DigitalHumanities2013

University of Nebraska–Lincoln, 16-19 July 2013

Abstracts

Fine-tuning Stylometric Tools: Investigating Authorship and Genre in French Classical Theater

July 17, 2013, 08:30 | Short Paper, Embassy Regents F

Schöch, Christof | christof.schoech@uni-wuerzburg.de | University of Würzburg, Germany

Topic(s):

- literary studies
- stylistics and stylometry
- text analysis
- french studies
- authorship attribution / authority
- data mining/ text mining

Keyword(s):

- stylometry
- authorship
- genre
- french classical theater

This paper is concerned with stylometric classification applied to French seventeenth-century plays. It reports on ongoing investigations into parameter setting and its impact on classification of such texts by author, genre, or form, using Eder & Rybicky's stylometric scripts for R (Eder & Rybicky 2011). Based on an investigation into the Corneille-Molière controversy, several methodological issues standing in the way of reliable results have been identified. One issue concerns the degree to which such authorship classification tasks are influenced by genre (here, comedy or tragedy) and form (here, verse or prose). Investigation of this issue shows that input parameters have indeed effects on the relative influence of authorship, genre and form in the classification of plays.

Stylometry today: advances and challenges

Stylometry has made significant advances in recent years, due no doubt to the increased availability of electronic texts, of sophisticated and accessible stylometric tools, and of proposed classification methods and distance measures. Based on this range of resources, researchers in stylometry are able to use various linguistic features as input for classification tasks and may adjust a wide range of parameters.

This situation, however, also brings renewed urgency to the issue of fine-tuning input parameters and distance measures, depending on the materials under scrutiny and the type of inquiry. Arguably, this is somewhat less of an issue today for a language such as English, where a well-established stylometric tradition exists. However, despite recent advances for some languages (Van Dalen-Oskam & Van Zundert 2007, Rybicky & Eder 2011), parameter setting remains an

insufficiently explored issue for languages such as German, French, Spanish, or Latin, and many more.

The Corneille-Molière controversy

This has been particularly apparent in the domain of seventeenth-century French drama, because work in this area has recently fuelled a controversy over whether or not Corneille was in fact the author of some or several plays traditionally attributed to Molière. In this controversy, traditional biographical and archival research (Boissier 2004) was complemented with results from stylometric analyses (Labbé & Labbé 2001; for a knowledgeable critique, see Brunet 2004; for a more recent approach, see Marusenko & Rodionova 2010). However, the methodological basis for stylometric analyses of this type of material seems to have been insufficiently investigated..

The conditions for reliable stylometric attribution results in this domain are challenging. The strong codification of classical literary discourse and the prevalence of stringent metrical forms mean that stylistic differences between authors are often subtle. At the same time, the available plays vary widely as to dramatic genre (e.g. comedy or tragedy) and form (i.e. verse or prose). Preliminary investigations into the Corneille-Molière corpus using Eder and Rybicky's stylometric scripts have indeed shown the fragility of the results. Depending on the composition of the text collection, on the linguistic material used as input, and on distance calculation measures, results vary widely.

Fine-tuning for author, genre and form

On the one hand, then, it can be challenging to make clear author attributions on material that is heterogeneous as to genre or form. For example, relevant research relying on the „unmasking“ technique showed unsatisfactory results for cross-genre authorship attribution (Kestemont et al. 2012). On the other hand, if only relatively homogeneous material is taken into account, the overall amount of data available for classification may be significantly reduced. What is needed is knowledge about how to limit the influence of factors other than the one of concern in any given classification task. The most relevant factors in the present case are authorship, genre (here, comedy or tragedy) and form (here, verse or prose). If the goal is to make reliable author attributions, how can the influence of genre and form be limited? The research reported on here was designed in order to explore such issues. All investigations are based on the *Théatre classique* collection (Fièvre 2007-2013), which provides XML/TEI versions of all plays. Texts were uniformely preprocessed to retain only character speeches, but no lemmatization was applied.

A first collection of plays was investigated limiting the number of relevant categories to just two, authorship and genre, and balancing the number of plays for each category. This resulted in a collection of 32 plays by Pierre Corneille and Thomas Corneille, with an equal number of comedies and tragedies by each author. The question at hand was to find out at which settings the classification would be dominated by either one of the author or genre category, and to what extent. Systematic variation was introduced as to the range of words from the frequency list taken into account: All runs relied on 100 words from the word frequency list, and each run took these from a moving onset point onwards, at an interval of 50 words.

For each run, the data was subjected to a distance measurement using Burrows' Delta (Burrows 2002), the results forming the basis of a cluster analysis. The distance tables were saved for each run, and the proportions of the different low-level pairs for each run were extracted. The low-level pairs found can be of the following types: author-and-genre match, author-only match, genre-only match, or pairs without a match. The proportions of author-only and genre-only matches is assumed to indicate to which extent the chosen settings give precedence to textual features associated with authorship or genre, respectively. Figure 1 visualizes the results from this investigation.

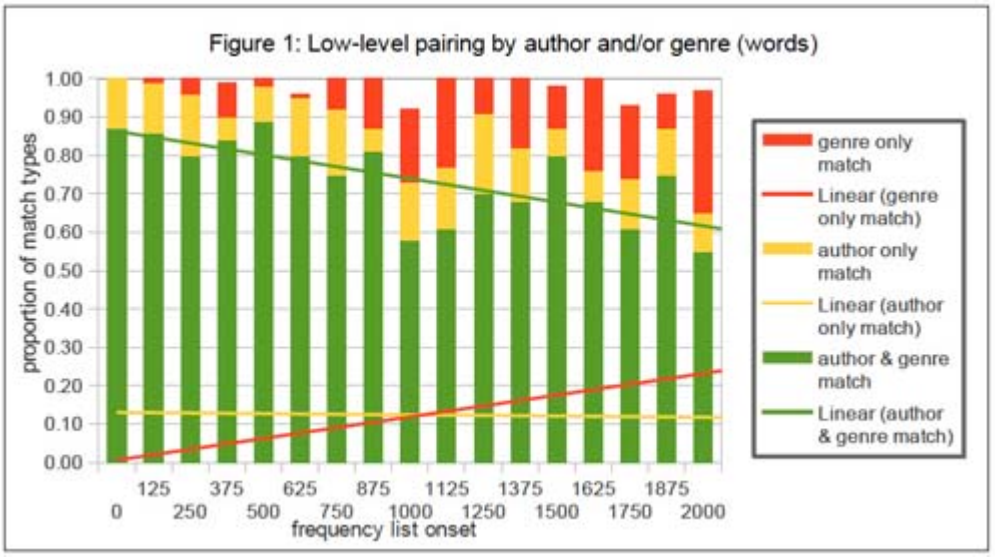

Figure 1

Author-and-genre matches decrease overall with increasing onset points, and the proportion of author-only matches remains relatively stable. However, the proportion of genre-only matches increases markedly with increasing onset points. In the range of onsets points between 0 and 1150 words (with two exceptions at 300 and 750 words), pairing is predominantly related to authorship, not genre. In the range of onset points from 1150 to 1650 words, pairing is related both to authorship and to genre, in varying proportions, while genre seems to be taking over more markedly beyond an onset point of 1650 words.

A similar investigation was run with a collection of plays with variation only as to authorship and form (verse or prose). The collection of plays consisted of 28 comedies, with an equal number of prose and verse plays by each of the following authors: Dufresny, Scudéry, Regnard, and Molière. Again, the nature and proportions of the different low-level pairs was assessed. With adjustment for the asymmetrical number of authors and forms, the graph shown here as figure 2 was constructed.

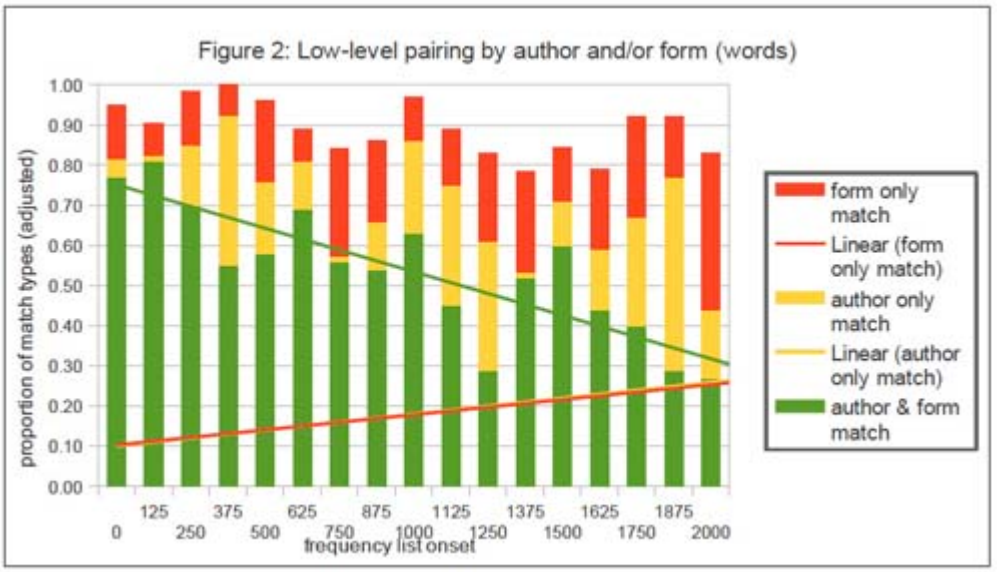

Figure 2

There is a very strong fall-off for author-and-form matches from an onset point of around 800 words. Although author-only and form-only matches increase somewhat over the range of onset points, this does not correspond to the decrease in author-and-form matches. The most important result is that over the entire frequency range, form-only matches are always present and in many though not all cases, they have a higher proportion than the author-only matches. Compared to the authorship vs. genre comparison (figure 1), there is certainly no clear cut-off point below or above which author-only matches would dominate form-only matches.

Conclusions

Despite their limitations, these preliminary results give some useful indications for authorship attribution studies in French classical verse drama, and may increase reliability of attributions. First, text collections of mixed dramatic sub-genre may be used in authorship classification tasks, provided that the wordlist used does not exceed the first 1150 most frequent words, so that influence from features related to genre remains limited. Second, form is a prevalent factor in the entire range of the frequency list, and should be controlled for when creating text collections. Applying these insights to the Molière-Corneille problem permits to enlarge the corpus of comparison texts beyond comedies, thus yielding a broader basis for classification tasks, but not beyond verse plays. While the procedure described here could be used for other languages and genre pairs, the results may be difficult to generalize: the best distinguishing parameters will likely be different from the ones found here for French classical drama.

However, more work needs to be done before the results obtained are sufficiently reliable. On the one hand, the approach taken here could be improved by enhancing the assessment of the dendrograms to take higher-level groupings into account as well. On the other hand, the fact that there are quite a few exceptions to overall trends shows the limit of this approach. In fact, a mechanism like feature selection may be more appropriate to solve the issue. Using supervised machine learning techniques with authorship, genre or form as separate target classes, and combining this with information gain analysis for each target class, would allow generating lists of features relevant for each target category.

Acknowledgements

I would like to thank Jan Rybicky and Maciej Eder for introducing me to their tools as well as João Guerra for helping me with some Python coding.

References

Boissier, D. (2004). L'affaire Molière. La grande supercherie littéraire, Jean-Cyrille Godefroy.
Brunet, É. (2004). Où l'on mesure la distance entre les distances. *Texto!*. (4) http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html (accessed 10 March 2013).
Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. LLC 17 (3) 267-287. 10.1093/llc/17.3.267
Eder, M., and J. Rybicki (2011). Stylometry with R. In *DH2011: Conference Abstracts*. Stanford University, Stanford, 308-11.
Fièvre, P., (ed.) (2007-2013). *Théâtre classique*, http://www.theatre-classique.fr/ (accessed 10 March 2013).
Kestemont, M., K. Luyckx, W. Daelemans, and T. Crombez (2012). Cross-Genre Authorship Verification Using Unmasking. *English Studies*
Labbé, C., and D. Labbé (2001). Inter-textual Distance and Autorship Attribution Corneille and Molière. Journal of Quantitative Linguistics. 8(3): 213-231. 10.1076/jqul.8.3.213.4100
Marusenko, M., and E. Rodionova (2010). Mathematical Methods for Attributing Literary Works When Solving the 'Corneille-Molière' Problem. *Journal of Quantitative Linguistics* 17(1): 30-54.
Rybicki, J., and M. Eder (2011). Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? *LLC*. 26(3): 315-321. 10.1093/llc/fqr031.
Van Dalen-Oskam, K., and J. van Zundert (2007). Delta for Middle Dutch. Author and Copyist Distinction in Walewein. *LLC*. 22(3): 345-362. 10.1093/llc/fqm012.

Links

- Alliance of Digital Humanities Organizations The Alliance of Digital Humanities Organizations (ADHO) promotes and supports digital research and teaching across all arts and humanities disciplines, acting as a community-based advisory force, and supporting excellence in research, publication, collaboration and training.
- The Center for Digital Research in the Humanities The Center for Digital Research in the

Humanities (CDRH) advances interdisciplinary, collaborative research, and offers forums, workshops and research fellowships for faculty and students in the area of digital scholarship.

- University of Nebraska–Lincoln The University of Nebraska–Lincoln, chartered in 1869, is an educational institution of international stature. UNL is listed by the Carnegie Foundation within the “Research Universities (very high research activity)” category. UNL is a land-grant university and a member of the Association of Public and Land-grant Universities (APLU).