Maciej EDER, Jan RYBICKI

Do birds of a feather really flock together, or how to choose training samples for authorship attribution

Pedagogical University of Kraków, Poland

Correspondence: Maciej Eder, Institute of Polish Studies, Pedagogical University of Kraków, ul. Podchorążych 2, 30-084 Kraków, Poland

E-mail: maciejeder@gmail.com

Abstract

This study investigates the problem of appropriate choice of texts for the training set in machine-learning classification techniques. Although intuition suggests picking the most typical texts (whatever ‘typical’ means) by the authors studied, any arbitrary choice might substantially affect the final results. Thus, to eschew cherry-picking, we introduce a method of verification of the choice of ‘typical’ samples, inspired by *k*-fold cross-validation procedures. Namely, we use a bootstrap-like approach to choose randomly, in 500 iterations, the samples for the training and the test sets. Next, we examine the obtained 500 attribution accuracy scores: if the density function shows widespread results, the corpus is assumed to be very sensitive to the permutations of the training set. To test this methodology empirically, we have selected roughly similar corpora in five languages: English, French, German, Italian and Polish. The results show considerable resistance of the English corpus to permutations, while the other corpora turned out to be more dependent on the choice of the samples; the Polish corpus produces both accuracy and consistency below any acceptable standards.

The Problem

In the house of non-traditional authorship attribution are many mansions, or methods for the statistical analysis of authorial style. They all compare text samples of disputed or unknown authorship to texts written by known authors, or ‘candidates.’ The degree of similarity or dissimilarity between samples allows informed guesses on the possible authorship of a given text.

Some of the techniques applied in attribution studies, usually called *unsupervised*, or *explanatory*, rely on the assumption that the obtained results ‘speak of themselves.’ It means that they require human interpretation of the degree of similarity between analyzed samples. These methods produce attractive and comprehensive graphs, but they are not sufficiently accurate in real attribution experiments due to being subjected to the attributor’s arbitrary decisions.

The state-of-the-art techniques, referred to as *supervised*, or *machine-learning*, perform an automated classification of input samples by assigning samples into classes; in other words, each text is linked to its presumed author. These machine-learning methods are supposed to be among the most effective; they include Support Vector Machines, Nearest Shrunken Centroid classification, K-Nearest Neighbor classification, Discriminant Analysis, Burrows’ Delta and so on (Burrows, 2002; Baayen et at., 2002; Koppel et al., 2009; Jockers et al., 2008; Schaalje et al., 2011; for a comparison of effectiveness of these methods cf. Jockers and Witten, 2010).

The feature common to the machine-learning methods is a two-step supervised analysis. In the first step, the traceable differences between samples produce a set of rules, or a classifier, for discriminating authorial ‘uniqueness’ in style. The second step is of predictive nature – using the trained classifier, the machine assigns other texts samples to the authorial classes established by the classifier; any disputed or anonymous samples will be assigned to one of the classes as well, provided that such a cassification is usually based on probabilistic grounds.

The procedure described above relies on an organized corpus of texts. Namely, the clue is to divide all the available texts into two groups: primary (training) set and secondary (test) set. The first set, being a collection of texts written by known authors (‘candidates’), serves as a sub-corpus for finding the best classifier, or discrimination rules, while the second set is a pool of texts of known authors, anonymous texts, disputed ones and so on. The better the classifier, the more samples from the test set are attributed (‘guessed’) correctly and the more reliable the attribution of the disputed samples.

Such procedures have been successful in social and medical studies; no wonder, then, that they soon made their way into authorship attribution. Yet contrary to, say, medical applications, where the researcher usually enjoys a high number of test samples (e.g. patients, various lab results, etc.), authorship attribution frequently has to struggle with a limited number of samples available to adequately train a classifier. This makes the classifier sensitive to statistical error. What is more, the generally-accepted division of data studied into a training set and a test set further limits the texts that can be attributed.

This sensitivity of machine-learning classifiers to the choice of samples in the training set has already been observed (Jockers and Witten, 2010: 220). Intuition suggests composing the training set from the most typical texts (whatever ‘typical’ means) by the authors studied (thus, for Goethe, *Werther* rather than *Farbenlehre*; for Dickens, *Great Expectations* rather than *A Tale of Two Cities*; for Sienkiewicz, his historical romances rather than his *Letters from America*). In practice, this can be quite complicated: in a small corpus, to change a single training set sample for another can upset the delicate mesh of interrelationships between all other texts. What is worse, a decision of what is ‘typical,’ is not trivial at all in many cases. Which is the ‘typical’ Joyce: the author of *The Portrait of the Artist* or of *Finnegan’s Wake*? Or which one novel to choose as ‘typical’ from the highly evolving *oeuvre* of James? The potentially heavy impact of sample selection has not been lost on Hoover: ‘As a reminder of how much depends upon the initial choice of primary and secondary texts, consider what happens if the same 59 texts are analyzed again, but with different choices for primary and secondary texts […]. If the analyses that are the most successful with the initial set are repeated, Delta successfully attributes only 16 of the 25 texts by members of the primary set’ (Hoover 2004a: 461).

Any manual selection of texts is highly arbitrary. To further quote Hoover: ‘The primary novels for this test are intentionally chosen so as to produce poor results, but one might often be faced with an analysis in which there is no known basis upon which to choose the primary and secondary texts, and nothing prevents an unfortunate set of texts like this from occurring by chance’ (Hoover, 2004a: 461–62). The Delta procedure – i.e. the technique used in the quoted study – does not include any validation of its results; Hoover’s statement indicates that this is not a minor matter.

A vast majority of machine-learning methods, however, routinely try to estimate the amount of potential error that may be due to inconsistencies in the training set samples. The general idea of such *cross-validation* tests is to replicate the original experiment multiple times with random changes to the composition of both the training and testing sets. In each of the iterations, the goal is to swap some of the samples from the test set with some of the samples from the training set. These permutations of the original data (often referred to as *folds* of *k*-folded cross-validation) are followed by a comparison of the classifier’s success. Ten-fold cross-validation is the standard solution (Tibshirani et al., 2003: 107; Zhao and Zobel, 2005; Baayen, 2008: 150; Baayen et al., 2002; Koppel et al. 2009; Jockers and Witten, 2010: 219; Luyckx and Daelemans 2011: 41). While this seems, for good reason, an accepted approach, the question might arise whether ten trials are sufficient for classifications tests that are based on only a small set of samples. With a relatively small amount of textual data, replication of stylometric experiments is difficult and requires special attention to the stability of obtained results (Rudman, 1998, 2003; Eder, 2010).

The classic ten-fold cross-validation, however, is not the only option. *Leave-one-out*, an extreme variant of *k*-fold cross-validation, relies on permuting the training and the test sets as many times as there are samples available; the result is a test in which all of the possible combinations of samples in the two sets are tested one by one (Good, 2006: 164). Despite some obvious advantages in this form of validation, there is also one inescapable drawback, i.e. the time needed for computing a given task is highly dependent on the number of samples to be analyzed. Assuming that the training set contains 10 samples by 10 authors, and the test set another 10 samples by these authors, there are 2^{10} (1024) possible combinations of members of the training set. For a corpus of 60 novels by 20 authors, this number becomes so large that testing all possible permutations of both sets is unrealistic.

Alternatively, the impact of the composition of the training set on attribution success can be assessed using a variety of bootstrap procedures and just several hundred random permutations (Good, 2006). The idea of bootstrapping is quite simple: in a large number of trials, samples from the original population are chosen randomly (with replacement), and this chosen subset is analyzed in substitution for the original population. It is entirely possible, then, that some of the observations are chosen several times, while some others are omitted by the procedure. The goal of the bootstrap, however, is to cover a wide range of original observations and to get rid of (possible) outliers. A bootstrap-like approach allows for random selection, in multiple iterations, of the samples for the training and the test sets.

The Experiment

To test this problem, we selected several corpora of similar size with a similar number of authors. We offer the obvious caveat that any comparison between different languages can never be fully objective. In the words of Juola, ‘it is hard to imagine ways to establish that two authorship attribution tasks are “comparably difficult” to enable such direct comparisons’ (2009: 163). As our own studies suggest, there are significant differences in performance between languages despite different sizes of analyzed corpora (Rybicki and Eder, 2011; Eder 2010, 2011). With this in mind, we decided that keeping the number of authors and texts exactly the same in corresponding corpora would be unpractical.

Five modern languages with established literary traditions have been chosen, and for each of the languages, a corpus of 19th- and/or 20th-century novels has been prepared. Then, each corpus has been studied three times: first as an entire corpus, second with a reduced number of texts written by particular authors, and third with a reduced number of authors. The corpora used are as follows:

– English novels: (a) 63 texts by 17 authors, (b) 53 texts by 17 authors, (c) 40 texts by 9 authors;

– French novels: (a) 71 texts by 25 authors, (b) 57 texts by 25 authors, (c) 44 texts by 12 authors;

– German novels: (a) 66 texts by 21 authors, (b) 59 texts by 21 authors, (c) 33 texts by 10 authors;

– Italian novels: (a) 77 texts by 9 authors, (b) 38 texts by 9 authors, (c) 31 texts by 4 authors;

– Polish novels: (a) 68 texts by 8 authors, (b) 39 texts by 8 authors, (c) 41 texts by 4 authors.

For each version of these corpora, we performed a controlled experiment of 500 iterations of attributive tests, each with random selection of the training and the test sets. The selections were made so that each author was always represented in the training set by a single (randomly-chosen) text, and that the test set always contained the same number of texts by each author. This was done to prevent a situation in which no works at all by one or more of the test authors would be found in the training set. We compared the number of correct author guesses in the context of our hypothesis that the more resistant a corpus is to changes in the choice of the two sets, the more stable the results.

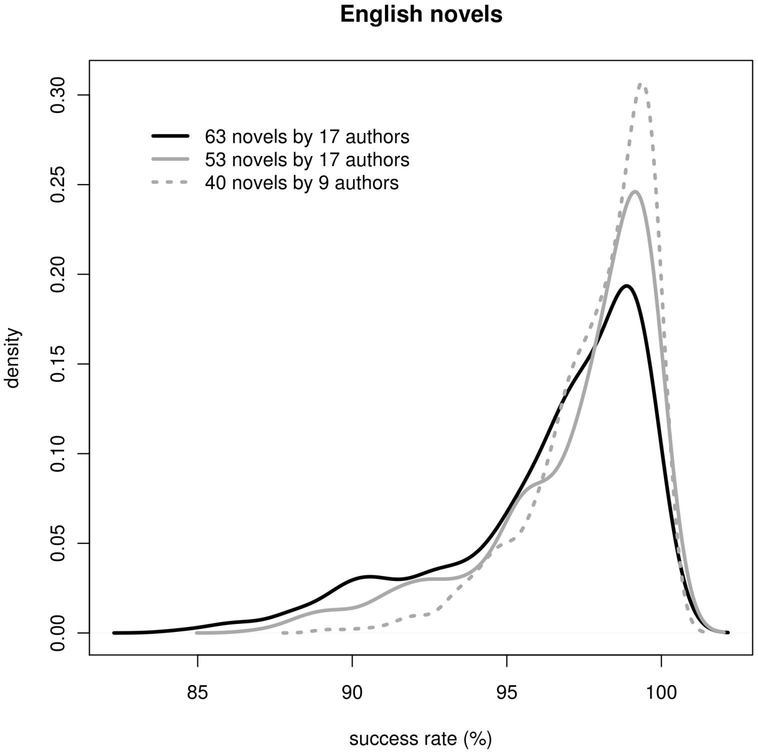
All tests featured the simplest, the very intuitive and the most frequently used of machine-learning attribution method: Burrows's Delta (Burrows, 2002; Hoover, 2004b).[1] To mitigate the problems associated with arbitrarily choosing the number of MFWs to be analyzed (Rybicki and Eder, 2011), and thus to obtain more reliable results, the controlled attribution test was run several times with different vectors of MFWs and with different settings for removal of too-characteristic words, or 'culling' (Hoover, 2004a). Thus, Delta was performed using 100 MFWs, then 200 and then, at increments of 100, all the way to 2,000 MFWs. This classing was performed at five different culling settings (0–100% incrementing by 20), giving a total of 1,000 results. The mean of these results was recorded, and the same procedure was then repeated for 500 random permutations

[1] For all the tests, an adjustable open-source statistical environment R was used (http://cran.r-project.org), together with a tailored multi-task R script developed by the authors (Eder and Rybicki, 2011).

of the texts in the training set. Contrary to the usual bootstrap approach, our interest concerned the entire distribution rather than just the means; a density function was then estimated for the final results thus obtained.[2]

It can be assumed that the distribution of each of these 500 final results should be Gaussian rather than anything else. The peak of the curve would indicate the real effectiveness of the method, while its tails – the impact of random factors. A thin and tall peak would thus imply stable results resistant to changes in the training set.

Figure 1 Density (vertical axis) of attributive success percentage rates (horizontal axis) in the English novel corpus



Our analysis of the results begins with the corpus of 63 English novels by 17 authors. As expected, the density of the 500 bootstrap results follows a (skewed) bell curve (Figure 1). At the

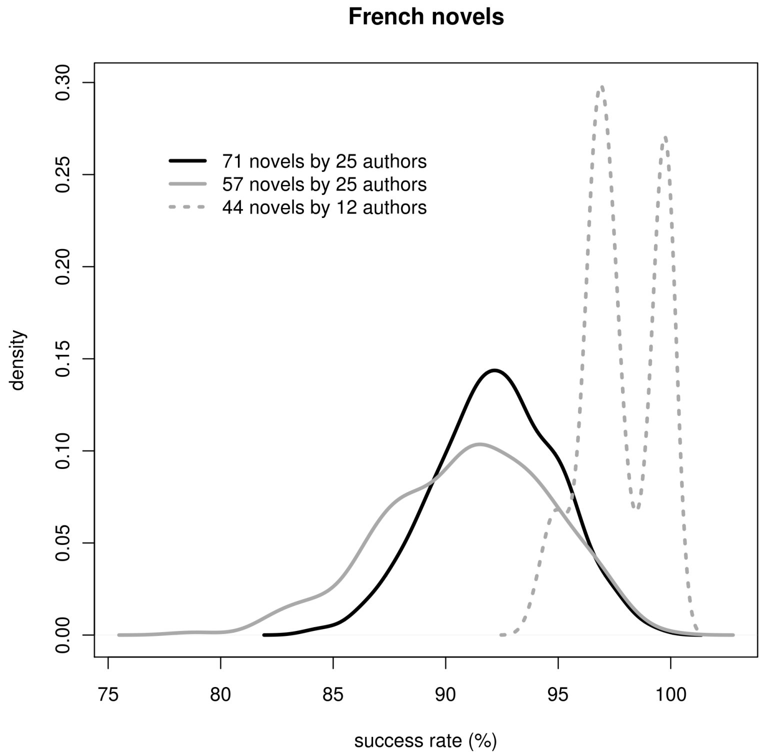
[2] Since the ‘classical’ Delta procedure counts z-scores based on the training set alone, and then applies the variables’ means and standard deviations to the test set, the permutation of both sets somewhat impacts the z-scores and thus, possibly, the final results as well. In view of this, we have performed a parallel set of tests with z-scores calculated for both sets; empirically, in this second approach, the success of attribution was slightly lower.

same time, its gentler left slope suggests that, depending on the choice of the training set, the percentage of correct attributions can vary, with bad luck, to below 90%.[3]

It is quite natural that the stability of the results might also depend on the number of authors and/or texts analyzed. The same Figure shows that, with fewer authors, a higher number of texts has no significant impact on the stability of the results at any permutation of both sets (the dashed line), a similar result has already been observed by Hoover and Hess (2009: 474). With more authors (i.e. when guessing becomes more difficult), the curve widens and a perfect match is even less frequent.

Nevertheless, this is still good accuracy and a fairly predictable model, as accuracies of less than 90% are exceedingly rare. However, it has to be remembered that Delta has been shown to be somewhat less perfect in other languages (Rybicki and Eder, 2011; Eder, 2010, 2011). This is quite visible in Figures 2–5.

Figure 2 Density of attributive success rates in the French novel corpus



[3] N.B. the density function curve that seems to exceed 100% is a result of smoothing and not of errors in procedure

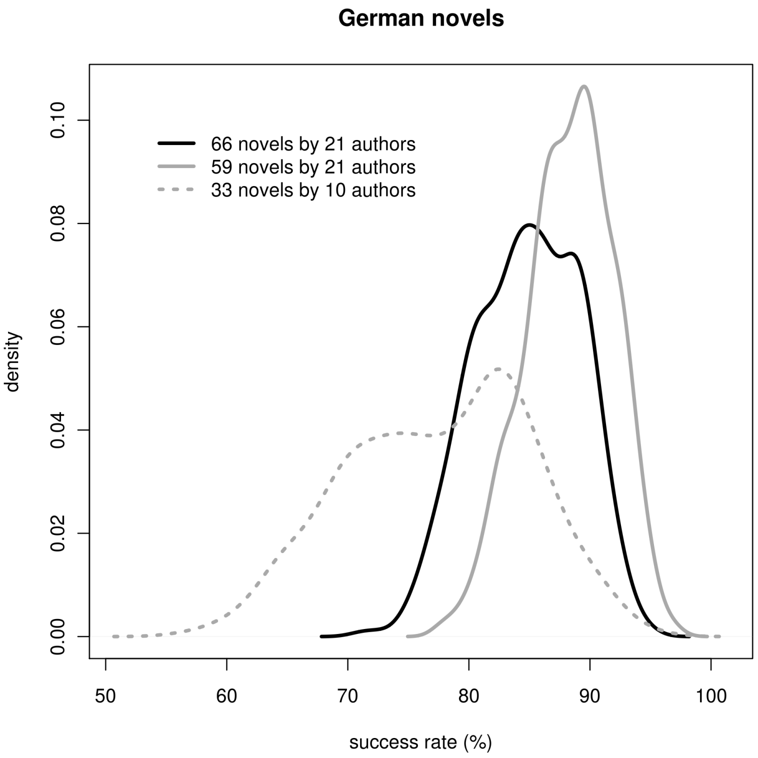


Figure 3 Density of attributive success rates in the German novel corpus

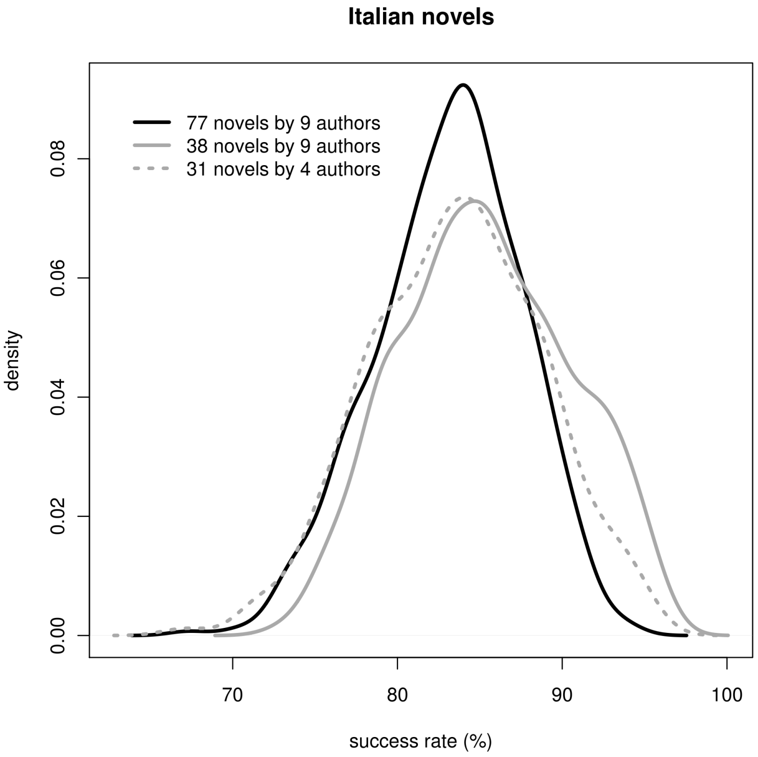


Figure 4 Density of attributive success rates in the Italian novel corpus

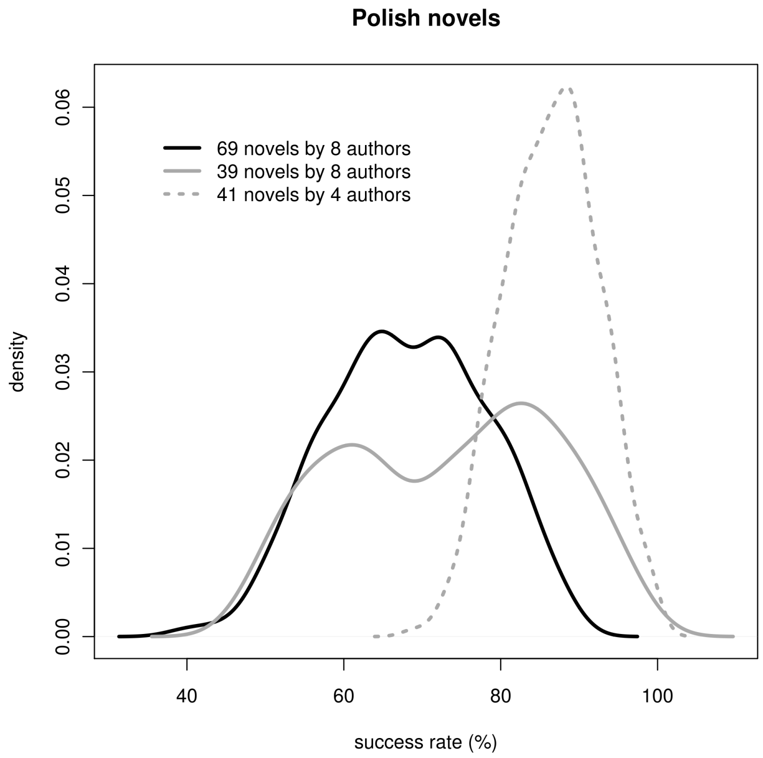


Figure 5 Density of attributive success rates in the Polish novel corpus

Indeed, the discrepancies in Figures 2–5 seem to call into question the validity of attribution tests that are based on arbitrary choice in training sets. The French corpus in Figure 2 exhibits a dramatic decrease in accuracy from the near-perfect guessing for fewer books by fewer authors (the dashed-line double-cone hat) to much more flattened curves when more authors and more texts *viennent semer la pagaïe*. Any hope that there is here, at least, a consistency in the effect of adding to or subtracting from the number of authors and/or texts in the corpus are *blitzkrieged* by the German texts, where the dashed line of the most extensive version of the corpus now reflects the worst accuracy and consistency of the three! By contrast, the Italian corpus is a *quintessenza* of consistency; but, at this time, the most frequent accuracy rate has receded to below 85%, i.e. into regions rarely visited even by the worst combinations of primary- and secondary-set texts in the English or even the French corpus. Much to the chagrin of the authors of this paper, the results for the Polish corpus are a true *katastrofa*: any attempt at broadening the corpus brings both accuracy and consistency below any standards acceptable.

Discussion

The results of this relatively simple experiment point in two different and somewhat discrete directions.

First comes the obvious observation from Figures 1–5 is that corpora exhibiting a lower overall guessing rate were also less consistent in that rate. The English corpus was not only stable but also stably achieved excellent results – showing, in a way, that it was lucky that Burrows developed his method with his native English as the main source of material! Polish literature attributes much worse and, adding insult to injury, it also exhibits the greatest dispersion. We have already been worried by the fact that literature in some languages guesses not as well as in some others (Rybicki and Eder, 2011; Eder, 2010, 2011), and we have tried to blame the different degree of inflection in the languages in question. Sadly, this was belied by very good guessing in Hungarian prose, and there is also the small matter of a tentative experiment by Rybicki with stemmed Polish texts giving even more disastrous results (Rybicki, 2009). The results of the experiment presented in this paper show that, indeed, the explanation might be much more complex: intuitively, the degree of inflection alone can hardly be a factor in such differences between the results for the various permutations. Further study is warranted here. A purely literary solution cannot be ruled out: while we tried to keep the corpora as similar and representative as possible for the individual national literatures, the fact that the attribution procedure stubbornly mistook historical romances by one Polish author for those by another, and equally stubbornly mixed up their novels on more contemporary themes might suggest that – for reasons as yet unexplained – genre is a much stronger stylistic factor in Polish 19th-century literature; or, even, that the principles of literary style in Polish are much more rigid than in the other literary traditions – and thus allowing less individual stylistic/lexical features.

Secondly, the huge discrepancies in guessing rate in some of our corpora beg for an inspection of the results to see *which* texts allow a good guessing rate when placed in the training set; in other words, which texts appear as the most ‘typical’ ones from the point of view of the machine (i.e. the computer running our Delta scripts); in yet other words, to stand the usual attributive task on its head and use authorship attribution to determine the most typical text by a selection of authors. This is just as interesting in the best-behaved English corpus: the combination of texts in the training set that produces the worst possible attribution accuracy for this corpus (86.99%) consists of Anne Brontë's *Tenant of Wildfell Hall*, Austen's *Pride and Prejudice*, Charlotte Brontë's *Villette*, Conrad's *Almayer's Folly*, Dickens's *Little Dorrit*, Eliot's *Middlemarch*, Fielding's *Tom Jones*, Galworthy's *To Let*, Hardy's *Far from the Madding Crowd*, James's *Ambassadors*, Kipling's *Kim*, Richardson's *Pamela*, Scott's *Waverley*, Sterne's *Sentimental Journey*, Swift's *Gulliver's Travels*, Thackeray's *Pendennis* and Trollope's *Prime Minister*. These can hardly be called 'untypical' texts for their authors, and should any eyebrows be raised by the presence of Hardy's relatively early work together with Sterne's travelogue, the two appear in another training set permutation that scores a full 100%. On the other end of the accuracy spectrum, in the Polish corpus, a poor 55% is scored by a training set that reads like an obligatory reading list for Polish secondary school students. While it might be difficult to pinpoint the exact mechanism of this phenomenon, it is worthwhile considering a certain element of machine learning methods: for the same collection of texts, the fact that certain samples are chosen for the training set has the consequence that they will not take part in the testing procedure. Elementary, one might say; yet this signifies that increases or decreases in the accuracy rate can be a result of *removing* some particularly difficult samples from the test set to the training set – or, inversely, the fortunate choice of the 'typical' samples for the training set.

In analyses of this kind, the arbitrariness cannot be entirely eliminated from the choice of texts. It is true that even this pre-selection might seem arbitrary – just as arbitrary as selecting novels irrespective of narrative mode, genre, or target reader (e.g. novels for adults vs. novels for children). On the other hand, since the same approach was used for all corpora, the possible distorting effect could be expected to be roughly similar on analyses in different languages.

Conclusion

In view of the above, the question arises whether the insight obtained in our experiment can be useful in a real-life attribution test. Since, especially for some of the corpora, the choice of both sets makes a true difference, truly a difference between heaven and hell (i.e. between a respectable 90% and a lowly 30%), such a test should consist of a series of several hundred iterations of various compositions of the training set, culminating with a comparison of all of the several hundred individual attributions.

Although peaks for some combinations of numbers of texts and authors may be at acceptable levels, the left slopes of the curves tend towards dangerously low values; and the wide tails of the curves show that high success rate outliers might be a stroke of luck rather than a consequence of the method, the data, or the statistical assumptions – the most ominous memento appearing here from the inexplicable dispersion in the corpus of 39 Polish novels by 8 authors (Figure 5, grey solid line). Therefore, the ideal authorship attribution situation is not only that of many texts by many authors; it is equally important to assess the validity of the training set with a very high number of trials. This might be the only way to escape the quandary of arbitrarily naming each author's 'typical' text. It seems, paradoxically, that the best way to be sure of a text's 'typicality' is to use the procedure described above to pinpoint the single text of an author that produces the best attributive results. Assuming, of course, that authors know how to write their own most typical texts.

References

Baayen, H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Baayen, H., van Halteren, H., Neijt, A. and Tweedie, F. (2002). An experiment in authorship attribution. *Proceedings of JADT 2002*. Université de Rennes, St. Malo, pp. 29–37.

Burrows, J. F. (2002). ‘Delta’: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Eder, M. (2010). Does size matter? Authorship attribution, small samples, big problem. *Digital Humanities 2010: Conference Abstracts*. King’s College London, pp. 132–35.

Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, **6** (in press).

Eder, M. and Rybicki, J. (2011). Stylometry with R. *Digital Humanities 2011: Conference Abstracts*. Stanford University, CA, pp. 308–11.

Good, P. (2006). *Resampling Methods*. Boston: Birkhäuser.

Hoover, D. L. (2004a). Testing Burrows’s delta. *Literary and Linguistic Computing*, **19**(4): 453–75.

Hoover, D. L. (2004b). Delta prime?. *Literary and Linguistic Computing*, **19**(4): 477–95.

Hoover, D. L. and Hess, S. (2009). An exercise in non-ideal authorship attribution: the mysterious Maria Ward. *Literary and Linguistic Computing*, **24**(4): 467–89.

Jockers, M. L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, **25**(2): 215–23.

Jockers, M. L., Witten, D. M. and Criddle, C. S. (2008). Reassessing authorship in the ‘Book of Mormon’ using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, **23**(4): 465–91.

Juola, P. (2009). Cross-linguistic transference of authorship attribution, or why English-only prototypes are acceptable. *Digital Humanities 2009: Conference Abstracts*. University of Maryland, College Park, MD, pp. 162–63.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**(1): 9–26.

Luyckx, K. and Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, **26**(1): 35–55.

Rudman, J. (1998). The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities*, **31**: 351–65.

Rudman, J. (2003). Cherry picking in nontraditional authorship attribution studies. *Chance*, **16**(2): 26–32.

Rybicki, J. (2009). Translation and Delta revisited: when we read translations, is it the author or the translator that we really read?. *Digital Humanities 2009: Conference Abstracts*. University of Maryland, College Park, MD, pp. 245–47.

Rybicki, J. and Eder, M. (2011). Deeper delta across genres and languages: do we really need the most frequent words?. *Literary and Linguistic Computing*, **26**(3): 315–21.

Schaalje, B., Fields, P., Roper, M. and Snow, G. L. (2011). Extended nearest shrunken centroid classification: a new method for open-set authorship attribution of texts of varying sizes. *Literary and Linguistic Computing*, **26**(1): 71–88.

Tibshirani, R., Hastie, T., Narashimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, **18**: 104–17.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. In: Lee, G. G., Yamada, A., Meng, H. and Myaeng, S.-H. (eds.) *Asia Information Retrieval Symposium 2005* (=Lecture Notes in Computer Science, vol. 3689). Berlin: Springer, pp. 174–89.