LESLI
Linguistic Evidence in Security, Law and Intelligence

Volume 3 (2019) | ISSN 2327-5596 (online) | DOI 10.5195/lesli.2019.20 | http://lesli-journal.org

Benchmarking Author Recognition Systems for Forensic Application

Hans van Halteren

Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

Abstract

This paper demonstrates how an author recognition system could be benchmarked, as a prerequisite for admission in court. The system used in the demonstration is the FEDERALES system, and the experimental data used were taken from the British National Corpus. The system was given several tasks, namely attributing a text sample to a specific text, verifying that a text sample was taken from a specific text, and verifying that a text sample was produced by a specific author. For the former two tasks, 1,099 texts with at least 10,000 words were used; for the latter 1,366 texts with known authors, which were verified against models for the 28 known authors for whom there were three or more texts. The experimental tasks were performed with different sampling methods (sequential samples or samples of concatenated random sentences), different sample sizes (1,000, 500, 250 or 125 words), varying amounts of training material (between 2 and 20 samples) and varying amounts of test material (1 or 3 samples). Under the best conditions, the system performed very well: with 7 training and 3 test samples of 1,000 words of randomly selected sentences, text attribution had an equal error rate of 0.06% and text verification an equal error rate of 1.3%; with 20 training and 3 test samples of 1,000 words of randomly selected sentences, author verification had an equal error rate of 7.5%. Under the worst conditions, with 2 training and 1 test sample of 125 words of sequential text, equal error rates for text attribution and text verification were 26.6% and 42.2%, and author verification did not perform better than chance. Furthermore, the quality degradation curves with slowly worsening conditions were not smooth, but contained steep drops. All in all, the results show the importance of having a benchmark which is as similar as possible to the actual court material for which the system is to be used, since the measured system quality differed greatly between evaluation scenarios and system degradation could not be predicted easily on the basis of the chosen scenario parameters.

Keywords*:* Author recognition, forensic linguistics, court admissability, evaluation, representativeness

Introduction

Author recognition and author profiling, i.e., attempts to deduce the identity or characteristics of the author of a text on the basis of observable properties of that text, have a venerable tradition. Originally, investigations were done by hand, as far back as the 15th century (Valla 1439/1440). But they were certainly not always ad hoc, as can be seen from the work of Wincenty Lutosławski (1890). Although manual investigation is applied even today, there appears

to be more research focus on computational methods. After the seminal work of Mosteller and Wallace (1964), much effort has been focused on identifying characteristic textual features and statistical techniques to use these features in automated recognition (see e.g., the overviews by Juola (2008) and Stamatatos (2009)).

On my first foray into the field of authorship attribution, in 1995, I was surprised that evaluation was mostly done on the basis of cases where the actual truth was not known. I proposed to apply generally accepted techniques to texts with undisputed authorship, and Harald Baayen and Fiona Tweedie agreed to my proposal (Baayen, van Halteren, & Tweedie 1996). To our surprise those accepted techniques did not perform as accurately as expected on what was a relatively simple attribution task. Since then, I have been in strong support of efforts in methodologically proper evaluation.

In a forensic context, this is obviously even more important than in scholarly investigations. Recently, various judicial systems have started to appreciate this point, and have set up guidelines for judging the admission in court of the application of forensic analysis systems. As an example, in the U.S., all Federal Courts and almost all State Courts adhere to the so-called Daubert Standard (1993), which (among other criteria) asks whether the technique in question can be and has been tested, and what its known or potential error rate is. In this paper, I show how such a test could be performed. Furthermore, I will stress the importance of controlling the conditions under which the test is performed, as the measured error rate can vary greatly with those conditions. In the following sections, I will first discuss how the benchmark was set up, then present the results, and finally conclude with a summary and discussion.

1. Experimental setup

In order to interpret the results of any evaluation, it is vital to document the data used and the steps taken in the evaluation process. In this section I intend to provide such documentation, although space does not allow me to include a full list of the text samples used. I will start with general deliberations. Next, I describe the data used, which were taken from the British National Corpus (BNC Consortium 2007). Then, although of less importance for the main thrust of the paper, I briefly describe the recognition system. I conclude the section with a detailed description of the actual tests.

1.1. Representativeness of the benchmark

In principle, an evaluation should replicate the situation (data and task) as closely as possible in the court case for which this evaluation is needed. It is my experience that this is hardly ever possible, as the necessary data are seldomly available. For the current paper, this problem is worth mentioning, but not of main importance. I can demonstrate a testing procedure (and make my points) without actual forensic material. But the reader should note that the benchmark in this paper has only limited value for argumenting admissibility. At most it could serve as a gatekeeper: if the system does not reach the desired standards for the data used here, it might not reach these standards in any specific case.

Now if actual forensic data cannot be used, what data can be? One extreme option is generated artificial data, with specific properties, as they are sometimes created when evaluating machine learning systems. This allows for a tight control of features and their distribution. However, natural language data is usually so unlike generated data that I refrained from using this option for my demonstration. Instead, I turned towards data taken from a broad language corpus, intended to be representative of the overall use of a specific language, namely the British National Corpus (BNC Consortium 2007; henceforth “BNC”).

Another important point is that authorship is only one of the factors that influence a text. Other factors include genre and topic, both of which are known to be sources of major interference in authorship studies using specific methods.

In order to test the recognition system at different levels, I used the BNC data for two different tasks. In the first one, I did not try to attribute a sample to an author, but to a text. In this way, the system's task is to recognize the triple author-genre-topic rather than just the author (assuming that these three are constant for a specific text). If a system cannot manage this, recognizing authorship by itself is hopeless. In the second task, I did actual author recognition, so that the system would have to deal with interference by genre and topic. This is more similar to what is expected of it in practice, but there was much less data, as there were not many authors with multiple texts in the corpus.

Then there is a point which is unrelated to data, but depends on the situation. In author attribution, there is a list of potential authors, from which the system has to choose one. In author verification, there is only one suggested author and the task consists of stating whether a text is by that author. Attribution is easier, as the differences between the authors in question can be identified and it is possible to focus on the more distinctive features. In the current demonstration, I benchmarked both attribution and verification. For attribution, the benchmark was restricted to choices between two authors.

1.2. Data

For the experiments in this paper, I took data from the BNC, a corpus of about 100 million words spanning most normal uses of the English language, collected in the 1990's in Britain. Given the nature of the experiments, only written texts from the BNC were used. Furthermore, in order to be able to use a sufficient number of training and test samples, I used only texts containing at least 10,000 words for the text attribution and verification experiments, allowing up to 7 training samples and 3 test samples of 1,000 words. There were 1,099 texts with these characteristics. For the author verification task, training was done for the 28 known authors for whom there were three or more texts, and testing for all 1,366 with known authors (1,213 different authors) that allowed taking 3 non-overlapping test samples.

1.3. Recognition system

For the purpose of this paper, it is not strictly necessary to know which technique is being benchmarked. Yet, I assume the reader might want to know its basics, in order to better interpret the results. The system used is the 2017 instantiation of FEDERALES (FEature DEviation RAting LEarning System), which is a further development of Linguistic Profiling (van Halteren 2004). There are two variants: FEDERALES-A does attribution, choosing between two authors; FEDERALES-V does verification. I only offer a brief explanation here, postponing a more detailed one to a later publication.

When scoring a test sample, FEDERALES first gives a penalty per feature. This is based on the z-score (distance to mean, divided by standard deviation) of the test sample measurement with regard to the statistics of the training sample(s). The penalty is tweaked with four hyperparameters, three simple ones being: the z-score, which is only counted when it is higher than t (*threshold*); then it is taken to the power dx (*deviation exponent*), and cut off at the maximum value of c (*ceiling*). The fourth one, used in FEDERALES-A, is that another power is applied, sx (*separation exponent*), being the difference of the two means divided by the sum of the standard deviations. Normally, the best values of the hyperparameters are determined in a tuning procedure, but here the values were simply chosen on the basis of past experience: t=0, c=10, for attribution dx=0.5 and sx=0.5, for verification dx=3.

The penalties for all features are added. Then, as for verification, a normalization is applied by factoring in a linear model which predicts the score on the basis of the average scores for the models and test samples in question, plus the number of feature comparisons made during scoring; the adjusted score is the deviation from the predicted value. Finally, all scores are normalized to z-scores with regard to all observed penalties for a model, in order to give some intuitive interpretation to presented scores and to make scores comparable between different author models. The latter goal is needed if we are to select a common threshold for verification.

Volume 3 2019 | DOI 10.5195/lesli.2019.20 | http:// lesli-journal.org

As for the features used in the current experiments, they were restricted to surface level, such as n-grams of words, word patterns, or parts-of-speech. In order to abstract away from topic, strongly topic-dependent words were masked. For the texts in question, this led to about 150 million features. In addition, 22 overall measures were included, referring to such matters as word and sentence length, vocabulary richness, and relative uses of various distribution ranges.

1.4. Test parameters

The evaluation experiments were executed with various settings for specific parameters. First of all, there was the sample size, which was set at 1,000, 500, 250 and 125 words. Then, there was the way of selecting samples from the texts, which could be either sequential blocks or collections of randomly selected sentences (using each sentence only once). Next, I varied the amount of training material on which the system could base its model, and the amount of test material on which it could then base its authorship decision. The test material could consist of either a single sample or three samples. The training material for text recognition could be 7, 5, 3 or 2 samples (non-overlapping, cumulative, all from the same text), and for author recognition 20, 10, 7, 4 or 2 samples (non-overlapping, cumulative, as spread out as possible over the various training texts for the author). Furthermore, the author verification tests were executed twice, in order to get an impression of the stability of the measurements.

2. Results

As discussed in the previous section, the only evaluation that is practically possible, without referring to a specific case, uses material from a broad language corpus. For a more detailed evaluation the experiments should use this material at different task difficulties. Here, I present three tasks: a) recognizing from which text samples were taken, choosing between two texts; b) verifying whether samples were taken from a specific text; and c) verifying whether samples were written by a specific author.

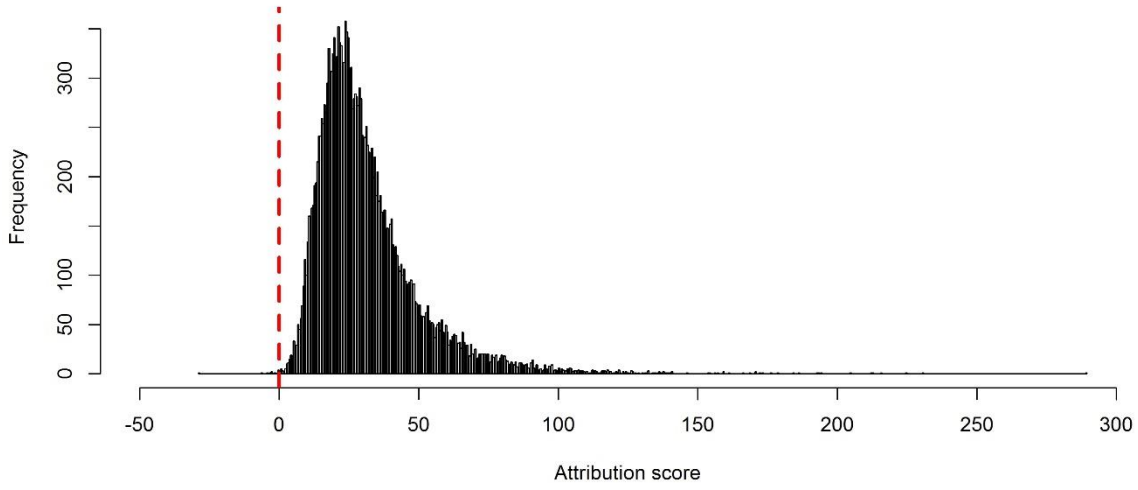

Figure 1. Attribution scores under the most favourable experimental conditions. Given 7 training samples for two texts and 3 test samples all from one of these texts, the system had to determine from which text the test samples were taken. Samples were built by randomly taking sentences totalling 1,000 words. The horizontal axis shows the

Volume 3 2019 | DOI 10.5195/lesli.2019.20 | http:// lesli-journal.org

attribution score, with a positive value indicating the correct choice. Out of 10,000 comparisons, the system erred only 6 times.

2.1. Attribution of samples to one of two texts

The first evaluation concerned the attribution of (1 or 3) test samples to a text (represented by 2 to 7 training samples). I randomly selected 10,000 pairs of texts and let FEDERALES-A assign an attribution score. A histogram of the 10,000 scores is plotted in Figure 1. A positive score (to the right of the dashed red line) means that the system chose the correct text from the two options. As the system could focus on differences, scores tended to be high, with a peak in the histogram around 25. The highest score was 289.2, for choosing between text B7B (*New Scientist*) and G2A (*Estate agents' property details*). Only 6 samples were incorrectly attributed. The highest erroneous score was 29.0, but on inspection this turned out to be between texts CSV and CSW, both sampled from the periodical *Unigram X*. Accepting such problems with the corpus-based evaluation, and assuming that this was indeed an error, the system had an error rate of 0.06% with this amount of data, which can be considered a very good result.

Table 1

Error rates in text attribution with two candidates

	7/3	5/3	3/3	2/3	7/1	5/1	3/1	2/1
Rand, 1000	0.06	0.2	0.4	1.1	0.2	0.3	0.7	2.2
Rand, 500	0.3	0.5	1.3	1.8	0.9	1.2	2.7	5.5
Rand, 250	1.4	1.8	3.9	7.0	3.2	4.1	7.1	11.5
Rand, 125	3.8	5.1	9.4	15.9	8.9	9.7	15.2	22.1
Seq, 1000	0.4	1.0	1.4	2.5	2.0	2.6	3.7	5.5
Seq, 500	1.4	2.4	3.5	5.5	4.1	5.1	7.2	10.2
Seq, 250	3.0	4.3	6.3	8.1	7.6	8.5	11.9	16.5
Seq, 125	6.0	8.1	13.5	20.6	11.8	14.3	20.6	26.6

Explanation: The values show the error rates in percents for 10,000 2-way attributions of test samples to a text, e.g. 1.3 means 130 errors were made. Columns indicate the number of training and test samples used. Rows indicate sample size and whether random sentence samples were taken or sequential blocks.

However, with lower amounts of data, the system results went down on all three size dimensions, as can be seen in Table 1. Furthermore, random (Rand) samples gave a better representation of the overall writing in a text - and

Volume 3 2019 | DOI 10.5195/lesli.2019.20 | http:// lesli-journal.org

hence better results - than sequential (Seq) samples, which rather represent local writing. Under the worst tested circumstances – that is to say, two training samples and one test sample, all of them sequential blocks of 125 words - the measured error rate was 26.6%, far lower than the 0.06% under the best circumstances.

2.2. Verification that samples belong to a text

Switching to verification, the system had a much harder task as it could not look for differences between two choices, but could only draw a comparison with one set of examples. For verification, also other evaluation measures are in order. To balance for a much lower number of author texts than non-author texts, it is measured how often a sample by the correct author was not verified (*False Reject Rate; FRR*), and how often a sample by another author was incorrectly verified (*False Accept Rate; FAR*). FAR and FRR are threshold-dependent, one going up when the other goes down because of choosing another threshold for verification. To get a single evaluation score, a common technique is to choose the threshold in such a way that FAR and FRR are the same, which is then called the *Equal Error Rate* (*EER*).

Table 2

Equal Error Rates in text verification

	7/3	5/3	3/3	2/3	7/1	5/1	3/1	2/1
Rand, 1000	1.3	3.5	5.0	7.4	1.6	4.0	5.6	9.6
Rand, 500	2.5	3.5	4.8	7.6	4.2	4.9	6.7	11.3
Rand, 250	7.2	9.3	10.0	12.3	11.6	13.5	14.5	17.6
Rand, 125	20.2	32.4	30.3	34.1	25.4	41.0	38.2	40.9
Seq, 1000	3.3	4.9	5.7	7.0	6.0	7.2	8.8	10.7
Seq, 500	5.8	7.8	9.0	10.0	9.6	11.8	13.3	15.6
Seq, 250	10.6	13.5	14.1	17.3	16.2	20.1	20.8	24.0
Seq, 125	22.5	35.2	35.1	37.6	29.1	42.4	42.0	42.2

Explanation: The values show the Equal Error Rates in percents in 1,099 x 1,099 (training texts x test texts) verifications whether the test samples are taken from a text, e.g. 1.3 means that at the optimal threshold setting, there are 1.3% false accepts and 1.3% false rejects. Columns indicate the number of training and test samples used. Rows indicate sample size and whether random (Rand) sentence samples were taken or sequential (Seq) blocks.

Table 2 shows the Equal Error Rates for various parameter settings for the text verification task. Compared to the attribution task, there was the same general degradation pattern. At the best settings, the EER (1.3%) was very

Volume 3 2019 | DOI 10.5195/lesli.2019.20 | http:// lesli-journal.org

good, but not perfect. At the worst ones, the EER (42.2%) was rather poor, but not entirely without worth. All in all, the values were high enough to proceed to the author verification task.

2.3. Verification that a text is written by an author

For the final experiment, in which I benchmarked author verification, we will first look at overall quality measurements, and then investigate more closely what happened at the author level.

Volume 3 2019 | DOI 10.5195/lesli.2019.20 | http:// lesli-journal.org

Table 3

Equal Error Rates in author verification

	20/3	10/3	7/3	4/3	2/3	20/1	10/1	7/1	4/1	2/1
Rand, 1000	7.5	7.5	7.6	8.5	10.4	8.0	8.5	9.1	9.1	13.1
	7.4	7.5	7.5	7.5	10.4	7.9	8.8	8.8	9.4	12.5
Rand, 500	9.4	11.3	11.0	11.3	13.2	11.2	11.7	12.3	14.1	15.5
	9.4	10.4	9.4	10.9	11.3	11.6	11.3	12.6	12.2	16.2
Rand, 250	11.4	11.4	15.8	16.0	15.3	13.8	14.4	18.9	20.1	21.4
	11.5	10.4	17.0	14.7	19.7	14.5	15.3	19.3	19.2	24.5
Rand, 125	15.8	17.0	40.6	37.7	43.0	19.5	21.2	40.3	41.7	44.0
	15.7	18.7	29.7	32.1	44.3	19.8	21.3	34.8	39.3	42.5
Seq, 1000	9.4	11.3	11.7	11.7	12.6	12.2	13.2	14.2	13.8	16.7
	10.4	12.3	12.9	13.2	15.8	14.1	14.6	15.6	14.9	17.9
Seq, 500	11.3	11.3	13.2	13.2	17.9	15.1	15.4	16.8	18.2	20.6
	13.2	14.2	16.0	16.0	15.2	15.8	16.0	17.9	19.2	20.4
Seq, 250	17.0	16.0	19.8	21.7	28.3	19.8	20.1	24.0	27.7	31.1
	17.0	15.9	20.8	23.2	25.0	21.6	20.1	25.7	27.0	29.5
Seq, 125	17.0	18.3	27.8	42.5	51.8	23.3	23.3	32.4	45.9	50.9
	19.8	20.9	40.6	44.8	41.5	23.9	25.8	40.9	48.3	45.2

Explanation: The values show the Equal Error Rates in percents in 106 (leave-one-text-out author models for 28 authors) x 1,366 (texts by 1,213 authors) verifications (executed twice with different training samples) whether the text is by the modeled author, e.g. 7.5 means that at the optimal threshold setting, there are 7.5% false accepts and 7.5% false rejects in the first run. Columns indicate the number of training and test samples used. Rows indicate sample size and whether (Rand) random sentence samples were taken or sequential (Seq) blocks. The first and second lines in each row block are the measurements for two different sets of training samples (but the same test samples).

Volume 3 2019 | DOI 10.5195/lesli.2019.20 | http:// lesli-journal.org

2.3.1 Quality measurements

The results for author verification are shown in Table 3. Different from Table 2, each cell holds two numbers, as the experiment was conducted twice, with different training samples (but identical test samples). As could be expected, verification quality for authors (Table 3) was lower than for texts (Table 2). After all, authors should show less variation within a single text than between texts, even more so if the texts belong to different genres and/or topics. Under the best circumstances, the equal error rate was around 7.5%. For academic purposes, this is quite good, but for court use it would be better to attempt to collect more data.

As in the other experiments, there was degradation of the quality as circumstances became less favourable. Under the worst circumstances, verification was not better than random: when verifying with random scores (1,000 runs), the measured EER had a mean of 49.9% and a standard deviation of 1.4% (with 1 test sample) and 2.5% (with 3 test samples).

More worrying, as it could lead to confusing information as to the actual quality of the system, was the non-monotonous degradation when the number of training samples decreased. The noise in the measurements caused by random selection of samples from the training texts appeared to be much stronger than expected. This was also visible in the large differences between the two measurements which were sometimes found. This means that, for a proper benchmark, the measurements would have to be taken far more times than the two shown in this paper, with varying random compositions of the training data.

Also unexpected was the shape of the degradation curve. Other than for text verification, author verification at first lost quality slowly, but at the lower sample sizes it showed a steep drop around 7 training samples. It is unclear why this would be so, and a detailed investigation of this phenomenon (for this system) is outside the scope of this paper. However, for this paper, the phenomenon also proves that it is unwise to benchmark under circumstances that are dissimilar from the circumstances for actual use in case work.

2.3.2. Verification for specific authors

Looking at specific authors, it became clear that some were more recognizable than others. Figure 2 shows four histograms of scores in individual tests, all with 20 training samples and 3 test samples, all random and with 1,000 words. The best system performance at these settings was the verification of text G04 on the basis of FRF and GUG, all by David Wingrove. The test text scores are in the top left pane of Figure 2. G04 is at the red line, with a score of slightly over 4, with texts by other authors well behind. However, closer examination showed that this was an unfair test, as all three texts were from the *Chung Kuo* series and in a sense from one long text. This was not the case for the second pane, concerning verification of text FYY by Stephen Gallagher, and with scores only slightly worse than those for David Wingrove. This is how well verification can go.

The third pane shows the worst case example, texts by Angela Robertson. GXK, the sample by Robertson, was outscored by quite a few texts by other authors. Closer examination here showed interference by genre: GXK was named *Notes for my nephews and nieces*, whereas the training texts were HD7 (*Creative writing; prose*) and HD8 (*English literature papers*). This suggests that the method is sensitive to genre differences, another issue to be considered for actual use in case work.

Volume 3 2019 | DOI 10.5195/lesli.2019.20 | http:// lesli-journal.org

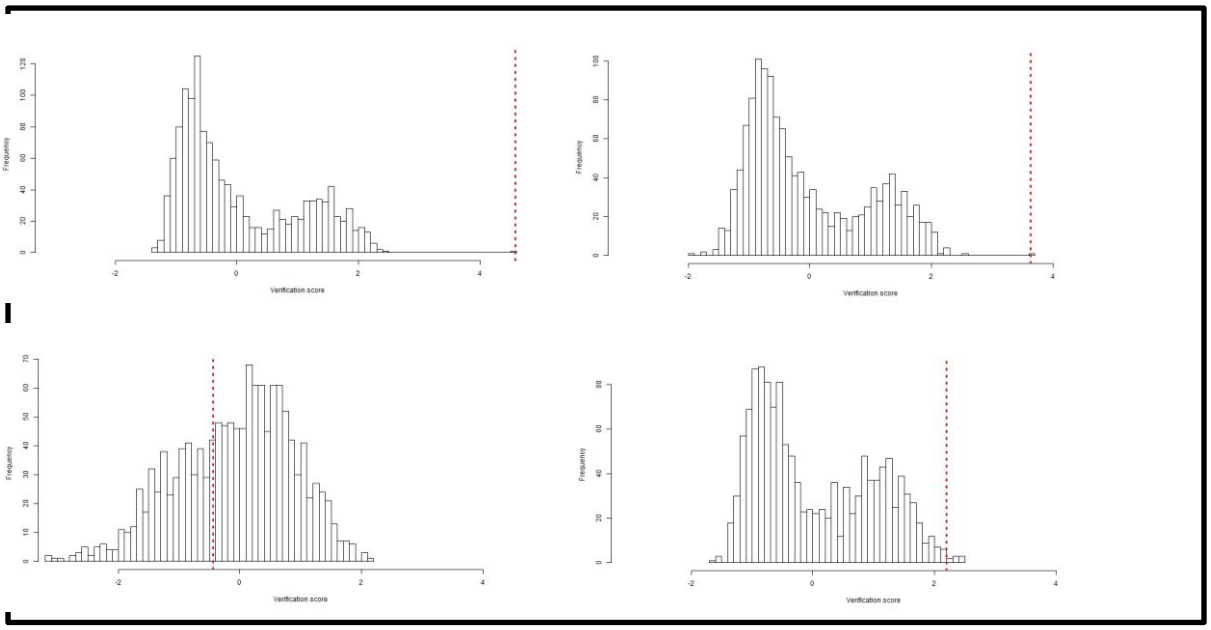

Figure 2. Verification scores for four authors. From top left to bottom right: a) David Wingrove, b) Stephen Gallagher, c) Angela Robertson, d) Iain Banks. The horizontal axis shows the verification score (higher scores indicate more similarity to the author model). The vertical axis shows how many of 1,268 test texts (excluding the training texts) receive that score. The red line indicates the score of the test text by the author in question. These results are with 1,000-word random samples, 20 training and 3 test samples.

Finally, an average example is shown in the final pane, concerning Iain Banks. Training samples were taken from FP6 (*Complicity*), G0A (*The crow road*) and HWC (*The wasp factory*); as for the test samples, they were taken from H7F (*Walking on glass*). H7F scored quite high, but there were several better scoring texts. A text with almost the same score as H7F was from another training author, Mike Ripley, namely HTL (*Angel Hunt*). In a direct comparison, attribution of H7F and HTL to either Banks or Ripley, the system managed perfectly again, with individual sample scores ranging from 3.2 to 5.7. However, in a real verification task, the luxury of training material for alternative authors is of course lacking.

Turning back to problem cases, an error analysis at the same settings showed that the EER corresponded to a threshold as low as 1.44. The low scores that forced this low threshold were found for only eight of the authors, and for all but one of these (Michael Frayn), there were special circumstances. Already mentioned was Angela Robertson, with rather varied text types. For Richard Dawkins, Grant Uden and Andrew Ashwood, two of their three samples were taken from one source, while the third one was not. For Trevor Barnes, one of the four texts is co-authored by Peter Dainty. J. King was even co-authoring all four samples with K. Bowry, and in an edited periodical (edited by T. Philpot). Finally, S. Townsend, listed as author for the four used samples from *Women's Art* magazine, was identified as the editor of that magazine rather than as an author in a fifth (unused) sample. Again we see how specific circumstances can substantially influence quality measurements. Without these problematic texts, the EER of 7.5% for the best measurement would have been much lower: with a threshold of 2.00, the FRR would be around 2.5% (still with errors because of Frayn) and the FAR around 1.5%.

3. Discussion

In several experiments I benchmarked an author recognition system, varying several aspects in order to examine the influence of changing circumstances on system quality. As expected, system quality degraded as the amount of training and test data decreased, going from very good to not significantly better than random assignment. For the system in question, the degradation curve was rather irregular, including a steep drop at specific factor values.

If such a benchmark is supposed to convince a court of admitting system output as proof on the authorship of a text, the dependence of the quality on circumstances is obviously a problem. In order to reach a good estimate of system quality in a specific case, the benchmark situation should imitate the case situation as closely as possible, a demand which is often impossible to satisfy because of a lack of sufficiently similar texts. A general benchmark with corpus data, as shown in this paper, is insufficient for providing the desired estimate. However, it could serve as a gatekeeper, assuming that systems performing unsatisfactorily on corpus data are likely to do so for case data as well.

For the FEDERALES system as presented here, this would probably mean that it should not be admitted in court, certainly for the smaller dataset sizes. With the given benchmark, the sometimes unstable measurements just cast too much doubt on any outcome on case material. Fortunately, the system was not tested at its best, and can improve with adding further features (such as syntactic patterns), hyperparameter tuning, feature selection, and a possible adjustment in score normalization which I am currently working on. Then the attribution task ought to perform sufficiently (at all but the worst circumstances). The verification task, however, should only be taken up when more data are present.

In general, if authorship recognition systems are to be respected in court, it has to be clear that they have been benchmarked rigorously, and that higher standards were required than the ones which are actually needed in practice. Especially in adversarial court systems like in the U.S., we have to make sure that any attempt to cast doubt can be rebutted easily. Such benchmarking would be substantially easier if there were datasets that are more similar to the data likely to occur in court cases. Then, the margin between the requested benchmark quality and desired court quality can be smaller. I therefore strongly encourage the forensic linguistic community to attempt to construct such a benchmark dataset. Once it is built, we can start working on a standard benchmark (data, tasks, methodology and possibly even a platform) specifically designed for forensic author recognition of various types. This would make a reliance on general corpus benchmarks as presented here superfluous, which is necessary as in the end such benchmarks are just not informative enough.

References

Baayen, F. H., Van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3), 121-132.

BNC Consortium. (2007). The British National Corpus, version 3 (BNC XML Edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Juola, P. (2008). Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3), 233-334.

Lutosławski, W. (1890). Principes de stylométrie. *Revue des études grecques*, 41, 61–81.

Mosteller, F. & Wallace, D. L. (1964) *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538-556.

Daubert v. Merrell Dow Pharm. Inc., 509 U.S. 579 (U.S. 1993).

Valla, L. (1439/1440). De falso credita et ementita Constantini Donatione declamatio. https://history.hanover.edu/texts/vallatc.html

Van Halteren, H. (2004). Linguistic profiling for author recognition and verification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (p. 199). Association for Computational Linguistics.