Linguistic correlates of style: authorship classification with deep linguistic analysis features

Michael Gamon
Microsoft Research
Microsoft Corp.
One Microsoft Way
Redmond, WA 98052
mgamon@microsoft.com

Abstract

The identification of authorship falls into the category of style classification, an interesting sub-field of text categorization that deals with properties of the form of linguistic expression as opposed to the content of a text. Various feature sets and classification methods have been proposed in the literature, geared towards abstracting away from the content of a text, and focusing on its stylistic properties. We demonstrate that in a realistically difficult authorship attribution scenario, deep linguistic analysis features such as context free production frequencies and semantic relationship frequencies achieve significant error reduction over more commonly used "shallow" features such as function word frequencies and part of speech trigrams. Modern machine learning techniques like support vector machines allow us to explore large feature vectors, combining these different feature sets to achieve high classification accuracy in style-based tasks.

1 Introduction

Authorship identification has been a long standing topic in the field of stylometry, the analysis of literary style (Holmes 1998). From a broader perspective, issues of style, genre, and authorship are an interesting sub-area of text categorization. Typically, text categorization concerns itself with classifying texts according to topics. For that objective, it is crucial to extract information about the content of a text. In contrast, issues of style, genre and authorship are about the "form" of a text. The analysis of style needs to abstract away from the content and focus on content-independent form properties of the linguistic expressions in a text. This makes style analysis a prime candidate for the use of linguistic processing to extract structural features. Viewed from a different angle, the abstractness of style assessment features makes them highly domain-independent and reusable, as long as they are used with a classification technique that tolerates large feature vectors.

Previously suggested methods of style categorization and authorship identification have made use of a number of content independent features:

- frequencies of function words (Mosteller et al. 1964)
- word length and sentence length statistics (dating back to 1851 according to Holmes 1998)
- word tags and tag n-grams (Argamon et al. 1998, Koppel et al. 2003, Santini 2004)
- "stability features" (Koppel et al. 2003) capturing the extent to which an item can be replaced by a semantically similar item
- rewrite rules in an automatic parse (Baayen et al. 1996)

In this paper, we demonstrate that a combination of features based on shallow linguistic analysis (function word frequencies, part of speech trigrams) and a set of deep linguistic analysis features (context free grammar production frequencies and features derived from semantic graphs) yields very high accuracy in attributing a short random text sample to one of the three Brontë sisters as its author. Through feature ablation experiments we show that both the syntactic information captured in syntactic rewrite rules and the semantic information from semantic graphs contribute to the final classification accuracy. We also argue that by using support vector machines as a machine learning technique, we can leverage a very large number of features, effectively combining the different feature sets into large feature vectors, eliminating the need for laborious manual search for features that may be correlated with style.

2 Data

To test our approach to authorship identification, we used texts from Anne, Charlotte and Emily Brontë. This decision was motivated by the fact that we could keep gender, education and historic style differences to a minimum in order to focus on authorship identification, and by the easy availability of electronic versions of several lengthy texts from these authors. The texts we used were:

Charlotte Brontë: *Jane Eyre*, *The Professor*

Anne Brontë: *The Tenant of Wildfell Hall*, *Agnes Grey*

Emily Brontë: *Wuthering Heights*

For each of the three authors we collected all sentences from those titles and randomized their order. The total number of sentences for each author is: 13220 sentences for Charlotte, 9263 for Anne and 6410 for Emily. We produced artificial documents of 20 sentences in length from these sets of sentences. We split the resulting 1441 documents 80/20 for training and test. This split yields 288 documents for test, and 1153 documents for training. All numbers reported in this paper are based on 5-fold cross validation.

3 Features

All linguistic features have been automatically extracted using the NLPWin system (for an overview see Heidorn 2000). Note that this system produces partial constituent analyses for sentences even if no spanning parse can be found. The only exception are sentences of more than 50 words which do not result in any assignment of linguistic structure.

3.1 Length features

We measure average length of sentences, noun-phrases, adjectival/adverbial phrases, and subordinate clauses per document.

3.2 Function word frequencies

We measure the frequencies of function word lemmas as identified by the NLPWin system. In order to be maximally "content-independent", we normalized all personal pronouns to an artificial form "perspro" in order to not pick up on "she" or "he" frequencies which would be linked to the gender of characters in the works of fiction rather than author style. The number of observed function word lemmas is 474.

3.3 Part-of-speech trigrams

We extract part-of-speech (POS) trigrams from the documents and use the frequencies of these trigrams as features. The NLPWin system uses a set of 8 POS tags. 819 different POS trigrams are observed in the data.

3.4 Syntactic productions

The parses provided by the NLPWin system allow us to extract context-free grammar productions for each sentence, similar to the features in Baayen et al. (1996). Examples of common productions are:

PP → PP DETP NOUN
INFCL → INFTO VERB NP
DECL → VP CONJ VP CHAR

For each observed production, we measure the per-document frequency of the productions. 15.443 individual productions (types) occurred in our data, the total number of production tokens is 618.500.

3.5 Semantic information

We extract two kinds of information from the semantic dependency graphs produced by the NLPWin system: binary semantic features and semantic modification relations. Examples of semantic features are number and person features on nouns and pronouns, tense and aspectual features on verbs, and subcategorization features (indicating realized as opposed to potential subcategorization) on verbs. There is a total of 80 such semantic features.

Semantic modification relations are represented in a form where for each node A in a semantic graph the POS of A, the POS of all its n daughters $B_{1..n}$, and the semantic relations $SR_{1..n}$ of all its daughters $B_{1..n}$ are given. Some common modification structures are illustrated below:

Noun Possr Pron (a nominal node with a pronominal possessor)

Verb Tsub Pron Tobj Noun (a verbal node with a pronominal deep subject and a nominal deep object)

Noun Locn Noun (a nominal node with a nominal modifier indicating location)

As with the previously discussed features, we measure per-document frequency of the observed modification structures. There are a total of 9377 such structures.

3.6 n-gram frequency features

The use of word n-gram frequencies is not appropriate for style classification tasks since these features are not sufficiently content-independent. In our experiments, for example, they could pick up on nouns referring to events or locations that are part of the story told in the work of fiction at hand. We included these features in our experiments only as a point of comparison for the purely "form-based" features. In order to prevent the most obvious content-dependency in the word n-gram frequency features, we normalized proper nouns to "NAME" and singular personal pronouns to "Perspro".

3.7 Feature selection

While the total number of observed syntactic and semantic patterns is very high, most of the patterns occur only very few times, or even only once. In order to eliminate irrelevant features, we employed a simple frequency cutoff, where the frequency of a pattern that occurs less than n times is not included as a feature.

4 The machine learning technique: Support vector machines

For our experiments we have used support vector machines (SVMs), a machine learning algorithm that constructs a plane through a multi-dimensional hyperspace, separating the training cases into the target classes. SVMs have been used successfully in text categorization and in other classification tasks involving highly dimensional feature vectors (e.g. Joachims 1998, Dumais et al. 1998). Diederich et al. (2003) have applied support vector machines to the problem of authorship attribution. For our experiments we have used John Platt's Sequential Minimal Optimization (SMO) tool (Platt 1999). In the absence of evidence for the usefulness of more complicated kernel functions in similar experiments (Diederich et al. 2003), we used linear SVMs exclusively.

5 Results

All results discussed in this section should be interpreted against a simple baseline accuracy achieved by guessing the most frequent author (Charlotte). That baseline accuracy is 45.8%. All accuracy differences have been determined to be statistically significant at the .99 confidence level.

5.1 Feature sets in isolation

Classification accuracy using the different feature sets (POS trigram frequencies, function word frequencies, syntactic features, semantic features) are shown in Figure 1. The four length features discussed in section 3.1 yielded a classification accuracy of only 54.85% and are not shown in Figure 1.

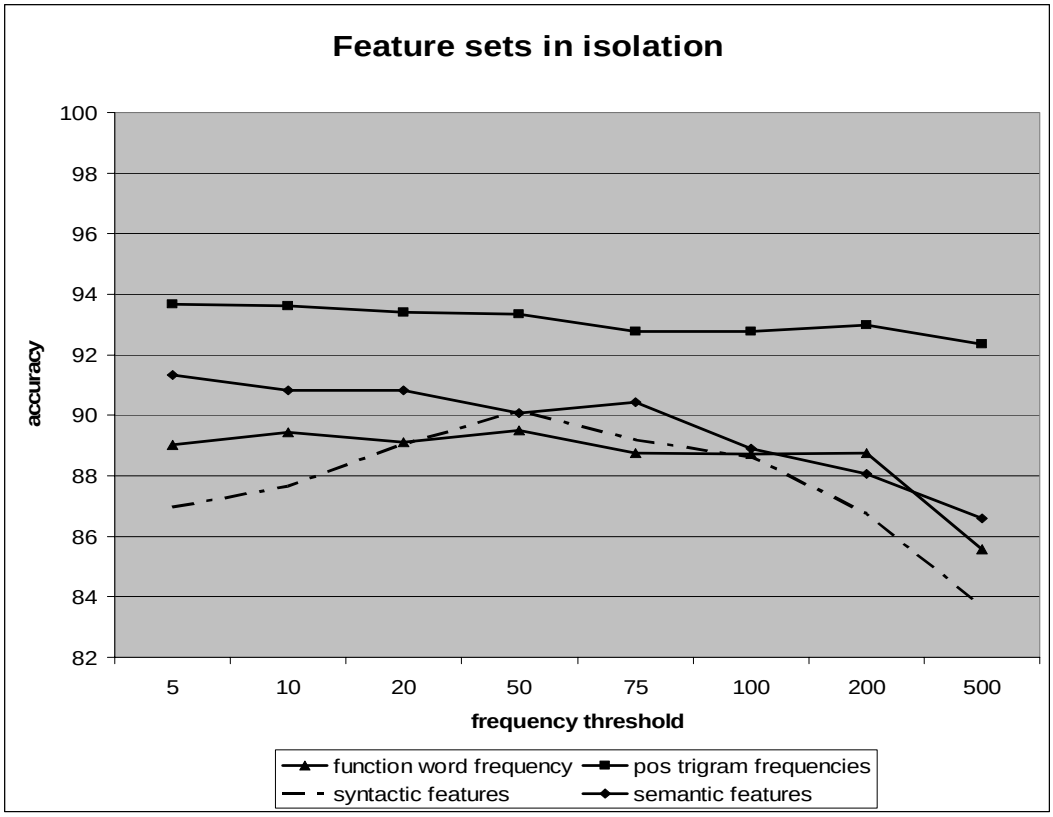

Figure 1: Classification accuracy using the feature sets in isolation

5.2 Feature sets combined

The combination of all feature sets yields a much increased classification accuracy across frequency thresholds as shown in Figure 2. Combining all features, including length features, consistently outperforms all other scenarios. Restricting features to those that only utilize shallow linguistic analysis, such as the POS trigram features and the function word frequency features reduces accuracy by about one percent. Interestingly, the use of syntactic and semantic features alone yields classification accuracy below the other feature combinations. In combination, though, these features contribute strongly to the overall accuracy. Semantic features which constitute the most abstract and linguistically sophisticated class, add to the accuracy of the classifier. This is evidenced by comparing the top two lines in Figure 2 which show the accuracy using all features, and the accuracy using all features except the semantic features.

Also included in Figure 2 is the accuracy obtainable by using "content-dependent" bigram and trigram frequency features. As stated above, these features are not adequate for style assessment purposes since they pick up on content, whereas style assessment needs to abstract away from content and measure the form of linguistic expression. It is noteworthy, however, that the true stylistic and "content-independent" features produce a classification accuracy that outperforms the ngram features by a wide margin.

Precision and recall numbers using all features with a frequency threshold of 75 (which yields the highest accuracy at 97.57%) are shown in Table 1.

Target	Precision	Recall	F-measure
Anne	97.20	98.08	97.64
Charlotte	98.18	98.20	98.19
Emily	96.81	95.52	96.16

Table 1: precision, recall and F-measure for the best model series with all features at frequency cutoff 75.

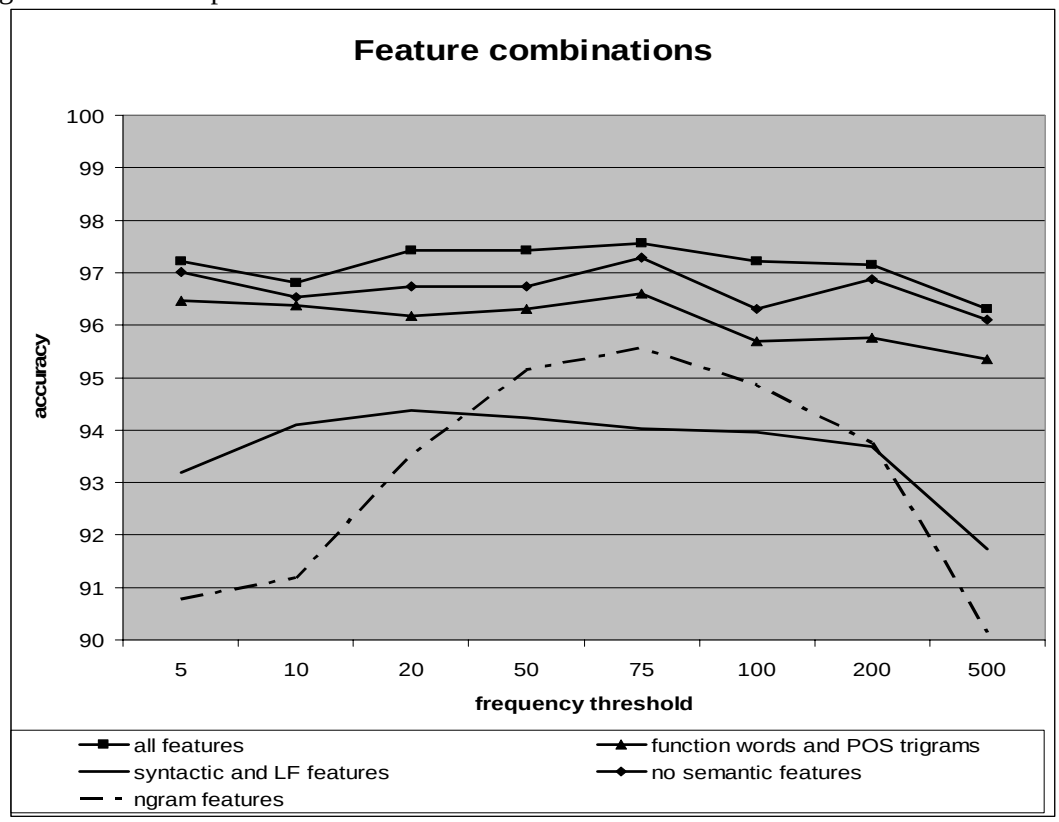

Figure 2: Classification accuracy based on combinations of feature sets

Table 2 shows error reduction rates for the addition of deep linguistic analysis features to the "shallow" baseline of function word frequencies and POS trigrams.

Frequency cutoff	+ syntactic features	+ syntactic and semantic features
5	15.70%	21.60%
10	3.80%	11.50%
20	14.50%	32.70%
50	11.30%	30.20%
75	20.40%	28.60%
100	14.50%	35.50%
200	26.20%	32.80%
500	16.40%	20.90%

Table 2: Error reduction rates achieved by adding deep linguistic analysis features to a baseline of POS trigram features and function word frequencies

5.3 Number of features and frequency threshold

Table 3 shows the number of features at each frequency cutoff. The total number of style-related features ranges from 6018 at a frequency cutoff of at least 5 observed instances to 546 at a frequency cutoff of 500. The size of these feature vectors is at the high end of what has typically been reported in the literature for similar experiments: For example, Argamon-Engelson et al. (1998) use feature vectors of size 1185 for newspaper style detection, Finn and Kushmerick (2003) have 36 POS features and 152 text statistics features for detection of "objective" and "subjective" genre, Koppel et al. (2004) use 130 features for authorship verification.

Frequency cutoff	All features	Function words	POS trigrams	Syntactic features	semantic features	Ngrams
5	6018	315	695	3107	1896	28820
10	3947	238	650	1885	1170	12312
20	2714	186	613	1176	735	5437
50	1730	140	542	623	421	1789
75	1421	125	505	442	345	1102
100	1233	116	466	355	292	781
200	870	88	385	201	192	357
500	546	62	257	101	122	114

Table 3: The number of features at different frequency cutoffs

6 Discussion

We believe that the results presented in the previous section allow a number of interesting conclusions for research into automatic style and authorship assessment. First, in our experiments the addition of deep linguistic analysis features increases classification accuracy.

From a linguistic perspective this is no surprise: it is clear that matters of linguistic form are those that can be captured by a syntactic and to some extent by a semantic analysis (as long as the semantic analysis is not so abstract that it completely abstracts away from any form properties of the sentence). It was less clear, though, whether an automatic language analysis system can be reliable enough to provide the necessary feature functions. This has been categorically denied in some of the literature (e.g. Stamatos et al. 2000). These statements, however, did not take into account that as long as a language analysis system is consistent in the errors it makes, machine learning techniques can pick up on correlations between linguistic features and style even though the label of a linguistic feature (the "quality" it measures) is mislabeled.

Secondly, we would like to emphasize that the results we have achieved are not based on deliberate selection of a small set of features as likely candidates for correlation with style. We have selected sets of features to be included in our experiments, but whether or not an individual feature plays a role was left to the machine learning technique to decide. Ideally, then, we would pass any number of features to the classifier algorithm and expect it to select relevant features during the training process. While this is possible with a large number of training cases, a smaller number of training cases poses a limit to the number of features that should be used to achieve optimal classification accuracy and prevent overfitting. In order to prevent overfitting it is desirable to reduce the vector size to a number that does not exceed the number of training cases. Support vector machines are very robust to overfitting, and in our experiments we find that classification results were quite robust to feature vectors with up to 4 times the size of the training set. However, it is still the case that optimal accuracy is achieved where the size of the feature vector comes close to the training sample (at a frequency cutoff of 75 for the vector containing all sets of features).

We also examined the features that carried high weights in the SVMs. Among the most highly weighted features we found a mix of different feature types. Below is a very small sample from the top-weighted features (recall that all features measure frequency):

- punctuation character starting a sentence (quote, double dash etc)
- *but*
- NOUN CONJ NOUN sequence
- *on*
- prepositional phrases consisting of preposition and pronoun
- VERB ADVERB CHAR sequences
- progressive verbs
- verbal predicates with a pronominal subject and a clausal object

In order to determine whether our results hold on sample documents of smaller size, we conducted a second round of experiments where document length was scaled down to five sentences per document. This yielded a total of 5767 documents, which we subjected to the same 80/20 split and 5fold cross-validation as in the previous experiments. Results as shown in Table 4 are very encouraging: using all features, we achieve a maximum classification accuracy of 85%. As in our previous experiments, removing deep linguistic analysis features degrades the results.

Frequency threshold	Number of all features	Number of shallow features	Accuracy using all features	Accuracy using shallow features
5	6018	1011	85.00	81.65
10	3947	889	84.96	81.56
20	2714	800	84.84	81.25
75	1421	631	84.53	80.59

Table 4: results on documents of a length of 5 sentences

It should also be clear that simple frequency cutoffs are a crude way of reducing the number of features. Not every frequent feature is likely to be discriminative (in our example, it is unlikely that a period at the end of a sentence is discriminative), and not every infrequent feature is likely to be non-discriminative. In fact, hapax legomena, the single occurrence of a certain lexeme has been used to discriminate authors. Baayen et al. (1996) also have pointed out the discriminatory role of infrequent syntactic patterns. What we need, then, is a more sophisticated thresholding technique to restrict the feature vector size. We have begun experimenting with log likelihood ratio (Dunning 1993) as a thresholding technique.

To assess at least anecdotally whether our results hold in a different domain, we also tested on sentences from speeches of George Bush Jr. and Bill Clinton (2231 sentences from the former, 2433 sentences from the latter). Using document samples with 5 sentences each, 10-fold cross-validation and a frequency cutoff of 5, we achieved 87.63% classification accuracy using all features, and 83.00% accuracy using only shallow features (function word frequencies and POS trigrams). Additional experiments with similar methodology are under way for a stylistic classification task based on unedited versus highly edited documents within the technical domain.

7 Conclusion

We have shown that the use of deep linguistic analysis features in authorship attribution can yield a significant reduction in error rate over the use of shallow linguistic features such as function word frequencies and part of speech trigrams. We have furthermore argued that by using a modern machine learning technique that is robust to large feature vectors, combining different feature sets yields optimal results. Reducing the number of features (i.e. the number of parameters to be estimated by the learning algorithm) by frequency cutoffs to be in the range of the number of training cases produced good results, although it is to be expected that more intelligent thresholding techniques such as log likelihood ratio will further increase performance. These results hold up even if document size is reduced to only five sentences.

We believe that these results show that the common argument of the "unreliability" of automatic linguistic processing used for feature extraction for style assessment is not as strong as it seems. As long as the errors introduced by a parser are systematic, a machine learning system presented with a large number of features can still learn relevant correlations.

Areas for further research in this area include experimentation with additional authorship and

style classification tasks/scenarios, experiments with different thresholding techniques and possibly with additional linguistic feature sets.

Additionally, we plan to investigate the possibility of training different classifiers, each of which contains features from one of the four major feature sets (function word frequencies, POS trigram frequencies, syntactic production frequencies, semantic feature frequencies), and maximally n such features where n is the number of training cases. The votes from the ensemble of four classifiers could then be combined with a number of different methods, including simple voting, weighted voting, or "stacking" (Dietterich 1998).

Acknowledgements

We thank Anthony Aue, Eric Ringger (Microsoft Research) and James Lyle (Microsoft Natural Language Group) for many helpful discussions.

References

Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. 1998. Style-Based Text Categorization: What Newspaper am I Reading? *Proceedings of AAAI Workshop on Learning for Text Categorization, 1-4.*

Harald Baayen, Hans van Halteren, and Fiona Tweedie. 1996. Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing 11(3): 121-131.*

Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass.2003. Authorship Attribution with Support Vector Machines. *Applied Intelligence 19(1):109-123.*

Thomas G. Dietterich. 1998. Machine Learning Research: Four Current Directions. *The AI Magazine 18(4): 97-136.*

Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. 1998. Inductive Learning Algorithms and Representations for Text Categorization. *Proceedings of the 7th International Conference on Information and Knowledge Management: 148-155.*

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics 19: 61-74.*

Aidan Finn and Nicholas Kushmerick. 2003. Learning to Classify Documents According to Genre. IJCAI-2003 Workshop on Computational Approaches to Text Style and Synthesis, Acapulco, Mexico.

George Heidorn. 2000. Intelligent Writing Assistance. In R. Dale, H. Moisl and H. Somers, eds., *Handbook of Natural Language Processing*. Marcel Dekker.

David I. Holmes. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing* 13(3):111-117.

Thorsten Joachims. 1998. Text Categorization with Support Vector Machines: Learning with many Relevant Features. *Proceedings of the tenth European Conference on Machine Learning:137-142.*

Moshe Koppel, Navot Akiva and Ido Dagan. 2003. A Corpus-Independent Feature Set for Style-Based Text Categorization. IJCAI-2003 Workshop on Computational Approaches to Text Style and Synthesis, Acapulco, Mexico.

Moshe Koppel, Jonathan Schler and Droz Mughaz. 2004. Text Categorization for Authorship Verification. Paper presented at the 8th Symposium on Artifical Intelligence and Mathematics, Fort Lauderdale, Florida.

Moshe Koppel, Shlomo Argamon, and Anat R. Shimoni. 2003. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing* 17(4): 401-412.

F. Mosteller. and D. L. Wallace. 1964. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers.* Addison-Wesley, Reading, MA.

John Platt. 1999. Fast Training of SVMs Using Sequential Minimal Optimization. In: B.Schölkopf, C. Burges and A. Smola (eds.) *Advances in kernel methods: support vector learning*. MIT Press, Cambridge, MA, 185-208.

Marina Santini. 2004. A Shallow Approach to Syntactic Feature Extraction for Genre Classification. *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics.*

Efstathios Stamatos, Nikos Fakotakis and George Kokkinakis. 2000. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics 26(4): 471-495.*