The effect of author set size and data size in authorship attribution

Kim Luyckx and Walter Daelemans
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

Abstract

Applications of authorship attribution 'in the wild' [Koppel, M., Schler, J., and Argamon, S. (2010). Authorship attribution in the wild. *Language Resources and Evaluation*. Advanced Access published January 12, 2010:10.1007/s10579-009-9111-2], for instance in social networks, will likely involve large sets of candidate authors and only limited data per author. In this article, we present the results of a systematic study of two important parameters in supervised machine learning that significantly affect performance in computational authorship attribution: (1) the number of candidate authors (i.e. the number of classes to be learned), and (2) the amount of training data available per candidate author (i.e. the size of the training data). We also investigate the robustness of different types of lexical and linguistic features to the effects of author set size and data size. The approach we take is an operationalization of the standard text categorization model, using memory-based learning for discriminating between the candidate authors. We performed authorship attribution experiments on a set of three benchmark corpora in which the influence of topic could be controlled. The short text fragments of e-mail length present the approach with a true challenge. Results show that, as expected, authorship attribution accuracy deteriorates as the number of candidate authors increases and size of training data decreases, although the machine learning approach continues performing significantly above chance. Some feature types (most notably character *n*-grams) are robust to changes in author set size and data size, but no robust individual features emerge.

Correspondence:
Kim Luyckx,
Universiteit Antwerpen,
Prinsstraat 13 (L 205),
B-2000 Antwerp,
Belgium.
E-mail:
kim.luyckx@ua.ac.be

1 Introduction

Authorship attribution has benefited from increased attention over the past decade, in both computational linguistics and digital humanities. The dominating approach in computational linguistics consists of a combination of text analysis for extracting document features that are predictive of the author, and text categorization using Machine Learning (ML) techniques. However, only limited attention has been paid to the fundamental issues in this approach. In a recent survey article on modern authorship attribution, Stamatatos (2009) collects several crucial open issues. One of these is data size. Whereas there is a consensus in ML-based research that 'There is no data like more data' (Moore, 2001), the *minimum* size requirements of a training text for authorship attribution have not been set. A second issue is the fact that the accuracy of any approach to authorship attribution also depends on the number of candidate authors. When applied to authorship attribution 'in the wild'

Literary and Linguistic Computing, Vol. 26, No. 1, 2011.
doi:10.1093/llc/fqq013 Advance Access published on 16 August 2010

(Koppel *et al.*, 2010), cases involving many candidate authors and limited data (e.g. in social networks), most existing approaches will fail to perform as expected from reported results. Several methods and algorithms have been suggested for discriminating between authors, such as PCA (Baayen *et al.*, 1996), Delta (Burrows, 2002, 2007), k-NN (e.g. Zhao and Zobel, 2005; Luyckx and Daelemans, 2008a), and nearest shrunken centroids (NSC) (Jockers and Witten, 2010), but SVMs (e.g. Argamon, 2008) are currently the method of choice. However, most of these methods have only been tested on small author set sizes, a factor making it difficult to estimate their validity outside the controlled data set.

In order to evaluate an authorship attribution method thoroughly, its performance should be measured under various conditions (Stamatatos, 2009):

- training corpus size, in terms of amount and length of training texts;
- test corpus size, in terms of text length;
- number of candidate authors;
- distribution of the training corpus over the authors (balanced or imbalanced).

Without addressing these issues, it is impossible to claim superiority of any type of features or any type of ML algorithm for authorship attribution. Equally essential are objective evaluation criteria and the comparison of different methods on the same benchmark corpora (Stamatatos, 2009), as is common practice in text categorization.

We consider writing style to be a reflection of the various interrelated aspects that characterize an individual. Among these aspects are gender, age, personality, education level, etc. Authorship attribution is thus conceived as an attempt to model individual style, whereas gender prediction, for instance, models an abstraction from that individual. Apart from individual style, various other factors determine variation in text, such as topic, genre, register, and domain. Ultimately, authorship attribution techniques should be sufficiently robust to discriminate between these interacting sources of variation. That said, keeping a maximum of these interfering factors constant, is a good strategy for finding reliable indicators of style, considering the current state of the art. We see topic as the most important variable interacting with authorship, which is why we report on experiments with single-topic and multi-topic data sets.

In this article, we present a systematic study of the effect of author set size and data size on performance and feature selection in a categorization approach to authorship attribution using Memory-Based learning (MBL). The short text fragments used for training and testing are a challenge to any approach to authorship attribution. We compare the behaviour of MBL and the predictive strength of different types of features using various (sizes of) author sets and varying data sizes in three balanced data sets that contain written texts in two languages. To our knowledge, this is the first systematic study of these aspects of authorship attribution on more than one data set.

This article is organized as follows. We begin by explaining why size has an effect on performance and feature selection in authorship attribution in Section 2. In Section 3, we provide a detailed description of the text categorization methodology underlying the experiments. In Section 4, we introduce the three data sets. The core of this study is in Sections 5 and 6, where we zoom in on the effect of author set size and the effect of data size, respectively. These sections are organized in a similar way. First, the experimental set-up is introduced, and then we present the results and discuss their implications. Finally, we formulate our conclusions and describe ongoing and future research in Section 7.

2 The Issue of Size in Authorship Attribution

The central question in authorship attribution is *Which of the candidate authors wrote the text under investigation?* ML-based authorship attribution starts from a set of training documents (documents with known authorship), extracts cues that are informative for the author, and trains a ML method that uses these features to identify the author of new, previously unseen, documents. The field of authorship attribution originates from a tradition of close reading by human experts investigating

disputed authorship in literary work like the œuvre attributed to Shakespeare. This line of research therefore tends to focus on small sets of authors and relatively large amounts of data per author. In this section, we introduce two important issues in the task that influence performance and feature selection: the number of candidate authors—the *author set size*—and the amount of data per candidate author—the *data size*. We present a survey of the studies where either the efffect of author set size and data size is examined or where a larger number of authors or smaller set of data than typical is used.

2.1 Author set size: the number of candidate authors

Trying to classify an unseen text as being written by one out of two or a few candidate authors is a relatively simple task that in most cases can be solved with high reliability and accuracies over 95%. An early statistical study by Mosteller and Wallace (1964) adopted distributions of function words as a discriminating feature to settle the disputed authorship of the Federalist Papers between three candidate authors (namely, Alexander Hamilton, James Madison, and John Jay). The advantage of distributions of function words and syntactic features is that they are not under the author's conscious control, and therefore provide good clues for authorship (Holmes, 1994). Frequencies of rewrite rules (Baayen *et al.*, 1996), *n*-grams of syntactic labels from partial parsing (Hirst and Feiguina, 2007), *n*-grams of parts-of-speech (Diederich *et al.*, 2000), function words (Miranda García and Calle Martín, 2007), and functional lexical features (Argamon *et al.*, 2007) have all been claimed to be reliable markers of style.

However, there is a difference between claims about *types* of features and claims about individual features of that type. For example, it may be correct to claim that distributions of function words are important markers of author identity, but the distribution of a particular function word, while useful to distinguish between one particular pair of authors, may be irrelevant when comparing another pair of authors.

Taking into account a larger set of authors brings us closer to the authorship verification task, where a model of individual style is built that, in the ideal case, is able to distinguish the author from the many other potential authors of a text. Authorship attribution is a simplification that allows us to zoom in on predictive features and methods.

Only recently, research has started to focus on authorship attribution on larger sets of authors. In a recent article, Koppel *et al.* (2010) found that performance decreases when the approach is applied to a large set (this study involved *thousands*) of candidate authors. However, this study would have benefited from broadening the scope to more than one data set and feature type. Argamon *et al.* (2003b) report on experiments on a set of Usenet posts, including the two, five, and twenty most active authors. Increasing author set size from two to twenty leads to a performance drop of 40%. Taking a quantitative rather than a computational perspective, Grieve (2007) also shows a significant decrease in performance when increasing author set size from two to forty authors. Abbasi and Chen (2008) investigate the effect of author set size in order to improve scalability of authorship attribution across authors. Their Writeprint system achieves a remarkable performance of 83% accuracy in 100-way authorship attribution, as a result of a rich feature set of several thousands of features. Finally, Madigan *et al.* (2005) focus on a data set of 114 authors released by Reuters, each represented by a minimum of 200 texts. The data sizes used in the Madigan *et al.* (2005) study are very different from the ones in this study, since we have available—depending on the data set—between 1,400 and 9,000 words per author.

In this article, we measure the influence of author set size in three evaluation data sets—in two languages—of which one is single topic and two are multi-topic. This selection allows us to investigate the effect in data sets of different dimensions in terms of author set size, data size, and topics.

2.2 Data size: the amount of data per candidate author

A second problem in traditional studies are the large sizes of training data that also make the task

considerably easier. The effect of data size has not been researched in much detail yet, since most stylometry research tends to focus on long texts per author or multiple short texts. Traditionally, 10,000 words per author is regarded to be 'a reliable minimum for an authorial set' (Burrows, 2007). When no long texts are available, for example in poems (Coyotl-Morales *et al.*, 2006) or student essays (van Halteren *et al.*, 2005), often a large number of short texts per author is selected for training. Some studies have shown promising results with short texts (Sanderson and Guenter, 2006; Hirst and Feiguina, 2007), but the minimum requirements for a text have not been set. When trying to set the minimum requirements for an authorial set, we consider it crucial to take into account aspects such as the domain, genre, number of topics, and the number of authors in the data set.

In cases with only limited data available, figures reported on by studies training their approach on more than 10,000 words of training data per author, cannot be relied on. In forensic applications, where researchers often need to deal with a single short text per candidate author, authorship attribution will be less reliable than expected from reported results.

A number of studies focus explicitly on data size in authorship attribution. Hirst and Feiguina (2007) present a study on authorship attribution of short texts in works by Anne and Charlotte Brontë. They find that using multiple short texts overcomes part of the obstacle of having only short texts, even when 'short' means only 200 words per author. Stamatatos (2007) investigated the class imbalance problem and tests several methods for compensation of imbalanced data sets. He concludes that the best method uses many short text samples for minority classes and less but longer ones for the majority classes. Sanderson and Guenter (2006) observed that the amount of training material has more influence on performance than the amount of test material. In order to obtain reliable performance, they find that 5,000 words in training can be considered a minimum requirement.

3 Methodology

Following the model of Stamatatos *et al.* (2000), we approach authorship attribution as an automatic text categorization task that labels documents according to a set of predefined categories. In this section, we describe how the text categorization model is applied to the authorship attribution task.

3.1 Text categorization model

Automatic text categorization (Sebastiani, 2002) labels documents according to a set of predefined categories. Most text categorization systems use a two-stage approach in which features are extracted that have high predictive value for the categories, after which a ML algorithm is trained to categorize new documents by using the features selected in the first stage, and tested on previously unseen data. Figure 1 shows a visualization of the model.

Stamatatos *et al.* (2000) translated and applied this text categorization methodology to the authorship attribution task. The model starts from a set of

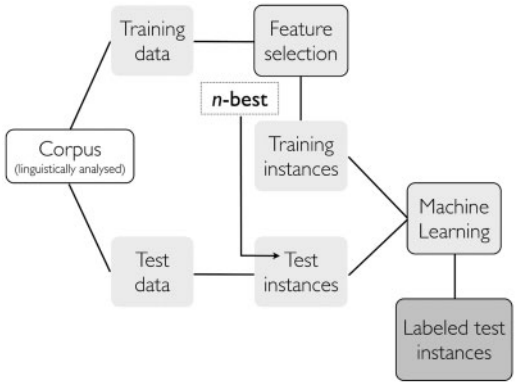

Fig. 1 Visualization of the text categorization model we apply to the authorship attribution task. Starting from a linguistically analysed data set, the data is separated in train and test. In a first stage, predictive features are extracted from the linguistically analysed training data, after which training and test instances are created based on these features. In the second stage, a ML model is generated from the training data, in order to be tested on unseen test data

documents of which the author is known (the so-called *training data*), automatically extracts features that are informative for the identity of the author, and trains a ML method that uses these features to do authorship attribution for previously unseen documents with unknown authorship (the *test data*). This approach has not only been applied in authorship attribution (e.g. Gamon, 2004; Houvardas and Stamatatos, 2006; van Halteren, 2007; Luyckx and Daelemans, 2008a), but also in authorship verification (Argamon *et al.*, 2003a; Koppel and Schler, 2004; Koppel *et al.*, 2007; Luyckx and Daelemans, 2008a), gender prediction (Koppel *et al.*, 2003), and personality prediction (Mairesse *et al.*, 2007; Nowson and Oberlander, 2007; Luyckx and Daelemans, 2008b).

3.2 Automatic linguistic analysis

In a pre-processing stage, the data is subject to automatic linguistic analysis. The data set is converted into UTF-8 format for easy processing, and then it is sent to a parser—a system that performs automatic linguistic analysis. To allow the selection of linguistic features rather than only (*n*-grams of) terms, robust, and accurate text analysis tools such as lemmatisers, part of speech taggers, chunkers etc., are needed.

The Memory-Based Shallow Parser (MBSP) (Daelemans and van den Bosch, 2005) returns a partial parse of the input text, consequently enabling the extraction of reliable linguistic features. Table 1 shows MBSP sample output for English and Dutch. MBSP tokenises the input, performs a part-of-speech analysis, looks for noun phrase, verb phrase, and other phrase chunks, and detects subject and object of the sentence and a number of other grammatical relations. This shallow parser exists for both English and Dutch.

3.3 Feature engineering

Four main types of features useful for authorship attribution research can be distinguished: lexical, character, syntactic, and semantic features. We report on experiments using the first three types of features. The features we use are listed in Table 2.

We implemented a number of basic lexical features indicating vocabulary richness, like type–token ratio—indicating the ratio between the number of unique words and the total number of words in a text—the Flesch-Kincaid metric indicating the readability of a text, and average word and sentence length. Most of these features are considered unreliable when used by themselves. We use them to complement the more complex features. Word *n*-grams and *n*-grams of lemmata are also part of this study. Character features—more specifically character *n*-grams—have been proven useful for Language Identification (Cavnar and Trenkle, 1994), Topic Detection (Clement and Sharp, 2003) and Authorship Attribution (Keselj *et al.*, 2003; Grieve, 2007; Hirst and Feiguina, 2007). We test them here for authorship attribution with varying author set size and data size. Syntactic features have been proposed as more reliable style markers than, for example, lexical features since they are not under the conscious control of the author (Baayen *et al.*, 1996; Argamon *et al.*, 2007). Part-of-speech (or PoS) *n*-grams are implemented in two ways: as fine-grained and as coarse-grained PoS. Fine-grained PoS tags provide more detailed

Table 1 Samples of MBSP output for English and Dutch

English	Dutch
The/DT/I-NP/O/NP-SBJ-1/the	De/De/LID(bep,stan,rest)/B-NP/I-SU
cat/NN/I-NP/O/NP-SBJ-1/cat	kat/kat/N(soort,ev,basis,zijd,stan)/I-NP/I-SU
jumped/VBD/I-VP/O/VP-1/jump	sprong/springen/WW(pv,verl,ev)/B-VP/I-HD
on/IN/I-PP/B-PNP/O/on	op/op/VZ(init)/B-PP/I-LD
the/DT/I-NP/I-PNP/O/the	de/de/LID(bep,stan,rest)/B-NP/I-LD
table/NN/I-NP/I-PNP/O/table	tafel/tafel/N(soort,ev,basis,zijd,stan)/I-NP/I-LD
././O/O/O/.	././LET()/O/O

Table 2 Features used in this study

Code	Feature	Type
tok	Type-token ratio V/N	Lexical
	Avg. word length	
	Avg. sentence length	
	Readability	
cwd	Content words e.g. *cat, jump, table*	
fwd	Function words e.g. *the, on*	
lex	Word n-grams e.g. *the cat, the cat jumped, cat jumped on* (lex3)	
lem	*n*-grams of lemmata e.g. *the cat jump (lex3)*	
chr	Character *n*-grams e.g. *t, th, the, he, e c, ca, cat* (chr3)	Character
cgp	Coarse-grained PoS *n*-grams e.g. *LID-N-WW, N-WW-VZ* (cgp3)	Syntactic
pos	Fine-grained PoS *n*-grams e.g. *LID(bep,stan,rest)-N(soort,ev,basis,zijd,stan)-WW(pv,verl,ev)* (pos3)	
chu	Chunk *n*-grams e.g. *B-NP, I-NP* (chu1)	
rel	Grammatical relations e.g. *SU-HD-LD* (rel3)	
lexpos	Concatenation of lex and pos e.g. *the_DT cat_NN jumped_VBD* (lexpos3)	

information about the subcategorization properties and morphological properties of words. Depending on the language of the data set, only coarse-grained PoS (for English) or both types (for Dutch) are available. *N*-grams of chunks and grammatical relations (e.g. subject, object, main verb) are also used as feature types. The last feature type is *lexpos*, a simple concatenation of the *lex* and *pos* feature types (e.g. *book_N*).

We also report on experiments with combinations of feature types that are successful in distinguishing between authors. Several studies have shown that combining features of several types has a positive impact on performance (Gamon, 2004; Grieve, 2007; Luyckx and Daelemans, 2008a). Providing a more heterogeneous feature set—by using lexical as well as syntactic features, for instance—gives a less restricted representation of the authorial set. We test whether this works when we increase the author set size.

In this study, the chi-squared metric is used as a feature selection method in order to identify features that are able to discriminate between the categories under investigation. Chi-squared (Equation 1) calculates the expected (E) and observed frequency (O) for every item (i) in every category (n). Ranking the chi-squared values per item allows us to select the most predictive items for the task at hand. This metric has been used in several studies in text categorization in general, and in authorship attribution specifically—a recent example is Grieve (2007).

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

The resulting numeric feature vectors represent frequencies of each of the selected features in the train or test instance, followed by the author label. All frequencies are normalized for text length. For *tok*, we calculate the metrics in the train or test instance and add the author label.

3.4 Machine learning and evaluation

In discriminative ML, a distinction is made between eager learning methods that abstract away from the training data to learn a model and apply the model to new data during testing, and lazy learning methods, that simply store training data at learning time, and use local similarity-based extrapolation during testing. It has been argued that lazy learning is at an advantage in language learning as it does not abstract from (potentially useful) low-frequency and low-typicality instances (Daelemans and van den Bosch, 2005).

For classification, we experiment with lazy supervised learning. We use MBL as implemented in TiMBL (Tilburg Memory-Based Learner) (Daelemans *et al.*, 2007), an open-source supervised software package for learning classification tasks

based on the *k-nn* algorithm with various extensions for dealing with nominal features and feature relevance weighting. MBL stores feature representations of training instances efficiently in memory without abstraction and classifies new instances by matching their feature representation to all instances in memory. From the closest instances (the 'nearest neighbours'), the class of the test item is extrapolated.

In order to be able to predict authorship in data sets with only one document per author, we split every text into ten fragments. The fragmentation is done randomly, meaning that we divide the text in ten equal-sized fragments and therefore do not try to end each fragment with a sentence boundary.

We perform ten-fold cross-validation (Weiss and Kulikowski, 1991), a common practice in ML research. Ten equally sized partitions (aka. folds) are created randomly from the data. Each partition in turn is used to test a model on. This model is trained on the remaining nine partitions. This way we ensure that there is no overlap between training and test data, and that all data is used for testing and training. The average over the ten experiments and the standard deviation from the average allow for more reliable estimation of the accuracy of a system.

Performance is evaluated by looking at standard evaluation metrics in text categorization. Accuracy is used to indicate the number of correctly classified texts. We evaluate results from *k*-fold cross-validation by counting the number of True Positives (TP) and True Negatives (TN), over all folds and experiments, and by calculating the average accuracy.

4 Data Sets

In order to investigate the influence of author set size and data size systematically, we test our hypotheses on three evaluation data sets for authorship attribution. These data sets conform to a great extent to the ideal evaluation corpus as described by Stamatatos (2009). He states that, in order to ensure that 'authorship would be the most important discriminatory factor between the texts' (Stamatatos, 2009), a good evaluation corpus should be controlled for genre and topic. The ideal corpus would also be controlled for factors like age, education level, nationality, etc. and the time period in which the texts were written. Such a corpus is ideal for discovering those features that are relevant for authorship attribution, but on the other hand they underestimate the complexity of the task, since in applications of authorship attribution, the data will not be controlled for all these factors producing confounding stylistic variation. As stated in the introduction, we consider factors like age, gender, and education level to be inseparable from the author's identity. Therefore, we only control for genre, register, and domain.

The datasets chosen contain fixed topic student essays written during the same time period (often a semester during the academic year), and by students with similar age, education level, and nationality. They all have a balanced distribution of texts over candidate authors. In Table 3, an overview of the characteristics and dimensions of the different data sets is given. Two of them are in Dutch, and the other one is in (American) English. Each author in a data set is represented by the same number of texts in the same topics and with a similar number of words per text. We use the term *topic* to refer to the topic assigned by the lecturer. This approach to topic disregards the actual outcome of the assignment. In fact, we consider the number of approaches to a given topic to be in direct proportion to the number of students in a classroom—provided that they do not cheat.

From the Ad-hoc Authorship Attribution Competition corpus or AAAC (Juola, 2004), we use problem set A. For AAAC_A, the students were asked to write four essays each on the following topics: work, the Frontier Thesis (by Frederick Jackson Turner), the American Dream, and national security. For Dutch, we selected the Dutch Authorship Benchmark corpus or ABC_NL1 (van Halteren, 2007) and the Personae Corpus (Luyckx and Daelemans, 2008b). For ABC_NL1, the students were asked to write three argumentative non-fiction texts (on Big Brother, the unification of Europe, and health

Table 3 Data sets used in this study

Data set	Language	Authors	Docs	Size in words	Topics/author	Data size per topic and author
AAAC_A	English	13	51	43,497	4	844
ABC_NL1	Dutch	8	72	72,721	9	1,017
PERSONAE	Dutch	145	145	205,277	1	1,413

risks of smoking), three descriptive non-fiction texts (on football, the (then) upcoming new millennium, and a recent book they read), and three fiction texts (a fairy tale about Little Red Riding Hood, a murder story at the university, and a chivalry romance). For PERSONAE, the students were asked to write an essay about a documentary on Artificial Life.

The variety of topics is an important aspect the data sets differ in. Using multi-topic and single-topic data sets allows us to investigate the consequences of multiple topics on performance and feature selection as well. In the original experiments with AAAC_A—in the framework of the competition—and ABC_NL1—in (van Halteren, 2007)—the authors decided to train on all-but-one topics, and to test on the remaining topic. The aim was to minimize the influence of topic on the trained model. We decided to deviate from this, in order to keep the model as blind as possible. Even when topic becomes a factor in authorship attribution, the system should be able to isolate the correct candidate author. Moreover, even with given topics, each author interprets the topic according to his or her own preferences. In the discussion of the results we will return to this matter.

As far as the amount of data per author is concerned, the three data sets allow for an interesting comparison. On the one hand, in ABC_NL1, each author is represented by more than 9,000 words, close to the traditional description of a reliable minimum (Burrows, 2007). On the other hand, AAAC_A and PERSONAE only have 3,000 and 1,400 words per author available, respectively. Both ABC_NL1 and AAAC_A contain respectively nine and four texts per author, while there is only one text per author available in PERSONAE. This results in instances representing very small—to be specific, about 100 words in average—fragments of text.

5 The Effect of Author Set Size in Authorship Attribution

5.1 Research objectives

Most studies in quantitative or ML-based authorship attribution focus on two or a few authors. We claim that this constraint makes it difficult to estimate performance with larger author set sizes. Two-way authorship attribution—i.e. with two candidate authors—is a significantly easier task than for example the ten-way scenario, with baselines of 50 and 10% respectively. In most cases, the first task can be solved with high reliability and accuracies over 95%, whereas the second task would be expected to present a bigger challenge for authorship attribution systems.

Moreover, testing an approach on small author set sizes exclusively also 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. Whereas it is possible that different *types* of features (e.g. character *n*-grams or function word distributions) are reliable for small as well as large sets of authors, the *specific* features may be very different in both conditions.

By increasing the number of authors to be predicted stepwise, we investigate the influence of author set size on performance and on the selection of feature types. Our expectation is a significant performance drop with increasing number of authors, similar to the findings presented in Koppel *et al.* (2010) and Luyckx and Daelemans (2008a). We also expect to see robustness of specific feature types to the effect of author set size. It is generally accepted in stylometry that syntactic features relate to the author's preferences more than lexical features do, which are used consciously and relate more to the topic of the text. Therefore, syntactic features might show robustness to author set size.

Nevertheless, the text classification literature suggests that character *n*-grams might be good predictors, since they have been used with success in a number of classification tasks (e.g. language identification (Dunning, 1994), authorship attribution (Keselj *et al.*, 2003), and composer classification (Juola, 2004b).

5.2 Experimental set-up

In order to answer the question *What effect does author size have on performance and on selection of features?*, we gradually increase the number of authors of whom we predict authorship. All data sets are subject to exactly the same procedures and experiments (Section 3) allowing us to draw conclusions that generalize over several data sets of different sizes.

In each fold, we train on nine fragments (i.e. 90% of the texts) per author and test on the remaining fragment (i.e. 10% of the texts) (see Section 3.2 on *k*-fold cross-validation). For PERSONAE, there is only one text per author, resulting in nine training instances and one test instance per author. An instance represents a fragment with a length of 100 words in average, about the length of an e-mail. In the other data sets, more than one text per author is available, resulting in more instances per author. Still, authorship attribution on short texts is a genuine challenge to any approach.

Since most studies in authorship attribution use up to five candidate authors, we mimicked these experiments by selecting two, three, four, or five authors randomly from our larger sets of candidate authors. This set-up allows for a good comparison. In order to get reliable estimates, we take several random selections of [two, three, four, five] authors and report on averaged scores. For the larger sets, we also repeated the experiments a number of times. The author set sizes and number of random selections for the different data sets are presented in Table 4.

We use MBL as implemented in TiMBL (Daelemans *et al.*, 2007) (Section 3.4). For all experiments, we use TiMBL with default settings for numeric features. The rationale behind using default settings is that we are not concerned here with optimal accuracy (optimization of algorithm parameters would lead to higher absolute results), but with measuring a relative effect (namely of author set size). The scores presented, are average accuracies.

5.3 Results and Discussion

Figure 2 shows the effect of author set size in authorship attribution using MBL in the three evaluation data sets. For reasons of clarity, this graph only represents part of the results. Per feature type (e.g. *lex1*, *lex2*, and *lex3*), only the one with highest score (in this case *lex1*) is shown.

Already at first sight, it seems clear that increasing the number of candidate authors leads to a significant decrease in performance. This effect is visible in all three data sets, regardless of their size, the language they are written in, and the number of topics. Nevertheless, the single-topic data set PERSONAE shows a steeper decrease in performance with increasing author set size than the multi-topic data sets. Character trigrams outperform the other feature types in the three data sets.

Results for the PERSONAE corpus are shown in Fig. 2a. In authorship attribution with two candidate authors we achieve an accuracy of about 80% with character trigrams (*chr3*). The chance baseline—the performance achieved by guessing the majority class for all test instances—is 50%, since we have an equal number of test instances for all authors. In authorship attribution with more

Table 4 Set-up for author set size experiments

Data set	Author set sizes × Number of random selections
PERSONAE	[2 × 100, 3 × 100, 4 × 100, 5 × 100, 10 × 10, 20 × 5, 50 × 2, 100, **145**]
AAAC_A	[2 × 20, 3 × 20, 4 × 10, 5 × 10, 10 × 10, **13**]
ABC_NL1	[2 × 20, 3 × 20, 4 × 10, 5 × 10, **8**]

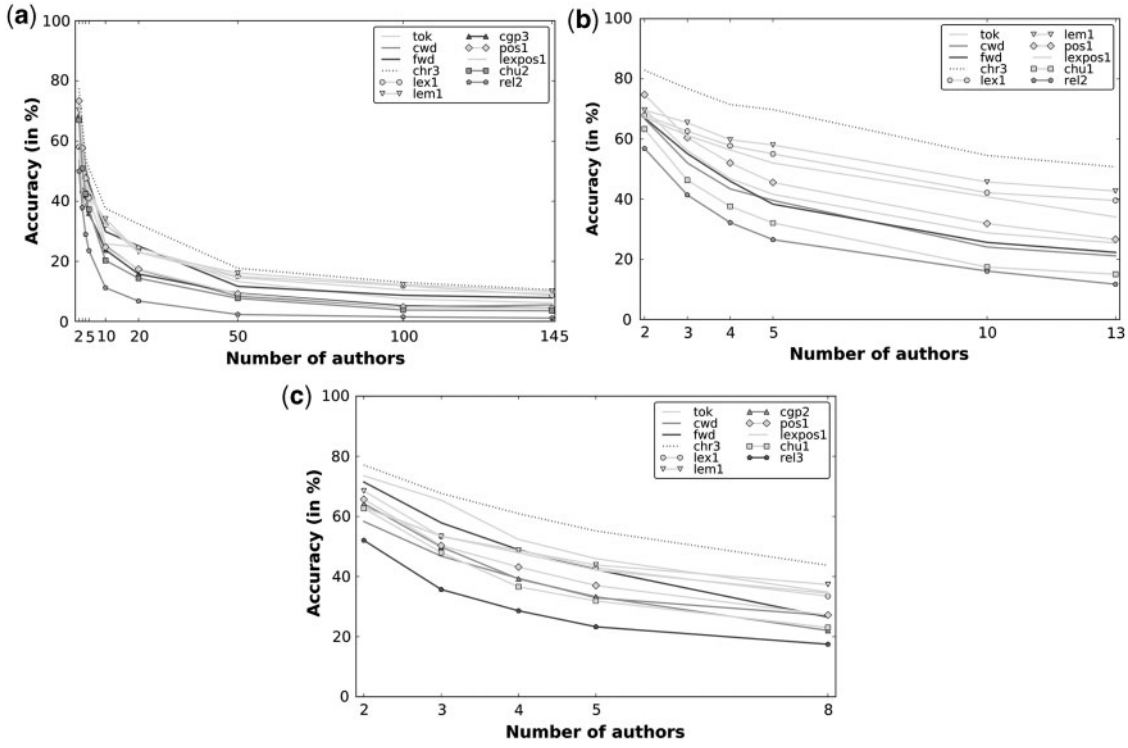

Fig. 2 The effect of author set size in authorship attribution in three data sets: (**a**) PERSONAE, (**b**) AAAC_A, (**c**) ABC_NL1

candidate authors, we see a steep decrease in performance. Five-way authorship attribution, for example, can be done with an accuracy of 51% with MBL. Increasing the number of candidate authors even more shows a similar effect. The 10-, 20-, 50-, 100-, and 145-way authorship attribution present a big challenge to the system. Authorship attribution with 145 candidate authors can be done with an accuracy of around 11%, which is still rather good, taking into account the difficulty of the task. Majority baseline performance in this task is 0.69% (1/145). In case of PERSONAE, we use around 1,260 words per author (i.e. nine fragments of 140 words) for training and 140 words for testing.

Figure 2b shows the influence of author set size in the AAAC_A data set. Whereas PERSONAE shows a steep decrease in performance when increasing the number of candidate authors, we see a less dramatic drop in the AAAC_A data set. Authorship attribution with five candidate authors still achieves a score of about 70%, while two-way authorship attribution can be done with 85% accuracy. Using the maximum number of candidate authors in this data set is possible with 51% accuracy.

Results for the ABC_NL1 corpus are presented in Fig. 2c, showing a steeper decrease in performance with increasing number of authors than AAAC_A, but less steep than PERSONAE. Scores for two- to five-way authorship attribution are overall lower than for the AAAC_A data set, and higher than for PERSONAE. Eight-way authorship attribution can be done with an accuracy of 44%, which is lower than the top score for thirteen-way authorship attribution in the AAAC_A data set. This could be an effect of the number of topics in the respective data sets—four in AAAC_A and nine in ABC_NL1—but other factors could be playing a role as well. We will discuss this further below.

When zooming in on the feature types, we see that the best scores are achieved by character

n-grams. This result can be found across the three corpora and over all author set sizes. In stylometry research, syntactic features like rewrite rules, *n*-grams of parts-of-speech, and function words have all been claimed to be reliable markers of style. Feature types of syntactic nature, such as *pos1*, and *fwd*, score well overall, but are unable to compete with character *n*-grams in AAAC_A and ABC_NL1. In PERSONAE, there is no significant difference in performance between character *n*-grams and syntactic features. In some cases, superficial lexical features like average word and sentence length or type-token ratios—implemented in *tok*—score well. Grammatical relations *rel2* hardly ever do better than majority baseline performance. The reason why character *n*-grams provide good clues to authorship could be that they capture and combine information on different linguistic levels: lexical, syntactic, and structure (Houvardas and Stamatatos, 2006).

Table 5 Author set size: the effect of providing a more heterogeneous feature set (in %) (**a**) PERSONAE, (**b**) AAAC_A and (**c**) ABC_NL1

Feature	**2 × 100**	**3 × 100**	**4 × 100**	**5 × 100**	**10 × 10**	**20 × 5**	**50 × 2**	**100**	**145**
(a)									
chr3	77.50	65.37	54.12	50.96	37.60	32.40	17.70	13.00	10.55
chr_tok	78.65	66.77	56.05	53.28	39.20	27.00	18.10	12.60	11.72
chr_fwd	78.45	67.07	55.42	53.56	42.00	25.50	18.40	12.40	12.07
chr_lex	76.80	57.17	51.23	49.08	40.20	31.40	19.80	13.10	10.48
chr_lem	78.95	54.33	50.95	48.54	40.30	25.60	20.00	13.10	10.69
chr_cgp	80.20	67.13	57.55	53.34	40.60	22.70	19.80	10.30	12.97
chr_pos	79.30	67.13	56.73	53.46	39.20	28.10	20.70	12.40	11.38
chr_lexpos	75.45	57.10	51.55	48.98	39.40	23.40	20.00	9.90	12.41
chr_chu	78.45	66.77	55.58	51.46	36.30	24.20	20.10	11.90	10.34
chr_rel	75.30	55.63	53.02	47.68	36.40	21.70	16.40	10.50	11.24
Feature	**2 × 20**	**3 × 20**	**4 × 10**	**5 × 10**	**10 × 10**	**13**			
(b)									
chr2	85.48	75.96	69.53	64.44	50.86	46.08			
chr3	82.91	76.80	71.49	69.80	54.57	50.78			
chr_tok	79.69	80.01	73.94	70.47	55.60	50.98			
chr_fwd	81.16	80.23	74.90	71.37	55.95	53.53			
chr_le ×	76.62	78.50	73.05	70.69	56.10	53.73			
chr_lem	77.50	79.92	73.49	72.08	58.72	54.90			
chr_pos	80.74	81.33	76.57	72.82	58.98	56.27			
chr_lexpos	76.14	79.36	74.41	70.54	56.79	53.33			
chr_chu	79.63	80.49	73.37	71.14	56.09	51.76			
chr_rel	78.75	75.73	68.58	65.05	53.98	43.73			
Feature	**2 × 20**	**3 × 20**	**4 × 10**	**5 × 10**	**8**				
(c)									
chr2	78.03	65.22	59.00	52.38	42.64				
chr3	77.08	67.67	60.94	55.18	43.75				
chr_tok	79.92	69.43	62.36	53.69	44.31				
chr_fwd	79.61	69.46	61.36	55.69	45.97				
chr_lex	80.53	69.65	59.97	59.04	49.86				
chr_lem	81.58	69.80	60.42	58.24	50.28				
chr_cgp	79.14	69.11	60.86	53.13	43.47				
chr_pos	77.14	68.91	60.67	52.09	40.14				
chr_lexpos	80.11	68.74	60.19	58.40	50.28				
chr_chu	78.83	66.98	61.00	52.87	43.61				
chr_rel	76.72	62.56	49.08	45.22	38.75				

Table 5 shows the effect of adding features of a different type (e.g. syntactic or lexical) to the best scoring feature type over all cases presented above, namely, character n-gram. The results show an increase in performance in most tasks and data sets. Although character *n*-grams capture nuances on different linguistic levels, adding syntactic information—such as *pos* and *cgp*—has the largest positive effect on performance in PERSONAE and AAAC_A. In ABC_NL1 (Table 5, panel c), lexical additions to the character *n*-grams seem to be the most successful. In 145-way authorship attribution, adding syntactic information scores 13% accuracy. Working with 13 candidate authors leads to a score of 56%—an increase of 6%. We see the same amount of increase in ABC_NL1. These results indicate that providing a more heterogeneous set of features improves the results significantly.

When we examine the results thoroughly, a number of interesting conclusions and issues emerge. First of all, our claim that studies focusing on a small number of authors may lead to an overestimation of the importance of extracted features on the one hand and of performance on the other hand, holds. From the results we described above, it is clear that increasing the number of authors leads to a significant decrease in performance. A system that scores an accuracy of over 80% on two-way authorship attribution will not be able to obtain similar results when tesed on for example twenty candidate authors.

We also find evidence to support the claim that providing a more heterogeneous feature set has a positive influence on performance. Adding extra information—of syntactic nature, for example—to the top scoring feature type—in this case character *n*-grams—increases the score in our three data sets. Author set size does not influence this positive effect, since it emerges regardless of the number of candidate authors.

As far as feature selection is concerned, we find that similar *types* of features tend to work well for small and large sets of authors in our corpora. Character *n*-grams outperform the other feature types in most cases. By looking at the *individual* features, we want to investigate the existence of robust features for authorship attribution.

While it may be correct to claim that distributions of function words are good clues for authorship, the distribution of a particular function word, however useful to distinguish between one particular pair of authors, may be irrelevant when comparing another pair of authors. This is a critique often heard with respect to, for instance, Chaski's work in the framework of forensic linguistics (Grant and Baker, 2001). In order to investigate this in our data, we first randomly selected two author pairs from PERSONAE and calculated the percentage of overlap between the two lists of top-100 features (ranked and selected by means of chi-squared). An overlap of 100% would indicate that the same features are selected for the two author pairs. We repeated this experiment a number of times while increasing the number of randomly selected author pairs. Figure 3 shows the results of these experiments. When comparing two author pairs, we find an overlap of almost 5%, but increasing the number of author pairs indicates a dramatic drop in overlap. This indicates that robust feature types exist, whereas robust individual features do not emerge.

Table 6 shows the a-priori distribution of the main feature types in PERSONAE on the one hand and in the top-1,000 features in 145-way authorship attribution on that data set as ranked by means of

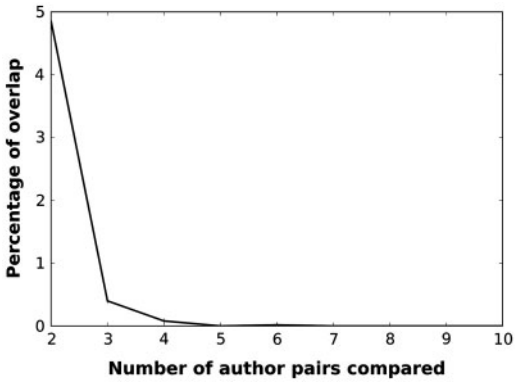

Fig. 3 Amount of overlap in *chr3* features over *n* pairs of authors. Taking the 100 *chr3* features with highest chi-squared value in *n* author pairs, overlap indicates the percentage of features that occur in all of the author pairs

Table 6 Distribution of feature types in the PERSONAE full feature set and in top-1,000 features (ranked on the basis of chi-squared values)

Feature type	Distribution	
	A-priori	Top-1000
chr	29.27%	16.10%
lex	9.46%	5.60%
cgp	7.51%	1.40%
pos	14.87%	4.00%
lem	9.94%	6.10%
rel	13.69%	59.70%
chu	5.79%	1.20%
lexpos	9.47%	5.90%

A feature is taken into account if it reaches a frequency threshold (set at 50) over all authors.

the chi-squared metric (see Section 3) on the other hand. These figures give an indication of the relative importance of the feature types. Most features are of the *rel* type, but these fail to perform well, as we indicated above. Most of these are very fine-grained and specific, and therefore unlikely to occur in test, which explains their underperformance. *Chr* features account for thirty percent of all features, and a quarter of the selected top-1000 features. Still, they outperform the other feature types. Chi-squared is a commonly used feature selection method, but it is known to be sensitive to very small expected counts (Forman, 2003), which is a typical characteristic of natural language (Zipf, 1935). It would therefore be very interesting the investigate the effect of the feature selection method (e.g. Forman, 2003).

Apart from the number of candidate authors, we find that a number of other factors impact performance, such as the variety of topics in the data set, corpus size, the choice of methodological set-up, and the choice of ML algorithm. We will elaborate on these factors here.

First of all, the results show that the type of data set has an effect on performance. For example, the influence of author set size is visible in the three data sets, but the single-topic data set seems to be affected more, while the multi-topic data sets undergo a smaller negative influence with increasing number of authors. Comparing results in five-way authorship attribution—a task the data sets have in common—leads to the conclusion that performance in AAAC_A is higher than in ABC_NL1, although the latter data set contains three times more data per candidate author (3,000 versus 9,000 words). It seems that, apart from the amount of data avalable per author, the variety of topics might also affect the results. Stamatatos (2009) indicates that the ideal corpus for authorship attribution should be single-topic. In order to mimic single-topic performance in ABC_NL1, we extracted nine single-topic data sets from the data set and ran eight-way authorship attribution experiments. Over all topics, we achieved an accuracy of 44% (Fig. 2c). In Table 7, we see that accuracies vary to a large extent, depending on the topic. The top-scoring data set achieves an accuracy of 90%, while the third topic (*T3*) scores only 26% accuracy with character trigrams. As far as feature types are concerned, *cgp2* appears to give the most consistent score over all topics with a standard deviation of 7.16% and an average score of 44.44%. Character trigrams score overall better with an average accuracy of 58%, but the variation is immense (namely a standard deviation of 24%). Figuring out the exact dynamics between the type of data set and performance of a given approach is not the focus of this study (see Mikros and Argiri (2007) for studies on the influence of topic in authorship attribution), but the results presented here indicate a role for topic.

Corpus size is another aspect of data that affects performance. On the one hand, it is generally accepted in ML that one can never have enough training data and that more data leads to an increase in performance (cf. 'There is no data like more data' (Moore, 2001)). On the other hand, studies show that, depending on the choice of Machine Learner, adding training data may lead to a plateau at some point in the learning curve. Such a learning curve demonstrates the existence of a reliable minimal set of training data leading to a good performance. The results for PERSONAE described above (Fig. 2a) indicate that authorship attribution with a small set of training data—about 1,200 words per author—is up to standards when comparing with performance on larger sets of data like ABC_NL1 (Fig. 2c). That said, topic also has an effect of performance and

Table 7 Single-topic simulation of ABC_NL1 (results in % of accuracy)

Feature	T1	T2	T3	T4	T5	T6	T7	T8	T9
tok	45.00	30.00	41.25	56.25	50.00	60.00	51.25	35.00	51.25
cwd	25.00	16.25	25.00	42.50	38.75	53.75	31.25	32.50	60.00
fwd	37.50	27.50	26.25	42.50	26.25	50.00	31.25	35.00	53.75
chr3	41.25	27.50	40.00	73.75	87.50	90.00	36.25	48.75	81.25
lex1	32.50	31.25	28.75	56.25	52.50	77.50	31.25	41.25	61.25
lem1	41.25	26.25	27.50	61.25	46.25	67.50	33.75	33.75	62.50
cgp2	30.00	28.75	23.75	46.25	32.50	28.75	31.25	25.00	40.00
pos1	20.00	35.00	25.00	40.00	37.50	57.50	23.75	31.25	51.25
lexpos1	37.50	27.50	35.00	56.25	55.00	71.25	32.50	38.75	61.25
chu1	33.75	18.75	10.00	27.50	28.75	26.25	22.50	25.00	22.50
rel3	11.25	8.75	15.00	12.50	13.75	18.75	12.50	17.50	15.00

ABC_NL1 is a corpus with texts by eight authors about nine different topics. We extracted nine single-topic data sets from this multi-topic corpus and report on results in eight-way authorship attribution using MBL.

Table 8 Comparison of AAAC and MBL results on AAAC_A and ABC_NL1 (in %)

AAAC results	Data set A (AAAC_A)	Data set M (ABC_NL1)
Baronchelli	3/13	5/24
Coburn	5/13	19/24
Halteren	9/13	21/24
Hoover	4/13	7/24
Juola	9/13	11/24
Keselj1	11/13	17/24
Keselj2	9/13	15/24
Obrien	2/13	5/24
Schler	7/13	4/24
Stamatatos	9/13	14/24
MBL	7/13 (chr3) 6/13 (lem1)	13/24 (cgp3) 12/24 (lexpos2)

unraveling the dynamics between the different factors is a research topic in itself, as we observed above.

A third factor is the choice of methodological set-up. By deviating from the AAAC competition set-up (see Section 4), it is difficult to compare our results with those of the competition. The AAAC set-up minimizes the influence of topic on the trained model by training on all-but-one topics and testing on the held-out topic. Table 8 shows the best scoring teams (Juola, 2008) and our results using the AAAC set-up with default MBL for numeric features. More information on the competing teams can be found in Juola (2008). When comparing the results, we see that MBL scores are situated in the top-half of the AAAC results (even without optimization of algorithm parameters).

A last aspect we want to highlight, is the choice of ML algorithm. Note that we used MBL experiments using TiMBL with default settings for numeric features, since our goal is to measure the effect of author set size, not that of optimization. However, we did compare MBL with a number of other ML algorithms. By using MBL, we do not mean to imply that MBL outperforms other algorithms when tested in authorship attribution. Table 9 shows performance of JRip, an implementation of the greedy Ripper algorithm (Cohen, 1995), SMO (Platt, 1998), an implementation of Support Vector Machines, Naive Bayes (John and Langley, 1995), and C4.5 (Quinlan, 1993) on the three evaluation data sets, using the maximum number of authors. This quick comparison teaches us that SMO scores the best results of all Machine Learners tested here. Interestingly, MBL seems to be the only learner that scores higher on thirteen-way than on eight-way authorship attribution.

6 The Effect of Data Size in Authorship Attribution

Now, we proceed with the second aspect of our study: the amount of training data per candidate author.

Table 9 Comparison of MBL with other ML algorithms on the three data sets with maximum number of authors and *chr3* (results in %)

Feature	TiMBL	JRip	SMO	Naive Bayes	C4.5
PERSONAE (145-way)	10.55	16.00	23.79	12.83	7.24
AAAC_A (13-way)	50.78	26.86	57.06	44.31	32.55
ABC_NL1 (8-way)	43.75	38.61	60.56	51.81	45.28

6.1 Research objectives

Most studies in authorship attribution use large amounts of data per candidate author. Distinguishing between a small set of authors based on large collections of data per author is a task that can be solved with high accuracy. However, when only limited data is available for a specific author, the authorship attribution task becomes much more difficult. By testing the system on very limited data—140 words for training, for instance—we can estimate its viability when applied to e-mails, letters, blogs, or tweets in forensic applications (e.g. fraud detection).

We present learning curve experiments in authorship attribution and expect to see an increase in performance when the system is trained on more data. As far as feature types are concerned, our expectation is that syntactic or character features are more robust to the influence of data size than lexical features. Indications for this can be found in Stamatatos (2008), where it is stated that character *n*-grams reduce the sparse data problems that arise when using word *n*-grams.

6.2 Experimental set-up

We investigate the effect of data size in the three data sets by performing authorship attribution while gradually increasing the amount of data the system selects features from and is trained on (from 10 to 90%), keeping test set size constant at 10% of the entire data set. The resulting learning curve will be used to compare performance when using different feature types on the three evaluation data sets. We present results in authorship attribution with the maximum number of authors (see Section 4 for a description of the data sets). The scores presented, are average accuracies.

6.3 Results and Discussion

The influence of data size is demonstrated by means of learning curves. Figure 4 shows that performance is positively affected by increasing the amount of data the system is trained on. This effect is prominent in the three corpora, regardless of their size or language. However, the gain appears to be higher in PERSONAE (Fig. 4a) than in ABC_NL1 (Fig. 4c). As far as feature selection is concerned, a first inspection teaches us that some feature types perform better than others with increasing data size.

Zooming in on the learning curve for PERSONAE in Fig. 4a, we see that performance increases from 3% accuracy with 10% of the data in training to around 10% with 90% of the data in training (in 145-way authorship attribution). It is worth remarking that 10% of the data equals one fragment of 140 words per candidate author, about the size of a (long) e-mail. Best results are obtained with character trigrams, and lexical features. These feature types also benefit most from the increased amount of training data.

In AAAC_A (Fig. 4b), we see similar behaviour as in PERSONAE in that the same types of features perform best. With only 10% of the data in training, thirteen-way authorship attribution achieves an accuracy of 27% with character trigrams. Using 90% of the data in training results in 50% accuracy. The increase in performance is less apparant than in PERSONAE, since the AAAC_A is a multi-topic data set (see Section 4)—a data set where every candidate author is represented by texts in multiple topics -, which means that the system is built on more data than in PERSONAE.

ABC_NL1 results are shown in Fig. 4c. This data set contains documents in nine topics per candidate author, totalling to about 9,000 words per author. The influence of data size is least visible in this data set since we only see a light increase in performance with increasing amount of training data.

When taking a close look at the feature types, we see that the best scores are achieved by character *n*-grams. This result can be found across the three corpora and over all data sizes. As with the author set size experiments, we see that character *n*-grams

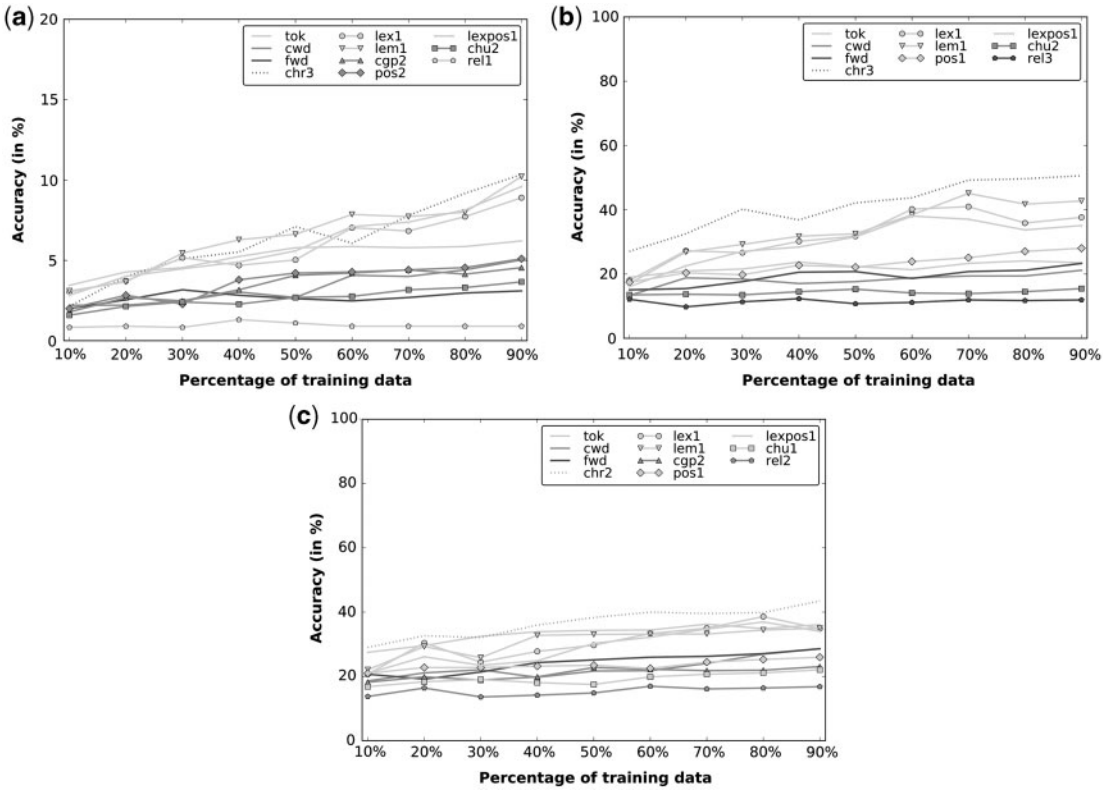

Fig. 4 The effect of data size in authorship attribution in three data sets (in %): (**a**) PERSONAE: 145 CANDIDATE AUTHORS, (**b**) AAAC_A: 13 CANDIDATE AUTHORS, (**c**) ABC_NL1: 8 CANDIDATE AUTHORS

show more robustness to the effect of data size than syntactic features. Lexical information—word and lemma *n*-grams—perform second-best.

Table 10 shows the impact of combinations of feature types in the three evaluation data sets. The PERSONAE results (Table 10, panel a) clearly indicate that providing a more heterogeneous set of features aids performance in 145-way authorshp attribution. In all sizes of training data—from 10 to 90%—this effect is present. In the other two data sets, AAAC_A (Table 10, panel b) and ABC_NL1 (see Table 10, panel c), this effect is only present when enough training data—i.e. between 70 and 90% of the data—is available. This could be a result of either the variety of topics, or the amount of data per candidate author. In most cases where this effect presents itself, providing syntactic features next to character *n*-grams is the best option.

A number of interesting conclusions and issues concerning the effect of data size emerge. Experiments in Section 5.3 have already indicated that authorship attribution can lead to reasonable results even when only limited data is available. The PERSONAE corpus consists of in average 1,400 words per author, but still 145-way authorship attribution can be done significantly above baseline with an accuracy of 10% using MBL. Of the three data sets, ABC_NL1 is the one with most data available per author, but scores with this data set (eight candidate authors) are lower than for AAAC_A (thirteen candidate authors), although the latter has three times less data per author. Bearing in mind that the train and test instances only represent texts of 100 words in length, topic seems to play a role here, since the amount of data per author is not the critical factor.

Table 10 Data size: the effect of providing a more heterogeneous feature set (in %): (a) PERSONAE: 145 CANDIDATE AUTHORS, (b) AAAC_A: 13 CANDIDATE AUTHORS and (c) ABC_NL1: 8 CANDIDATE AUTHORS

Feature	10%	20%	30%	40%	50%	60%	70%	80%	90%
Panel a									
chr2	2.69	4.48	5.38	6.28	6.14	6.55	7.10	7.24	7.66
chr3	2.14	4.00	5.10	5.52	7.10	6.07	7.79	9.17	10.34
chr_tok	2.97	4.90	6.48	6.55	7.72	7.03	8.55	10.14	12.07
chr_fwd	4.16	4.55	6.34	6.69	7.45	7.10	8.14	9.31	10.00
chr_lex	3.03	5.31	6.69	7.10	7.79	8.55	8.55	9.86	11.52
chr_lem	3.03	5.72	5.79	7.17	8.21	8.28	9.24	10.00	12.07
chr_cgp	3.31	5.17	5.31	6.00	8.00	7.03	8.83	9.45	9.79
chr_pos	2.83	4.62	5.86	7.17	9.45	8.90	10.14	11.24	12.48
chr_lexpos	3.17	5.38	7.03	7.03	7.72	8.83	9.03	10.48	11.52
chr_chu	2.83	4.90	5.79	6.55	7.17	7.52	7.79	9.45	11.45
chr_rel	3.95	4.07	5.31	5.66	6.69	6.55	8.14	8.28	10.69
Feature	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Panel b									
chr3	27.06	32.55	40.20	36.86	42.16	43.73	49.22	17.65	50.59
chr_tok	27.89	30.72	40.31	40.20	40.52	42.70	48.80	18.08	51.63
chr_fwd	26.36	28.43	37.69	33.33	39.87	37.47	40.20	18.04	52.35
chr_lex	19.61	22.35	24.94	27.65	32.94	36.08	38.24	21.57	51.57
chr_lem	20.59	23.33	27.45	29.41	35.29	34.31	34.51	25.49	55.88
chr_pos	20.15	32.75	34.12	28.43	38.43	36.86	36.67	17.45	55.69
chr_lexpos	18.52	23.33	28.82	27.25	34.31	35.49	36.27	22.94	52.55
chr_chu	26.36	28.43	31.76	35.29	40.59	40.39	43.33	18.63	52.94
chr_rel	19.80	32.94	36.86	28.43	30.98	33.53	32.55	17.65	43.53
Feature	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
Panel c									
chr2	29.03	32.64	32.08	35.97	38.33	40.00	39.58	39.86	43.47
chr3	24.58	29.72	30.56	31.81	32.22	37.08	36.81	43.33	42.36
chr_tok	28.19	32.92	33.61	36.94	37.22	39.03	40.00	44.72	42.64
chr_fwd	25.28	30.42	31.25	32.36	33.19	36.39	31.39	45.83	33.75
chr_lex	24.03	29.17	29.44	27.50	30.69	32.22	32.08	49.58	35.42
chr_lem	25.56	27.50	27.36	30.69	30.42	31.81	31.81	46.81	33.75
chr_cgp	28.19	27.36	32.64	35.56	31.67	39.86	40.14	42.92	34.58
chr_pos	23.19	28.06	28.47	31.11	31.94	26.94	29.17	42.50	30.42
chr_lexpos	25.28	28.47	28.33	27.50	30.69	29.72	31.25	46.53	33.61
chr_chu	23.33	27.92	29.44	30.83	32.08	38.47	38.89	43.47	41.81
chr_rel	23.47	27.92	30.28	32.08	33.61	35.42	37.36	35.97	35.97

By showing learning curves in three data sets, we presented a systematic study of the effect of data size. On the one hand, the results confirm the idea that 'There is no data like more data' (Moore, 2001) as far as the percentage of data in training is concerned. On the other hand, a factor like the amount of topics seems to play an important role as well. At this point, it is not possible to assess whether extracting features based on 1,400 words per candidate author from PERSONAE has similar predictive power as doing the same on 1,400 words from the other two data sets. The dynamics between the number of topics, the amount of data, and the number of candidate authors cannot be evaluated from the results presented above. Nevertheless, it is clear that the systematic analysis of the effect of data size in authorship attribution is crucial in assessing the robustness of our approach to data size.

Testing authorship attribution on small data sets (consisting of blogs, tweets, e-mails, short essays, etc.) could lead to interesting insights concerning the size of the 'minimal set' for reliable authorship

attribution. PERSONAE and AAAC_A are relatively small corpora (with 1,400 and 3,000 words per author, respectively), while ABC_NL1 contains about 9,000 words per author (see Section 4 for a description of the data sets). One of the basic assumptions underlying stylometry is the idea that stylistic choices are present in all end products of an author, on the one hand. On the other hand, short texts include less of the author's specific style preferences, hence providing a genuine challenge for most of the state-of-the-art authorship attribution approaches.

On top of that, the extent to which results can be called 'reliable' might be different when comparing for example a task such as settling disputed authorship between novels (Argamon *et al.*, 2003a) and large-scale weblog analysis (Koppel *et al.*, 2006, 2010).

7 Conclusions and Further Research

In this article, we presented the first systematic study in authorship attribution of the effect of author set size and data size on performance and feature selection. In order to estimate the viability of a given approach when applied 'in the wild' (Koppel *et al.*, 2010), typically involving large sets of candidate authors and limited amounts of data, it is vital to investigate how it is affected by the number of potential authors and the amount of text data available.

We approach authorship attribution as a text categorization task, meaning we build a model based on training data, and confront that model with texts of unknown authorship. The text categorization approach is challenged by the limited training data, consisting of short text fragments about 100 words in length, representing an approximation of the length of an e-mail.

Most studies in authorship attribution focus on small sets of authors with typically less than ten candidate authors. As expected, an approach that achieves an accuracy of 95% on such a small author set, will not to be able to deliver a similar performance with a large number of authors. We have shown that performance, while still significantly above baseline, decreases with increasing number of authors to a level where practical usability is no longer realistic. As far as feature selection is concerned, we find that similar *types* of features tend to work well for small and large sets of authors in our data sets. In most cases, character *n*-grams outperform the other feature types. We found evidence to support the claim that providing a more heterogeneous feature set—for example by adding syntactic information—has a positive effect on performance. Whereas robust feature types seem to exist, robust *individual* features do not emerge. A feature with a specific distribution may have good predictive power for a one set of authors, but will not generalize towards other author sets.

The harmful effect of increasing author set size is visible in the three data sets we presented, irrespective of their size, language, or number of topics. The extent to which it occurs, however, is influenced by a number of factors. The most pertinent factors are the number of topics in the data set, corpus size, and the choice of methodological set-up.

As far as data size is concerned, 10,000 words per author is traditionally regarded a 'reliable minimum for an authorial set' (Burrows, 2007). Setting the minimum requirements for an authorial set necessitates taking into account the characteristics and dimensions of the data set, such as the domain, genre, number of topics, and the number of candidate authors. In that respect, gaining insight in the effect of data set size on performance and feature selection is very important. We presented learning curve experiments that show how performance increases with increasing amounts of training data, an effect visible in the three data sets. Even on very small data sets, such as PERSONAE, which contains 1,400 words per author for 145 candidate authors, MBL scores relatively well. Similar to the author set size experiments, character *n*-grams work best, only the difference with the runners-up is smaller. Although the results confirm the idea that 'There is no data like more data' (Moore, 2001), factors like the number, variety, and type of topics seem to play an important role as well.

In further research, we will investigate the robustness of different types of ML algorithms for

tasks with many authors and small data sets. Some types of algorithms may have a better bias than others for handling this type of learning problem. We will also expand the scope of the study to more corpora and investigate additional (combinations of) features. Finally, whereas homogeneous datasets keeping topic, genre, register, and domain constant, facilitate evaluation of author style characteriztics, they also represent an idealized situation that will not be found in real-life problems, and results will be over-optimistic. We plan to systematically investigate the interaction of topic detection and authorship attribution to get a firmer grip on these issues.

Funding

This study was carried out in the framework of the 'Computational Techniques for Stylometry for Dutch' project, funded by the National Fund for Scientific Research—Flanders (FWO) in Belgium.

References

Abbasi, A., and Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2): 7:1–29.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. (2003a). Gender, genre, and writing style in formal written texts. *Text*, 23(3): 321–46.

Argamon, S., Saric, M., and Stein, S. (2003b). Style mining of electronic messages for multiple authorship discrimination: First results. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington DC: Association for Computing Machinery, pp. 475–80.

Argamon, S., Whitelaw, C., Chase, P. *et al.* (2007). Stylistic text classification using functional lexical features. *Journal of the American Society of Information Science and Technology*, 58(6): 802–22.

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): 121–31.

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

Burrows, J. (2007). All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22(1): 27–47.

Cavnar, W. and Trenkle, J. (1994). N-gram-based text categorization. *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas, NV: University of Nevada, pp. 161–75.

Clement, R. and Sharp, D. (2003). Ngram and Bayesian classification of documents for topic and authorship. *Literary and Linguistic Computing*, 18(4): 423–47.

Cohen, W. (1995). Fast effective rule induction. *Proceedings of the 12th International Conference on Machine Learning*. Tahoe City, CA: Morgan Kaufmann, pp. 115–23.

Coyotl-Morales, R., Villaseñor Pineda, L., Montes-y Gómez, M., and Rosso, P. (2006). Authorship attribution using word sequences. *Proceedings of the 11th Iberoamerican Congress on Pattern Recognition*. Cancun, Mexico: Springer, pp. 844–53.

Daelemans, W. and van den Bosch, A. (2005). *Memory-Based Language Processing. Studies in Natural Language Processing*. Cambridge: Cambridge University Press.

Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. (2007). TiMBL: Tilburg Memory Based Learner, version 6.1, reference guide. *ILK Research Group Technical Report Series no. 07-07*. The Netherlands: University of Tilburg.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2000). Authorship attribution with support vector machines. *Applied Intelligence*, 19(1–2): 109–23.

Dunning, T. (1994). Statistical identification of language. *Technical Report MCCS 94-273, Computing Research Lab (CRL) MCCS-94-273*, New Mexico State University.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3: 1289–305.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland: Association for Computational Linguistics, pp. 611–7.

Grant, T. and Baker, K. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics*, 8(1): 1350–771.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, 22(3): 251–70.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4): 405–17.

Holmes, D. (1994). Authorship attribution. *Computers and the Humanities*, 28(2): 87–106.

Houvardas, J. and Stamatatos, E. (2006). N-gram feature selection for authorship identification. *Proceedings of Artificial Intelligence: Methodology, Systems, and Applications (AIMSA)*. Varna, Bulgaria: Springer, pp. 77–86.

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

John, G. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. Quebec, Canada: Morgan Kaufmann, pp. 338–45.

Juola, P. (2004a). Ad-Hoc Authorship Attribution Competition. *Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH) Book of Abstracts*. Göteborg, Sweden: Göteborg University.

Juola, P. (2004b). On composership attribution. *Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH) Book of Abstracts*. Göteborg, Sweden: Göteborg University.

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

Keselj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based author profiles for authorship attribution. *Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics*. Halifax, Canada: Pacific Association for Computational Linguistics, pp. 255–64.

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

Koppel, M. and Schler, J. (2004). Authorship verification as a one-class classification problem. *Proceedings of the 21st International Conference on Machine Learning*. Banff, Alberta, Canada: Association for Computing Machinery, pp. 489–95.

Koppel, M., Schler, J., and Argamon, S. (2010). Authorship attribution in the wild. *Language Resources and Evaluation.* Advanced Access published January 12, 2010:10.1007/s10579-009-9111-2.

Koppel, M., Schler, J., Argamon, S., and Messeri, E. (2006). Authorship attribution with thousands of candidate authors. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*. Seattle, WA: Association for Computing Machinery, pp. 659–60.

Koppel, M., Schler, J., and Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8: 1261–76.

Luyckx, K. and Daelemans, W. (2008a). Authorship attribution and verification with many authors and limited data. *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester, UK: COLING 2008 Organizing Committee, pp. 513–20.

Luyckx, K. and Daelemans, W. (2008b). Personae: a corpus for author and personality prediction from text. *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco: European Language resources Association.

Madigan, D., Genkin, A., Lewis, D., Argamon, S., Fradkin, D., and Ye, L. (2005). Author identification on the large scale. *Proceedings of the 2005 Meeting of the Classification Society of North America*. St. Louis, MO: Classification Society of North America.

Mairesse, F., Walker, M., Mehl, M., and Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30: 457–500.

Mikros, G. and Argiri, E. (2007). Investigating topic influence in authorship attribution. *Proceedings of the 30th SIGIR, Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN)*. Amsterdam, The Netherlands: CEUR-WS, pp. 29–35.

Miranda García, A. and Calle Martín, J. (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1): 49–66.

Moore, R. (2001). There's no data like more data (but when will enough be enough?). *Proceedings of the IEEE International Workshop on Intelligent Signal Processing*. Budapest, Hungary: IEEE.

Mosteller, F. and Wallace, D. (1964). Inference and disputed authorship: The *Federalist. Series in Behavioral*

Science: Quantitative Methods Edition. Reading, MA: Addison-Wesley.

Nowson, S. and Oberlander, J. (2007). Identifying more bloggers. Towards large scale personality classification of personal weblogs. *Proceedings of the 1st International Conference on Weblogs and Social Media*. Boulder, CO: AAAI.

Platt, J. (1998). Fast training of Support Vector Machines using Sequential Minimal Optimization. In Schxlkopf, B., Burges, C., and Smola, A. (eds), *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press, pp. 185–208.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Sanderson, C. and Guenter, S. (2006). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Sydney, Australia: Association for Computational Linguistics, pp. 482–91.

Sebastiani, F. (2002). Machine learning in automated text categorization. *Association for Computing Machinery Computing Surveys*, 34(1): 1–47.

Stamatatos, E. (2007). Author identification using imbalanced and limited training texts. *Proceedings of the 18th International Conference on Database and Expert Systems Applications*. Regensburg, Germany: IEEE Computer Society, pp. 237–41.

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing and Management*, 44(2): 790–9.

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., Fakotakis, N., and Kokkinakis, G. (2000). Text genre detection using common word frequencies. *Proceedings of the 18th International Conference on Computational Linguistics*, vol. 2. Saarbrücken, Germany: Association for Computational Linguistics, pp. 808–14.

van Halteren, H. (2007). Author verification by linguistic profiling: an exploration of the parameter space. *Association for Computer Machinery Transactions on Speech and Language Processing*, 4(1): 1–17.

van Halteren, H., Baayen, H. R., Tweedie, F., Haverkort, M., and Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1): 65–77.

Weiss, S. and Kulikowski, C. (1991). *Computer Systems that Learn*. San Mateo, CA: Morgan Kaufmann.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Proceedings of the 2nd Asia Information Retrieval Symposium*. Jeju Island, Korea: Springer, pp. 174–90.

Zipf, G. (1935). *The Psycho-biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.