Natural Language Engineering **23** (5): 789–806. © Cambridge University Press 2017
doi:10.1017/S1351324917000249

Syntactic methods for topic-independent authorship attribution

JOHANNA BJÖRKLUND and NIKLAS ZECHNER
Deptartment of Computer Science, Umeå University, 901 87 Umeå, Sweden
e-mails: johanna@cs.umu.se, zechner@cs.umu.se

(*Received 8 April 2016; revised 9 June 2017; accepted 9 June 2017*)

Abstract

The efficacy of syntactic features for topic-independent authorship attribution is evaluated, taking a feature set of frequencies of words and punctuation marks as baseline. The features are 'deep' in the sense that they are derived by parsing the subject texts, in contrast to 'shallow' syntactic features for which a part-of-speech analysis is enough. The experiments are made on two corpora of online texts and one corpus of novels written around the year 1900. The classification tasks include classical closed-world authorship attribution, identification of separate texts among the works of one author, and cross-topic authorship attribution. In the first tasks, the feature sets were fairly evenly matched, but for the last task, the syntax-based feature set outperformed the baseline feature set. These results suggest that, compared to lexical features, syntactic features are more robust to changes in topic.

1 Introduction

Authorship attribution consists in identifying the author of a document, given a set of candidate authors and sample texts for each. There are also variations such as document clustering or author verification which avoid a closed-world assumption with respect to the set of authors, but those will not be considered here. Authorship attribution has a large number of applications: in authorship verification, to establish whether a text was written by a certain author (Koppel and Schler 2004), in plagiarism detection, to find similarities between texts (Eissen, Stein and Kulig 2007; Stein and Eissen 2007), in author profiling, to extract information about the age, education, and gender of the author of a text (Koppel, Argamon and Shimoni 2002), and finally, in the detection of stylistic inconsistencies, as may easily happen in collaborative writing (Collins *et al.* 2004; Graham, Hirst and Marthi 2005). Further uses are listed in the well-known survey by Stamatatos (2009).

A more recent use of authorship attribution is the detection of so-called 'troll armies'. These are made up of Internet users who are paid to promote political agendas under the guise of civilians expressing private opinions. As an example, the Dutch company 'Subvert and Profit' declare themselves to have a payroll of 25,000 users that vote, rate, and post as instructed. Methods for authorship attribution also

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

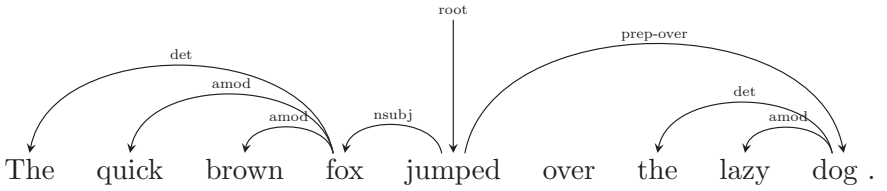


Fig. 1. The dependency analysis assigned by the Stanford parser to the sentence 'The quick brown fox jumped over the lazy dog'.

have applications to other forms of creative output, for example, source code and musical scores (Frantzeskou *et al.* 2006).

A central question is the extent to which we can recognise authors as they change topics. Previous studies by Menon and Choi (2011) suggest that function words are more effective than part-of-speech (POS) information to separate between authors when the topic varies. This can be seen as a case for lexical features over syntactic features. However, there is more to syntax than POS, so in this article, we go one step further and consider what happens when we do a full syntactic analysis and derive our features from the resulting parse trees.

Consider the dependency analysis shown in Figure 1, obtained by applying the well-known Stanford parser (Klein and Manning 2003), version 2.0.5, to the sentence 'The quick brown fox jumps over the lazy dog'. We note that the resulting structure does not represent the full stop explicitly, and that the word 'over' is captured as an edge label rather than as a vertex. There are several ways of extracting syntactic features from the analysis, and in this paper, we use the most frequently occurring relations (det, amod, nsubj and so forth). As a baseline, we take the most frequently used words and punctuation marks with respect to the training set, which overlaps with the standard sets of function words. Our results suggest that when limiting the influence of topic, syntactic features are better than lexical for recognising authors.

The paper is organised as follows. The remainder of this section covers related work. Section 2 presents our choice of features, classifiers and experimental method. Section 3 reports and discusses the results, and provides statistical analysis where needed. The concluding Section 4 summarises the results and outlines future work.

1.1 Syntactic features

Before we proceed, let us review some previous work that relates to the current effort. We begin with those that concern author attribution in general, and in the next section, look at those that address variations in topic.

One of the first to consider syntactic features in authorship attribution was Baayen, Halteren and Tweedie (1996). They were motivated by previous successes using function words, that is, common words like *the*, *of*, *and*, etc., which was argued to tap into the syntactic structure of a sentence. In their study, the classification task consisted in distinguishing between samples from a pair of books written by different authors, by means of principal component analysis with respect to either

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

the fifty most common words or syntactic features. Baayen *et al.* (1996) argued that the syntactic approach led to a better separation of samples from different authors.

Other early contributors were Stamatatos, Kokkinakis and Fakotakis (2000) who looked at stylistic features in author and topic classification. Their input consisted of online texts, and the features, for example POS tags, keywords, and overall phrase structure, were derived through automatic analysis without human intervention. Classification was done through multiple regression and discriminant analysis. In the case of topic identification, their system achieved an accuracy of eighty per cent in the case of ten topics. For authorship attribution, the accuracy was around seventy per cent, for ten authors.

When Baayen *et al.* (1996) and Stamatatos *et al.* (2000) began to introduce syntactic features, a complete syntactic analysis was not considered practical. For this reason, Luyckx and Daelemans (2005) restricted their attention to features obtainable through shallow parsing, for example, POS tags and verb forms, and used these to separate between pairs of authors. The test samples were chosen to be on the same topics for both authors and in the same genre. Luyckx and Daelemans (2005) found that lexical and syntactic features lead to comparable performances, and that the combination of the two outperformed either. Another example is the work of Argamon and Shimoni (2003) who studied combinations of simple lexical and syntactic features to infer the gender of the author of a literary text.

Improvement in natural language parsing technology led to a renewed interest in deeper syntactic features. Such an application of syntactic features is exemplified in the case of authorship attribution by Gamon (2004), who improved an existing algorithm based on shallow features by incorporating information about rewrite rules and semantic relations. Lučić and Blake (2015) continued to investigate how the local syntactic dependencies that arise when an author refers to named entities can be used for recognition.

Ayala *et al.* (2013) compared lexical and shallow syntactic features, with a slightly deeper, graph-based, form of analysis in combination with the data-mining tool Subdue (Olmos, Gonzalez and Osorio 2005), and evaluated these on different test sets covering between three and fourteen authors. On the eight datasets considered, the former method performed better in three cases, and the latter in five cases, but the authors acknowledge that the results are inconclusive and suggest further study.

The syntactic features became yet more complex in the work of Tschuggnall and Specht (2014). The authors mine so-called *pq*-grams (Augsten, Böhlen and Gamper 2005) from parse trees and use these as features. Here, p and q are a pair of natural numbers, and a *pq*-gram is a subtree whose shape is parameterised by p and q. The notion is often presented as a generalisation of n-grams to trees. Tschuggnall and Specht (2014) evaluated their system on various data sets; for example, for a data set written by four authors they achieve an accuracy rate of seventy-two per cent. Hollingsworth (2012) studied various forms of lexical and syntactic features on a corpus of novels written by three authors and had similar results across the board, where the main difference appears to be the number of features, rather than the kind.

Raghavan, Kovashka and Mooney (2010) trained probabilistic context-free grammars (PCFG) to recognise the works of target authors. This method achieves, for

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

example, a classification accuracy of seventy-eight per cent for a data set with six authors. Feng, Banerjee and Choi (2012) acknowledged the novelty of their work, but questioned how much deep syntactic features really contribute, compared to lexical productions that contain all lexical information. For this reason, they explored the usefulness of different deep syntactic features, in combination with an SVM classifier, for authorship attribution. Evaluations were made on scientific writing and classical literature. The first data set spans ten authors, each represented by at least eight scientific articles. The second data set spans five authors, this time represented by sixty documents each, created by selecting the 3,000 first words from one of the author's novels and dividing them into blocks of fifty sentences each. Feng *et al.* (2012) obtained a per-author accuracy exceeding ninety per cent even when disregarding lexical productions (and with them all words).

Fuller, Maguire and Moser (2014) combined the approaches of Raghavan *et al.* (2010) and Feng *et al.* (2012), in that they worked with PCFGs stripped of lexical information, so as to emulate traditional stylistic analysis. Fuller *et al.* argued that syntactic features are particularly interesting in the study of literature. Their method performs significantly below an SVM classifier for some of the smaller data sets tested, but as the sample size grows, the relative difference becomes smaller. In the case of a set of ten contemporary authors of suspense novels, the accuracy is eighty-three per cent for the SVM and eighty per cent for the PCFGs. In the case of a set of ten authors of classical novels, the difference disappears altogether – both classifiers give an average per-author accuracy of eighty-four per cent.

Most studies in statistical or machine-learning-based authorship attribution focus on two or a few authors. This leads to an overestimation of the importance of the features extracted from the training data and found to be discriminating for these small sets of authors. Most studies also use sizes of training data that are unrealistic for situations in which stylometry is applied, and thereby overestimate the accuracy of their approach in these situations. A more realistic interpretation of the task is as an authorship verification problem, which can be approximated by pooling data from many different authors as negative examples.

A measure of syntactic difference is developed by Wiersma, Nerbonne and Lauttamus (2011), which makes it possible to identify syntactic differences between documents in a collection. With this method, they identify under and overuse of specific constructs, which is shown to be useful for distinguishing between the English spoken by learners as compared to natives.

1.2 Topic-independent classification

Many publications on authorship attribution recognise the impact of topic and register on the results. The topic is the information that the text is meant to convey, and the register the variety of language used to express it, such as a formal or informal manner of writing. A common way to avoid the difficulties caused by changes in topic and register is to restrict experiments to a relatively homogeneous corpus. We also do that to some extent, as we only compare results within a particular register, but allow the topics to vary.

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

A promising branch of topic-independent authorship attribution originates from an article by Blei, Ng and Jordan (2003) on latent Dirichlet allocation (LDA), a generative model suited for text corpora. LDA are based on hierarchical Bayesian networks and include information about low-level features such as stylistic markers, topics and documents. The LDA was later extended by Rosen-Zvi *et al*. (2004) to include authorship information, and applied by Seroussi, Bohnert and Zukerman (2012) to authorship attribution. In doing so, they showed that by including information in the classification process about the *a priori* likelihood that an author will write on a particular topic, and when he or she does so, how this influences the style markers, the classification accuracy goes up. The difference between this approach and ours, is that we do not need a closed world assumption when it comes to topics.

Mikros and Argiri (2007) investigate how robust different types of features are to changes in topic. For this purpose, they assembled a corpus of Modern Greek newswire articles written by two authors. Using a two-way ANOVA test, they found that features such as lexical richness, variations in sentence and word lengths, character frequencies and function words correlated considerably with the topic, and must therefore be used with caution. The study does not cover syntactic features, neither shallow nor deep.

Another evaluation of different features across topics was made by Menon and Choi (2011). In their experiments, they use an SVM classifier and work with a corpus of classical novels written by fourteen authors. The features considered include n-grams of words and POS, mood words, and function words. The experiments are varied by the disjointness of topics and the likelihood of a certain topic. In all experiments, function words achieve the best results, though they sometimes fall short of the combination of all considered features.

2 Method

As mentioned in the introduction, we are interested in the robustness of deep syntactic features to changes in topic. The aim of our experiments is not novelty, but a better understanding of existing methods, and therefore our choices of data sets and classifiers tend towards the conservative.

2.1 Data sets

We work with data from three different corpora, which we refer to as Boards, Blogs and Novels; see Table 1 for an overview, and Figure 2 for samples. The Boards corpus[1] has been collected from an Irish web forum, and covers 10 years of discussions between 130,000 users. In our experiments, we used the data from 2007 and 2008, and removed users with fewer than sixty posts, leaving 5,450 authors. The Blogs corpus[2] contains the collected productions of 19,320 bloggers, all written in

[1] Collected from `boards.ie`, available at `http://data.sioc-project.org/`
[2] Collected from `blogger.com`, available at `http://u.cs.biu.ac.il/~koppel/`

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

Table 1. *Data characteristics for the Boards, Blogs, and Novels corpora*

	Boards	Blogs	Novels
Authors	5,450	19,320	20
Mean words per author	22,372	8,719	106,117
Written	1998–2008	2004	1880–1920

Doesn't that club have a reputation for being an absolute shambles of a setup?

Name and shame imo. Oh, and - so is your face. So there.

You are allowed to lie about your hand. But you cannot declare your hand while there is still action to take place. Simple really.

i didnt say it was a rule i thought was good.

huhwah? I think it's a silly rule. Read my last post again!

Hold on everyone before you get your knickers in a twist. Irrespective of the rights and wrongs of this ruling (and I agree the rule is ridiculous) the fact remains that it is in the rulebook for GJP events (and in other clubs and events around the country).

(a) Boards

Every day should be a half day. Took the afternoon off to hit the dentist, and while I was out I managed to get my oil changed, too. Remember that business with my car dealership this winter? Well, consider this the epilogue. The friendly fellas at the Valvoline Instant Oil Change on Snelling were nice enough to notice that my dipstick was broken, and the metal piece was too far down in its little dipstick tube to pull out. Looks like I'm going to need a magnet. Damn you, Kline Nissan, daaaaaaammmnnn yooouuuu....

Today I let my boss know that I've submitted my Corps application. The news has been greeted by everyone in the company with a level of enthusiasm that really floors me.

(b) Blogs

When Mary Lennox was sent to Misselthwaite Manor to live with her uncle everybody said she was the most disagreeable-looking child ever seen. It was true, too. She had a little thin face and a little thin body, thin light hair and a sour expression. Her hair was yellow, and her face was yellow because she had been born in India and had always been ill in one way or another. Her father had held a position under the English Government and had always been busy and ill himself, and her mother had been a great beauty who cared only to go to parties and amuse herself with gay people.

(c) Novels

Fig. 2. Sample extracts from each of the corpora Boards, Blogs, and Novels

2004. For lack of better information, we take each blogger to be a unique author. The Boards and Blogs corpora are similar in that (i) they are taken from contemporary web-based sources, (ii) they contain a substantial number of authors, and (iii) the amount of data for each individual author varies greatly. The usefulness of Boards and Blogs in this study is to benchmark the performance of word- and syntax-based classification in the absence of topic information.

The third corpus is different. It consists of 290 English novels, written between 1880 and 1920, by twenty-five different authors. To investigate how the accuracies

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

of the different methods depend on the amount of training data, we run our tests on samples of varying size while keeping the number of authors constant. Two equally sized data sets for the same number of authors can be different from a training perspective, if in one data set almost all data belong to one author, whereas in the other the data is divided equally over all authors. Clearly, the latter case is more beneficial for training. To avoid an in-depth analysis of the structure of the data sets, we include an equal amount of data for all authors. This means that the size of the greatest test set is limited by the size of the smallest data set for any single author. By omitting the five least productive authors, we are able to explore a larger variation in the size parameter. Both the uniformity and the amount of training data is unrealistic for applications such as forensics, but it provides insight into how classification accuracy relates to data size. A practically motivated study on authorship attribution with limited data is due to Luyckx and Daelemans (2008).

In order to simulate the effect of varying topics, we treat each novel as a separate topic. This is not perfect – some novels may of course be on the same topic, and they are in the same medium and style – but large corpora marked for topic as well as author are difficult to come by, particularly since topic is much more vague than author, so we consider this a reasonable place to start. To our advantage, separate novels are at least much more well defined than topics.

2.2 Features

With the web-based corpora in particular, it is possible to consider various paratextual data, such as posting time and markup, but we exclude those things and extract only the text itself from each corpus. We also ignore the information that could be extracted from looking at the posts separately. It is likely that average length of posts would have been a useful feature, and one might also use algorithms that count distributions of feature values over posts rather than looking at a single feature value for the entire text, but we will use neither in this study.

In order to simulate unknown identities, we divide the collected texts of each candidate in two samples (more on that in the next section). For each sample, we extract two different sets of features. The first feature set consists of word frequencies, starting with the most common words. As words, we include punctuation marks, and disregard capitalisation. Table 2 shows the ten most common words for each corpus. We have chosen not to lemmatise the words – that is, replace inflected forms with the basic dictionary form – but as can be seen in the table, most of these very common words cannot be inflected, so there would likely be little difference. We can see in Table 2 that there are many similarities, but also a few differences. In Blogs, we have a high number of *I*, as expected, since blogs often contain diary-like texts about the author. It is also worth noting that these blogs are from a time when the concept of blogs was quite new; perhaps, today they have branched out to other subjects with fewer *I*. In Novels, we see more commas than full stops. It is expected that these more formal texts have longer sentences, and the writing style of the time period may also contribute to the greater number of commas. The lack of apostrophes can be similarly explained, and more quotation marks are also

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

Table 2. *The ten most common words and punctuation marks for each corpus*

	Boards	Blogs	Novels
1	.	.	,
2	the	,	the
3	,	i	.
4	to	the	”
5	a	to	and
6	i	and	of
7	and	’	to
8	of	a	-
9	in	of	a
10	it	it	i

expected in a novel. In Boards, we might note that *a* is more prominent than in the others; perhaps, in a setting where many people write together, more new entities are introduced, which could increase the frequency of the indefinite forms.

As a second feature set, we extract syntactic patterns. We use the Stanford parser v. 2.0.5 (Klein and Manning 2003) to produce the dependency tree, and extract only the relation types. For example, in the sentence ‘The quick brown fox jumped over the lazy dog’, we get the analysis shown in Figure 1. This is decomposed into the dependencies:

det(fox-4, The-1)
amod(fox-4, brown-3)
root(ROOT-0, jumped-5)
amod(dog-9, lazy-8)
(end of sentence)
amod(fox-4, quick-2)
nsubj(jumped-5, fox-4)
det(dog-9, the-7)
prep_over(jumped-5, dog-9)

From these, we extract the number of occurrences of each relation type, in this case:

2x det (determiner)
3x amod (adjectival modifier)
1x nsubj (nominal subject)
1x root
1x advmod (adverbial modifier)
1x prep_over (the preposition ‘over’)
1x end of sentence

The parser treats some words as specific relations, such as ‘over’ in the example above, which means that some of the lexical information is also included in the syntax features, but that only applies to a small number of function words. It could also be argued that some relations, such as the determiner relation, are only used with a rather limited number of words, and therefore contains some lexical information. Other than that, there is no overlap between the two feature sets.

Note that since the parser does not give us a separate data point for words like ‘over’, nor for punctuation (other than in the form of ‘end of sentence’), we get slightly fewer data points for syntax than for words. There is also a percentage of

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

sentences which the parser fails to analyse. This explains why there is less data for syntax, as we will see in Section 3.

We consistently aim for simplicity, including in the choice of syntactic features. More complex features could provide deeper insight in tasks where syntactic information is valuable, but we leave this investigation for future work.

2.3 Classifier

For each candidate, we combine all their texts, and divide this data in two text samples, *a* and *b*, with equal amounts of data. We can think of these as the training sample and the test sample, although the setup is symmetric, so there is no systematic difference between *a* and *b*. We divide the texts chronologically; some other studies have divided the data by taking alternate sentences or posts, but in our experience, this leads to inflated and hence unrealistic results (Zechner 2015). The algorithm then compares each *a*-sample to each *b*-sample, calculating a similarity rating for each pair. For each *a*-sample, we find out which *b*-sample has the highest similarity. If it is the one which actually is by the same candidate, that is considered a successful match. The accuracy is computed as the percentage of successful matches. This way, we have a method that can answer questions like *which of these authors wrote this text*, but since we have similarity measures as an intermediate step, we can also ask *how likely are these texts to be written by the same author*.

Classification is done by compiling a feature vector for each text sample, and then applying cosine similarity between pairs of vectors. For each of the text samples, we first build a representative vector of relative frequencies. In the case of words, this means that if one per cent of the words are 'the', the feature value for 'the' will be 0.01, and so forth. As we would like to give equal weight to all features, regardless of how common they are, we then apply a standardisation step. This consists in iterating over every feature and adjusting its distribution so that the average value is 0 and the variance is 1. For each feature, we set the feature value x to

$$x_{\text{new}} = \frac{x - \mu}{\sigma},$$

where μ is the average value for this feature, and σ is the standard deviation, as computed over all training samples.

The training and test sets are standardised separately. This is reasonable since the distribution of text between different authors is the same in both sets. In an applied setting, where nothing is known about the distribution of the test set, there are other, better options. One option is to use only the training set as the basis of standardisation, while still applying it to both sets. Another is to base it on the combination of both sets. Since the standardisation is not affected by class labels, this will not skew the test results. It is also possible to include a large corpus of unannotated text purely for the standardisation step, taking care to ensure that this auxiliary data is in the same format and context as the training and test data.

The cosine similarity between vectors u and v is computed as

$$\frac{u \cdot v}{|u|\,|v|},$$

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

where $\cdot$ is the inner product between u and v, and $|u|$ and $|v|$ are the lengths of u and v, respectively. As the computations are straightforward, we do not make use of any specialised machine-learning software. Having calculated the similarity between the pairs, we match up each a-sample with the b-sample having the highest similarity; we treat this as a prediction that those samples are from the same candidate. For each experiment, we report the percentage of successful matches as the accuracy.

To statistically test whether there is a significant difference between the two feature sets, we use the non-parametric Wilcoxon signed-rank test (Wilcoxon 1945). The required assumptions are that data is paired and from the same population, that the pairs are chosen randomly and independently, and that data is measured on an ordinal scale. It does not require any assumptions about the data being normally distributed. We apply the test to the interval on the x-axis from 10,000 up to the last point which is available in both curves. We check whether the difference is statistically significant for $p < 0.0001$.

3 Results

In this section, we compare the classification accuracy obtained by working with words and syntactic features, respectively. We begin by considering the Boards and Blogs corpora, and then continue to the Novels corpus.

3.1 Boards and Blogs

Figures 3 and 4 show the accuracy for both methods, applied to the Boards and Blogs corpora and given as a function of the number of words in each sample. We see that the word-based classifier performs better than the syntax-based classifier on both corpora. The results are statistically significant, as defined above. This is in line with the previous body of work, see for example Fuller *et al.* (2014), in which purely syntactic approaches perform worse than lexical ones.

One explanation for why authorship attribution appears to be an easier problem for Boards than for Blogs is that the texts are less formal and more personal, whereas the authors have made greater efforts in Blogs to write in a literary style. According to Stamatatos (2009), this should be less of an advantage for the syntactic classifier, which has difficulties with slang and loose grammar. However, it is not clear from our results whether the difference between the two methods increases when we go from Blogs to Boards.

Note that we have limited the feature sets used here to the twenty most common features. One reason is to avoid obviously topic-dependent words, which might show an inflated accuracy. Another is improving the speed of the analysis, which could be a concern in many applications. A larger numbers of features would increase the accuracy (Zechner 2015), but that is not the focus of this study.

Another remark concerns the shape of the curves. These would have been smoother if we had run several experiments for each data size and taken the average, but then the greatest data size investigated could not have been as large. For the statistical analysis, the rough shape of the curves is not problematic.

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

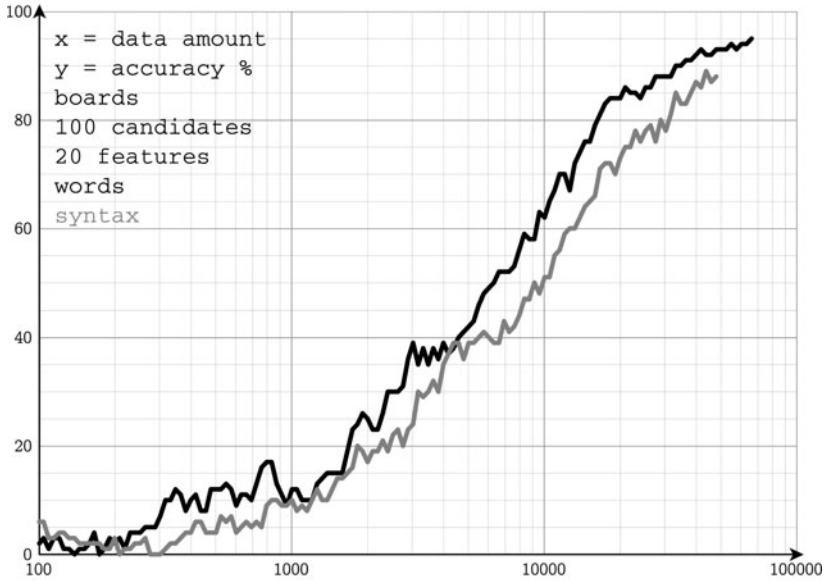


Fig. 3. The accuracy of word-based and syntax-based classifiers, applied to the Boards corpus, as a function of amount of data. The difference is statistically significant, as detailed in Section 2.3.

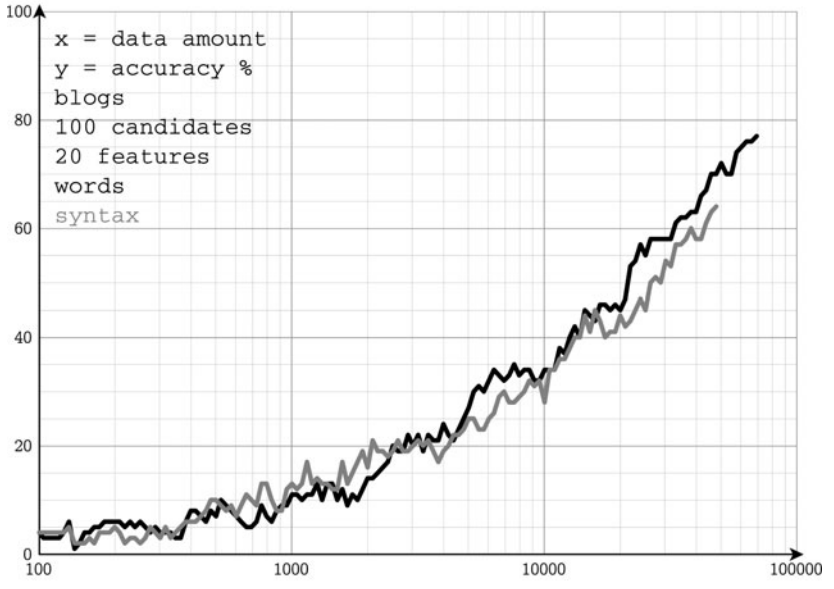


Fig. 4. The accuracy of word-based and syntax-based classifiers, applied to the Blogs corpus, as a function of amount of data. The difference is statistically significant, under the same conditions as the previous experiment.

3.2 Novels

As previously mentioned, we make use of the separate books in Novels to simulate topic identification. In the first case, we choose a data set consisting of one book per

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

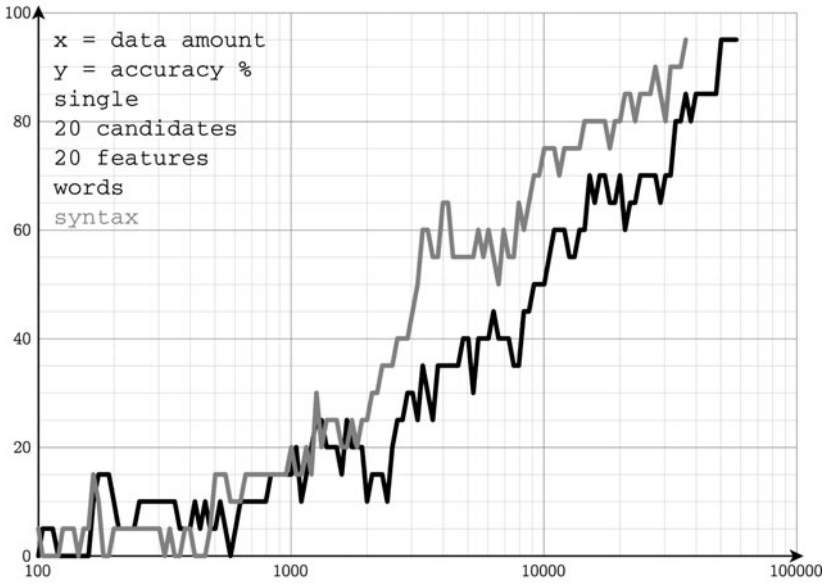


Fig. 5. The accuracy of word-based and syntax-based classifiers as a function of amount of data, with a subset of the corpus containing exactly one book per author. The difference is statistically significant.

author. This simulates the common situation in authorship attribution studies, where we are effectively identifying author and topic at the same time. In the second case, we compare only books by the same author, which means we have isolated topic differences, and avoided author differences. Finally, in the third case, we attempt to identify the author of some text from one or more novels by comparing it only with text from other novels, thus isolating author differences instead. If we can see that some methods work better in the last case than the previous one, it suggests that that method is more author dependent than topic dependent.

We first look at the case where we pick one book per author, extract two text samples from each of those books, and then try to match samples from the same book. We thus classify with respect to author and topic at the same time, which should be easy compared to the later tasks. The outcome of this experiment is shown in Figure 5 and suggests that the syntax-based classifier outperforms the word-based one. The difference is statistically significant. This differs from the other two corpora; it is not clear if this is due to the nature of the text, or coincidence.

We continue by looking at the case where we have a single author, and try to identify individual books. This means that we try to identify topics without help from author idiosyncrasies. The results are shown in Figure 6. The accuracy is now lower, which is to be expected since the data is more homogeneous. Words and syntax seem to be about equally effective, the difference is not statistically significant. Comparing with Figure 5, this suggests that syntax responds more to author differences, and words comparatively more to topic differences.

In the final case, we try to identify the author of text from one set of books by comparing with texts from other books; in other words, we identify the author

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

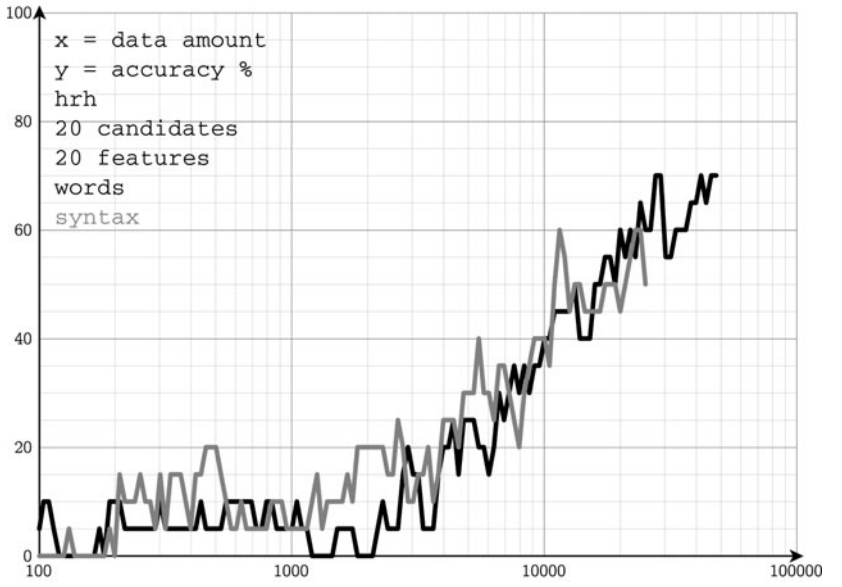


Fig. 6. The accuracy of word-based and syntax-based classifiers as a function of amount of data, when the corpus only contains books by a single author. The difference is not statistically significant.

without help from the topic. The results are given in Figure 7. Compared to the previous graphs, syntax rises considerably faster, whereas the word-based method falls behind. The difference is again statistically significant. This shows that syntax is less topic dependent than word counts, and therefore more reliable for identifying authors. Since topics are so vaguely defined, controlling for topic is virtually impossible, so this effect has bearing on any kind of author identification.

Up until now, syntax has seemed like a poor option, being computationally demanding to extract and performing worse than a simple word count, but these results show that there are situations where syntactic features give a better result.

It should be noted here that since the sample only includes twenty candidates, the expected accuracy of a random guess is five per cent, so we can also expect random fluctuations of at least that much. Looking at the first part of the graph, it would appear that the curves are both equal, but in fact all we can see is that they are both too low to get meaningful data. In the second half of the graph, the syntax curve is markedly higher.

We can see this effect from a different perspective, by going through all the books in the Novels corpus, and measuring similarities between the first and second halves of the books in terms of words and syntax. For each first half, we compare with each second half – the one from the same book, the ones from other books by the same author, and the ones from books by other authors. The corresponding plot is shown in Figure 8. The highest curves are from pairs consisting of the two halves of the same book. The curves in the middle are for pairs from different books by the same author, and the bottom curves are for pairs of texts by different authors. For each curve, the points have been rearranged in order of descending accuracy

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

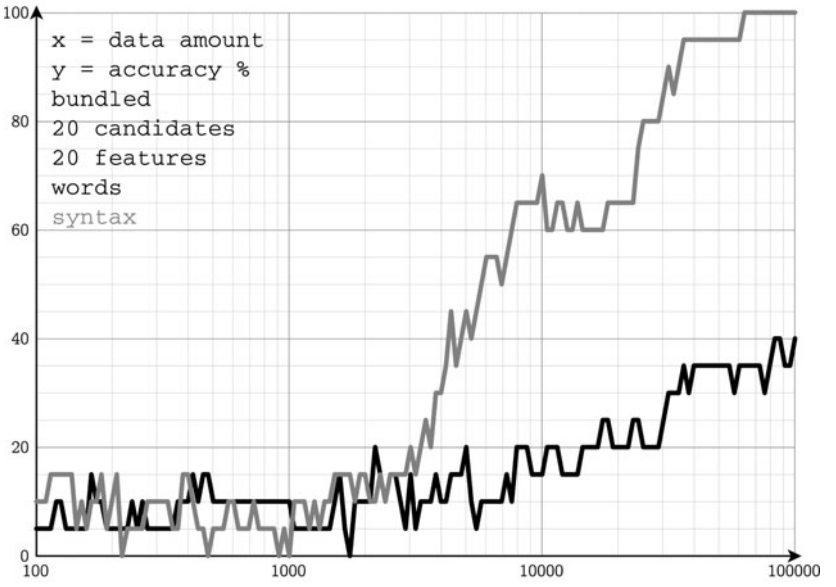


Fig. 7. The accuracy of word-based and syntax-based classifiers, when the training set and the test set both contain books from the same twenty authors, but none of the same books are present in both sets. The difference is statistically significant.

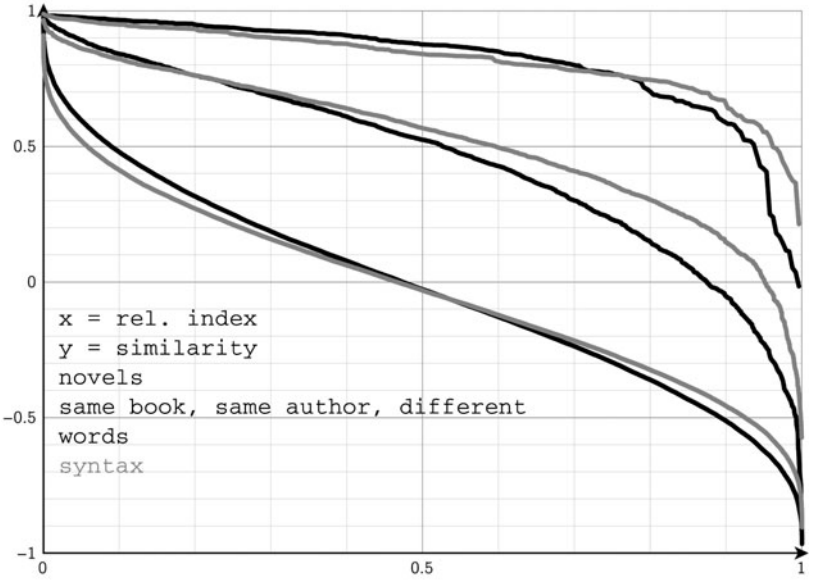


Fig. 8. The distributions of similarities for the different methods and pair types. From top to bottom: pairs from the same book, from different books by the same author, and from books by different authors.

score, and normalised so the x-axis goes from 0 to 1. This gives us a nice overview of what the similarities might look like; we see, for example, that if the similarity score is negative, we can be fairly certain the texts are not from the same book.

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

What is interesting here is the small but clearly visible difference between words and syntax: For texts from the same book, the words method mostly shows a higher similarity, but for texts from different books by the same author, syntax shows a higher similarity. This strengthens the hypothesis that a syntax-based approach is less topic dependent.

4 Conclusion

Our experiments suggest that deep syntactic features are more robust than word- and punctuation-based features against changes in topic, and therefore lead to more accurate predictions. This is in line with the intuition that the choice of topic has a greater impact on the actual words than how they are combined into sentences.

The results support the hypothesis that syntactic features are useful for topic-independent authorship attribution, but further work is needed. The corpus used in our experiments consisted of turn-of-the-century English novels, and it remains to verify how well the results generalise to other corpora and registers. Another open question is the relative performance of syntactic and lexical features as the number of authors grows. In the absence of manually annotated data, we took each novel to denote a particular topic. The soundness of this approach is open for discussion, and it would be interesting to repeat the experiments on corpora with manually annotated topic information. Moreover, the efficiency of combining lexical and syntactic features remains largely unexplored.

Future work should also evaluate other classification models, for example, Neural Networks or Support Vector Machines (possibly using kernel methods). These have a regularised optimisation model that makes them robust to irrelevant features, at the expense of a more demanding learning algorithm.

References

Argamon, S., and Shimoni, A. R. 2003. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* **17**(4): 401–12.

Augsten, N., Böhlen, M. H., and Gamper, J. 2005. Approximate matching of hierarchical data using *pq*-grams. In *Proceedings of the 31st International Conference on Very Large Data Bases*, Trondheim, Norway: Norwegian University of Science & Technology, pp. 301–12.

Ayala, D. V., Pinto, D., Gómez-Adorno, H., León, S., and Castillo, E. 2013. Lexical-syntactic and graph-based features for authorship verification notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, D. Tufis, and N. Ferro (eds.), *Working Notes for CLEF 2013 Conference, Valencia, Spain, 2013*, vol. 1179. Aachen, Germany: Sun SITE Central Europe, 1–6.

Baayen, H., van Halteren, H., and Tweedie, F. 1996. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* **11**(3): 121–32.

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**: 993–1022 (see http://www.jmlr.org/papers/v3/).

Collins, J., Kaufer, D., Vlachos, P., Butler, B., and Ishizaki, S. 2004. Detecting collaborations in text: comparing the authors' rhetorical language choices in the federalist papers. *Computers and the Humanities* **38**(1): 15–36.

Feng, S., Banerjee, R., and Choi, Y. 2012. Characterizing stylistic Elements in syntactic structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012*, Jeju Island, Korea, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1522–33.

Frantzeskou, G., Stamatatos, E., Gritzalis, S., and Katsikas, S. 2006. Effective identification of source code authors using byte-level information. In *Proceedings of the 28th International Conference on Software Engineering,* New York, NY, USA: ACM Press, pp. 893–6.

Fuller, S., Maguire, P., and Moser, P. 2014. A deep context grammatical model For authorship attribution. In N. Calzolari *et al.* (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14),* Reykjavik, Iceland: European Language Resources Association, 4488–4492.

Gamon, M. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics.

Graham, N., Hirst, G., and Marthi, B. 2005. Segmenting documents by stylistic character. *Natural Language Engineering* **3**(11): 397–415.

Hollingsworth, C. 2012. Using dependency-based annotations for authorship identification. In *Proceedings of the 15th International Conference, Brno, Czech Republic, 2012* P. Sojka, A. Horák, I. Kopeček, and K. Pala (eds.), vol. 7499, pp. 314–19. Berlin, Heidelberg: Springer.

Klein, D., and Manning, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 423–30.

Koppel, M., Argamon, S., and Shimoni, A. R. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* **4**(17): 401–12.

Koppel, M., and Schler, J. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the 21st International Conference on Machine Learning,* New York, NY, USA: ACM Press, p. 62.

Lučić, A., and Blake, C. L. 2015. A syntactic characterization of authorship style surrounding proper names. *Digital Scholarship in the Humanities* **30**(1): 53–70.

Luyckx, K., and Daelemans, W. 2005. Shallow text analysis and machine learning for authorship attribution. In *Computational Linguistics in the Netherlands 2004: Selected Papers from the 15th CLIN Meeting,* D. Scott and H. Uszkoreit, pp. 149–60. Utrecht, the Netherlands: Netherlands Graduate School of Linguistics.

Luyckx, K., and Daelemans, W. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 513–20.

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

Menon, R., and Choi, Y. 2011. Domain independent authorship attribution without domain adaptation. In G. Angelova, K. Bontcheva, R. Mitkov, and N. Nicolov (eds.), *Recent Advances in NLP,* pp. 309–15. Sofia, Bulgaria: Bulgarian Academy of Sciences.

Mikros, G. K., and Argiri, E. K. 2007. Investigating topic influence in authorship attribution. In B. Stein, M. Koppel, and E. Stamatatos (eds.), *Proceedings of the International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection,* vol. 276. Aachen, Germany: Sun SITE Central Europe, 29–35.

Olmos, I., Gonzalez, J. A., and Osorio, M. 2005. Subgraph isomorphism detection using a code based representation. In *FLAIRS Conference*, Miami, FL, USA: The Florida Artificial Intelligence Research Society, pp. 474–9.

Raghavan, S., Kovashka, A., and Mooney, R. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference (Short Papers),* Uppsala, Sweden, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 38–42.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, United States: AUAI Press, pp. 487–94.

Seroussi, Y., Bohnert, F., and Zukerman, I. 2012. Authorship attribution with author-aware topic models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 264–9.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* **60**(3): 538–56.

Stamatatos, E., Kokkinakis, G., and Fakotakis, N. 2000. Automatic text categorization in terms of genre and author. *Computional Linguistics* **26**(4): 471–95.

Stein, B., and zu Eissen, S. M. 2007. Intrinsic plagiarism analysis with meta learning. In *Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection*, Aachen, Germany: Sun SITE Central Europe, pp. 45–50.

Tschuggnall, M., and Specht, G. 2014. Enhancing authorship attribution by utilizing syntax tree profiles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 195–9.

Wiersma, W., Nerbonne, J., and Lauttamus, T. 2011. Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing* **26**(1): 107–24.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* **1**(6): 80–3.

Zechner, N. 2015. *Formal Foundations of Authorship Attribution.* Licentiate Thesis, Umeå, Sweden: Umeå University

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press

zu Eissen, S. M., Stein, B., and Kulig, M. 2007. Plagiarism detection without reference collections. In R. Decker and H.-J. Lenz (eds.), *Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, 2006* Springer, Berlin, Germany, 359–366.

https://doi.org/10.1017/S1351324917000249 Published online by Cambridge University Press