# Local stylometric features for authorship attribution in French fiction

**Michael Haaf**
260846673
michael.haaf@mail.mcgill.ca

**Étienne Fortier-Dubois**
260430244
etienne.fortier-dubois@mail.mcgill.ca

Group 44
December 2018

## Abstract

The performance of a classifier for authorship attribution is highly dependent on the features extracted from the texts. In this work, we explore various approaches for feature selection in the context of novel-length fiction in French. We find that stylometric features that can be computed locally on individual document segments perform as well as contemporary authorship attribution techniques. Character $n$-grams outscore other models on our small and homogeneous dataset.

## 1 Introduction

Authorship attribution is a classification task that aims at correctly finding the author of a given piece of text. While authorship attribution has several practical applications, including plagiarism detection, forensics, and author profiling, its classical focus since the 19th century has been in resolving literary questions (Stamatatos, 2009). With the increased efficiency of statistical analysis and development of powerful machine learning techniques, there is renewed interest in the problem as part of the field of digital humanities.

Many text classification tasks, such as sentiment analysis or topic categorization, focus on the semantic contents of the text, using some representation of that to classify the piece as part of a sentiment class, topic, etc. By contrast, authorship attribution focuses on *style* or, as Argamon et al. (2007) put it, "how" a text is written as opposed to "what it is about". The most useful features for classification may be very different between the two types of tasks. For instance, the use of commas in a news article may tell us nothing regarding whether it is about politics or sports, but be very indicative of the style of a particular author. Stamatatos (2009) surveys the types of these *stylometric* features in the context of authorship attribution.

The starting point for our work was a corpus of 47 classic novels by 27 French authors. This corpus was assembled in the Digital Humanities Lab at McGill and is similar to a corpus that was statistically analyzed by Brunet (1989). While Brunet was concerned with validating a high-level idea that some authors, across all of their works, discuss certain topics more than others, we use this corpus instead to compare the effect of various stylometric feature types that can be computed on document segments for classification. These feature types involve different domains of language: morphological (characters), lexical (words), syntactic (for instance, parts of speech), and semantic (for instance, using WordNet; not considered in this work). We show that these local stylometric techniques can reliably classify authors on this corpus.

## 2 Related work

Brunet (1989) was among the first to study French literature with a quantitative angle. Using the work of the author Colette and the theme of animals, he performed statistical comparisons of the occurrences of various animal names within Colette's body of work and with other authors from the same period. This was meant to demonstrate the use of statistics and computer technology in literary analysis, but did not involve any further applications such as authorship attribution.

A study by Laroche (2010) is claimed to be the first to examine authorship attribution in French-language literature. Laroche's corpus is larger than ours (114 texts and 53 authors), but also more heterogeneous, containing essays, poems and other genres in addition to novels. She compared two types of models, language models (word $n$-grams) and stylometric models (ratios of parts of speech frequencies), finding the former to give better recall.

Word and character $n$-grams are the typical feature types used for authorship attribution, and are regarded as efficient (Sari et al., 2017). However, several studies describe approaches that go beyond them. Argamon et al. (2007) argue that the *function* makes use of features built upon the principles of Systemic Functional Grammar (SFG). Their system considers the base frequency of function words (prepositions, pronouns, etc.) as well as the frequencies of functional units as defined in SFG-based taxonomies. These units express rhetorical functions such as elaboration, enhancement, probability, obligation, appreciation, judgment, or attitude. Many of these features are found to be useful for an array of stylistic classification tasks, including authorship attribution in a small corpus of British and American novels.

More recently, Pokou et al. (2016) described an extension of the $n$-gram models applied to parts of speech (POS). POS are interesting in stylistic classification because they represent syntactic information, as opposed to lexical or semantic information, and may be indicative of an author's style. Pokou et al. created a system using skip-grams, a generalization of $n$-grams that allows gaps between units, and therefore includes non-consecutive patterns. Deriving the a global stylometric 'signature' of an author by their most frequently used POS skip-grams across all an author's work, and differentiating each signature with the common POS skip-grams used by all authors, allowed them to reach 67% accuracy identifying authors from an English fiction corpus.

Word and character $n$-grams are discrete structures. Sari et al. (2017) considered a model in which $n$-grams are represented as continuous vectors instead. They found their continuous character $n$-grams approach comparable or superior to state-of-the-art methods for some datasets. In an ablation study, they determined that the continuous $n$-grams with the most impact on performance were those containing punctuation and white space.

## 3 Methodology

Our code and corpus are available for download or extension on GitHub[1].

### 3.1 Corpus

Our corpus consists of 47 classic French novels written from 1801 to 1954. There are 27 authors, most represented by a single work, although a few authors have several (for instance, Colette has 9 novels). Each work is novel-length (at least 100 pages), but there are disparities in length and thus in word count per author. The corpus was assembled by the Digital Humanities Lab from free ebook repositories such as Project Gutenberg[2] and TV5 Monde[3].

### 3.2 Data preparation

***Tokenization.*** When using character $n$-grams, we worked with the raw text. For the other features, each novel was processed using TreeTagger[4] (Schmid, 1994) for French. TreeTagger is a tool that tokenizes text and produces tag annotations, where each tag contains the token itself, the POS for that token, and the lemma. In contrast to many text classification tasks, we did not remove stop word nor punctuation tokens, motivated by the notion that stop word and punctuation usage vary characteristically by author. We see this notion borne out in our results. We additionally used the NLTK MosesDetokenizer[5] to reconstruct raw text.

***Instance-based document segmentation.*** There are two approaches to model construction for in authorship attribution (Stamatatos, 2009): *profile-based*, in which all texts from an author are concatenated in an attempt to determine the "profile" of the author, and *instance-based*, in which each text is considered separately. Given that authorship was already known for all of our texts, instance-based authorship attribution allowed us to test our corpus without having to arbitrarily select books to leave out of the training/test sample. Argamon et al.

[1] github.com/michaelhaaf/comp550-term-project
[2] www.gutenberg.org
[3] bibliothequenumerique.tv5monde.com/livres/1
[4] www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[5] github.com/alvations/sacremoses

(2007) took a similar approach by manually splitting all novels into chapters, motivated by the cohesion and independent discourse properties of novel chapters. Manually segmenting novels by chapters is tedious work requiring human expertise; we sought to automate a similar process by segmenting each novel into 1000-token "documents". While there is no consensus on document size for automated segmentation (Stamatatos, 2009) Feiguina and Hirst (2007) conclude that instance-based classification accuracy decreases at segments with less than 1000 tokens. We independently verified this heuristic.

***Balancing the corpus.*** Since authors do not produce uniform amounts of text, and not all texts produced are publicly available, it is common in author attribution problems to have unbalanced corpora. The 'true' distribution of authors on unknown data, however, is not proportional to the distribution of authors on known data. Stamatatos (2009) argues that, in training/test data splits, the test data set should be balanced; that is, uniformly distributed across authors. In our experiments, we ensured that all authors had just 20 randomly selected documents (corresponding to 20,000 tokens) available for each in the test set (using a standard training/test split of 80/20). Any authors with less than this amount of data were excluded from analysis. This resulted in a test set of 16 authors and 320 documents. We verified the intuition that balancing the test data does improve the performance of classifiers for authors less represented in the corpus.

### 3.3 Classifier and evaluation

For each of several feature sets, we classified documents using support vector machines (SVM) with stochastic gradient descent, implemented in the scikit-learn library (Pedregosa et al., 2011) under the name `SGDClassifier`. Since our balanced corpus gives a multilabel classification problem with 16 authors, we trained 16 different SVM classifiers (one per author) each employing the one-vs-rest strategy. We performed 5-fold cross-validation to tune the hyperparameters `alpha` (over values $\{0.00001, 0.000001\}$) and `max_iter` (over values $\{10, 50, 80\}$). We used the F1 score to evaluate the performance of the model for each feature set.

Given that many of our feature sets produce a large number of features, and given that stylometry in author attribution is a problem with linguistic interest, we forced our classifier to use L1-norm regularization. L1-norm regularization emphasizes highly performant features while zeroing out inconsequential features. In doing so, we produced feature sets that are easier to comprehend, emphasizing a relatively smaller number of features.

### 3.4 Feature sets

***Character $n$-grams.*** Character $n$-grams are all sequences of $n$ characters found in a text. They are considered useful in stylistic classification (Stamatatos, 2009; Sari et al., 2017) in part because they are resistant to spelling variations, which can be indicative of author style. According to Stamatatos (2009), the best value of $n$ varies by language and is dependent on typical word length; $n = 4$ is considered optimal for English. Since French has generally similar word lengths to English, we tried two models both including $n = 4$: a short $n$-gram model has $n$ in the range of $(2, 4)$, and a long $n$-gram model has $n$ in the range $(4, 8)$.

***Token $n$-grams.*** Word $n$-grams are the classical feature type for text classification and can be useful in identifying and distinguishing characteristic author word choice. We considered all tokens (words and punctuation) and used $n$ values in the range of $(1, 2)$. We tried this for both lemmatized and raw versions of the input.

***Function words.*** Function words are words that express grammatical relationships or attitudes rather than lexical meaning. They can be effective in stylistic classification (Argamon et al., 2007). As Stamatatos (2009) notes, "the selection of the specific function words that will be used as features is usually based on arbitrary criteria and requires language-dependent expertise." We identified as function words those that were POS-tagged as determiners, interjections, conjunctions, prepositions, and pronouns. The feature value of a function word $w$ in a document $d$ is

$$f(w) = \frac{\text{count}(w, d)}{\sum_{w' \in FW} \text{count}(w', d)}$$

where $FW$ is the set of all function words (Argamon et al., 2007).

***POS skip-grams.*** We implemented the POS skip-gram approach from Pokou et al. (2016) with some

| Feature set | Precision | Recall | F1 score |
|---|---|---|---|
| Baseline (majority class) | 0.06 | 1.00 | 0.12 |
| *Proper nouns included* | | | |
| Character (2,4)-grams | 0.97 | 0.97 | 0.97 |
| Character (4,8)-grams | 0.99 | 0.99 | 0.99 |
| Token (1,2)-grams (lemmatized) | 0.98 | 0.98 | 0.98 |
| Token (1,2)-grams (raw) | 0.98 | 0.98 | 0.98 |
| *Proper nouns removed* | | | |
| Character (2,4)-grams | 0.98 | 0.97 | 0.98 |
| Character (4,8)-grams | 0.99 | 0.98 | 0.98 |
| Token (1,2)-grams (lemmatized) | 0.97 | 0.97 | 0.97 |
| Token (1,2)-grams (raw) | 0.95 | 0.95 | 0.95 |
| Function words | 0.85 | 0.77 | 0.79 |
| POS skip-grams | 0.91 | 0.89 | 0.90 |
| POS skip-grams + function words | 0.93 | 0.90 | 0.91 |

**Table 1:** Precision, recall and F1 score (macro average) per feature set.

differences. Using the best results from that paper, we chose bigrams with gaps of at most 1. Only those bigrams appearing at least twice in a document were considered. The feature value of skip-gram $x$ in document $d$ is

$$f(x) = \frac{\text{count}(x, d)}{\sum_{x' \in SG} \text{count}(x', d)}$$

where $SG$ is the set of skip-grams appearing more than once. Unlike in Pokou et al. (2016), we did not attempt to derive author signatures: we input skip-grams into the classifier as a proxy for signatures, and considered each document separately (no global accumulation of skip-grams per author).

## 4 Results

The precision, recall and F1 score for each feature set are shown in Table 1. After running token and (long) character $n$-grams and obtaining high performance, we examined the features and found that many of the top features were proper nouns, which would typically be unique to a novel. To mitigate this effect, we ran these models again while excluding words that were POS-tagged as proper nouns. This gives slightly reduced performance by all metrics, but is presumably more generalizable.

The best performance is achieved when using long character $n$-grams, with $n$ in the range $(4, 8)$. Fig. 1 shows the features from that set with the largest negative and positive coefficients; that is, coefficients across all SVM classifiers that are the most useful in determining that a text does not belong to an author (negative) or does (positive). Several of these top features include punctuation, while many of them include the characters "et". These features also seem to be in the lower part of the range (4 instead of 8 characters). This and the similar performance of $(2, 4)$-grams suggest that 4 is the (or close to the) optimal value of $n$ for French like in English.

Token and character $n$-grams perform well. The alternative approaches (function words and POS skip-grams) failed to match their performance. This may be because they are simpler models: there are a limited number of functional words and POS tags compared to word and character $n$-grams. Combining these two feature types improved performance; thus the union of feature sets may be a valid strategy to capture more information and create stronger models for authorship attribution.

## 5 Discussion and conclusion

In our limited setting and with our small and homogeneous dataset, authorship attribution is not a hard problem. The language modelling approaches using character and token (word and punctuation) $n$-grams provided near-perfect performance. This is higher than in Laroche (2010), whose best model has a recall of 75%, although the comparison is not very meaningful, considering that her dataset is larger and more heterogeneous. Laroche suggested that the homogeneity of the corpus is an important factor in the success of authorship attribution, and our results support this.

Punctuation seems to be common among the top

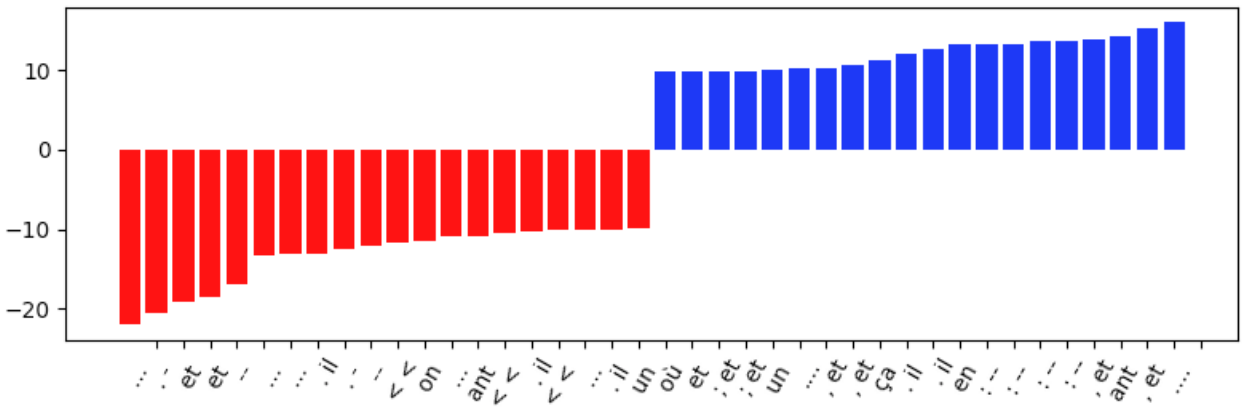


**Figure 1:** Character (4,8)-grams (when proper nouns are excluded) with the 20 largest negative and positive coefficients (the y-axis shows the coefficients).

features of a character $n$-gram feature set. However, punctuation signs are not found among the top features of token $n$-grams. This suggests that while the mere presence of punctuation signs may not be particularly indicative of style, the way these signs are combined is. This is in line with the finding that excluding punctuation from token $n$-grams in Sari et al. (2017) decreased performance.

One concern with our methods is that the test set comes from the same set of literary works as the training set. As a result, our classifiers may be overfitting to global authorship features (such as character names within a novel, or topics that authors discuss across works) that are available in the training data but could not generalize to unseen novels. The fact that performance improved when we removed proper nouns is a clue that this effect is real to some extent: there are likely other lexical features within the novels that, without being proper nouns, are specific enough to be found only in one author's work. The fact that the most significant features found in n-gram analysis were punctuation, however, contradicts this concern.

There are of course many types of features we did not consider in our analysis: examples include continuous vector representations (Sari et al., 2017), word and character skip-grams (Pokou et al., 2016), and advanced syntactic features (Argamon et al., 2007). While our survey of features is not exhaustive, it demonstrates that local stylometric features that are successful in English authorship attribution generalize to French and novel-length fiction.

## Statement of contributions

MH focused on the implementation and ÉFD focused on the writing, but both authors contributed to all steps of the project.

## References

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Étienne Brunet. 1989. L'exploitation des grands corpus : le bestiaire de la littérature française. *Literary and linguistic computing*, 4(2):121–134.

Olga Feiguina and Graeme Hirst. 2007. Authorship attribution for small texts: Literary and forensic experiments. In *PAN*.

Audrey Laroche. 2010. Attribution d'auteur au moyen de modèles de langue et de modèles stylométriques. *Actes RECITAL*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Yao Jean Marc Pokou, Philippe Fournier-Viger, and Chadia Moghrabi. 2016. Authorship attribution using small sets of frequent part-of-speech skip-grams. In *FLAIRS Conference*, pages 86–91.

Yunita Sari, Andreas Vlachos, and Mark Stevenson. 2017. Continuous n-gram representations for author-

ship attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 267–273.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. *International Confernce on New Methods in Language Processing*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.