Madigan, D., Genkin, D., Lewis, D.D., and Fradkin, D. (2005). *Bayesian multinomial logistic regression for author identification*. Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 803, 509-516.

It is worth noting the complexity of the models considered here. Model 3 has 2424 parameters (each genre has an intercept for each author and an 8 by 100 matrix of author-word coefficients). Fitting the model requires maximizing a function with 2424 parameters, something that was challenging a decade ago. To be clear, the maximization is not taxing; it requires roughly 300M of memory.

Bray, R. (1927). *La formation de la doctrine classique en France*. Paris.

Scherer, J. (1951). *La Dramaturgie classique en France*. Paris: Nizet.