Stylometry in the Modern Era: Coreference and Voice for Authorship Attribution

Yong Sik Cho, Samuel Sharpe, Harry Smith
Columbia University
New York, NY
{yc3522, sbs2193, hs3061}@columbia.edu

Abstract

With the emergence of online media content in various forms, data are readily available online for researchers to explore the features of writing that define writing styles of authors. However, with the simultaneous increase in anonymity, identifying authorship has become a task of interest for many practical reasons ranging from forensics, text re-use, and even recommendations. Previous studies (Argamon et al., 2007; Stamatatos, 2009) show that simple features of writing such as word choice and character n-grams, as well as more complex features like semantic relations between words and intra-sentence syntax help capture nuanced differences in writing styles between authors and thus enhance accuracy in the task of authorship attribution (AA). These studies, however, focus on classic literature with lexical and semantic features or online texts and employed only lexical features. Our research aims to close this gap, improving AA on modern texts across various topics while also augmenting these features with long range syntactic structures and author voice. We perform experiments with articles from the Guardian and find that while coreference features and passive voice usage contain some signal of author style; furthermore, we confirm the success of simple lexical and syntactic features for AA on modern texts. We make our data, models, and code available on Github[1].

1 Introduction

Stylometry, the analysis of literary style between authors, is based on the ability to quantify and measure style. Although *style* is a nebulous concept, progress in literary analysis have shown style can be measured through carefully designed text-based metrics. Stylometry is often linked with and assessed through an Authorship Attribution (AA) task, a problem concerned with determining the author of a sample of text. AA is, in general, a supervised learning task wherein a corpus of documents of known authorship are used to attribute authorship of anonymous or disputed works. An investigator might use AA models for a variety of tasks: identifying the author of a threatening post on an extremist web forum (Stamatatos, 2018), discerning information about a document's author without guessing their actual identity, or confirming if one particular person is truthful in claiming authorship of some text (Stamatatos, 2009). Other potential uses of AA include formulating book recommendations or even applying the 'writing style' captured via AA in automated chatbot responses.

Several levels of information are available for completing the classification task, ranging from syntactical and lexical to semantic and functional (Argamon et al., 2007). Prior research in AA has demonstrated the effectiveness of the "lower level" lexical features for classification, but research continues developing efficient and widely applicable analysis of the semantic information contained in text. Furthermore, as media of writing have evolved with the ubiquity of computers, recent work has demonstrated that the success of AA strategies depends strongly on the genre of writing (Stamatatos, 2018). While traditional techniques found success on classic works of fiction, the emergence of new genres of writing (e.g. Tweets or forum posts) has necessitated the discovery of new methods in AA.

Our research attempts to assess how an author's voice and use of higher level syntactical patterns contribute to that author's unique style and expression. In particular, we attempt to answer the following question: *do coreference patterns and passive voice usage characterize writing styles?* We hypothesize that the addition of these syntac-

[1] https://github.com/ssharpe42/AuthorStyle

tic features may improve AA model performance on texts from The Guardian ranging from book reviews to political opinions, and are distinguishing elements of author writing style.

2 Previous Research

Stamatatos (2009) provides a useful summary of AA methods in his review, categorizing the available methods by the types of features that they explore. Stamatatos notes that the problem of AA is essentially two-fold: first, investigators must determine the *stylometric features* of interest from a document or corpus; second, they must decide the means by which they apply these features to actually attribute an author. Although the focus of our research is mainly on the former component of the AA problem, it remains important to consider exactly how the features an AA model incorporates will be used. Indeed, to properly address the hypotheses that we have posed, it will be crucial to control for the effects of latter component by comparing the performance of feature choices in only one model of attribution.

In his review, Stamatatos enumerates the different types of stylometric features and provides a working definition for each that we will adopt in this study. First, he defines the *lexical features*, which are those that can be extracted simply by using a tokenizer. Examples of these features include word counts, n-gram counts, and measures of vocabulary richness. Stamatatos notes, through summary of previous studies, that a "simple and successful method to define a lexical feature set" (Stamatatos, 2009) is to include only the counts of the most frequently used words, where the threshold for frequency rank is chosen between 100 and 1000. Second, he defines *character features,* which are gathered by observing the sequences of the characters in the text independent of the words that these make up. These features include character n-grams of fixed or variable length and character classes. Stamatatos notes that many studies through the past several decades have had good success with character features. Third, he defines *syntactic features* as those which examine the roles of words and sentence structures in writing. These features include part-of-speech (POS) tags and sentence or phrase structures. While Stamatatos highlights several studies which have successfully used these features, he notes that they are more difficult to derive from the original text than lower-level features. Stamatatos defines *semantic features* as those which depend on the meaning of the text. Examples of semantic features include the use of synonyms, the topic of the writing, and the functions of sentences. He notes that these features are often noisy, that they are the most difficult to extract from the text, and that they have the most sparse record of successful use in the field of AA. Nonetheless, as the extraction of more complex features from text becomes more feasible with advancements in machine learning, experiments with semantic style have shown promising results.

Feng and Hirst (2013) utilize *local coherence* as syntactic feature for AA models employing state-of-the-art neural coreference architectures. The authors build from earlier computational linguistic research showing how to model the references that authors naturally make to the entities they introduce in their text: for example, in this sentence, we reference "the authors", wherein we invoke "Feng and Hirst", mentioned in the sentence prior, as subjects. In order to model these patterns of coherence, the authors adopt the model of a document as an entity grid, where the rows represent the sentences of the text and the columns represent the entities that an author has referenced. Entries into the grid take one of four forms: *subject, object, other, none,* where the entries represent the type of reference made to the column's entity in the sentence represented by the row. The authors then use bigram probabilities of role transitions as features.

Feng and Hirst evaluate the performance of AA tasks with lexical features only, coherence features only, and both feature types together. In the task of pairwise AA for several 19th century novels written in English, the model is unable to match the performance of lexical features when provided only the coherence features; however, for many of the pairwise tasks, the model's performance with *both* lexical and coherence features improves with statistical significance over cases lexical features only. Similarly, in the task of one-vs-all classification (e.g., Did Nathaniel Hawthorne author this document or not?), the combined feature set corresponds with a boost in performance as well. While this represents yet another dimension for AA, we note that the extraction of coherence features require the use of a separate machine learning model (Argamon et al., 2007; Feng and Hirst, 2013).

Uzuner and Katz (2005) implement another successful incorporation of features above the complexity of lexical features. They created a feature set comprised of three broad categories: surface features, syntactic features, and semantic features. Surface features were mainly comprised of baseline features used in other AA literature, namely word counts, word lengths, sentence lengths, and number of sentences in a document. Syntactic features served to tap into the modes of expressiveness of a document by measuring the frequency of declarative sentences, interrogatives, imperatives, as well as passive voice. Finally, semantic features focused on capturing negations and uncertainty markers. They showed that the standard deviation of sentence lengths, passive voice, and standard deviation of word lengths were among the top ten most useful features. Although the AA task in this experiment was rather unique in that the objective was to distinguish between translators who translated the same content, the results are no less impressive in that it sheds light on key features that differentiate writers.

To overcome difficulties in ways of capturing complex features, researchers have also looked at alternative ways to represent and transform text to simultaneously preserve informative features and speed up computation. Gomez-Adorno et al. (2016) explore how sentences can be represented as interlinked syntactic graphs (ISG) which allows for a compact way of encapsulating many structural language features. This kind of representation makes it easy to measure and capture various features that fall within four broad categories: lexical, syntactic, morphological, and semantic relations. The lexical category includes features such as character counts independent of the sentences, while the morphological category captures features such as morphemes, roots, stems, and part-of-speech tags. The latter two categories serves to represent the more complex features such as dependencies and meanings of sentences. By using the C10 corpus, a subset of the Reuters Corpus Volume 1 (RCV1), experiments leveraging this mode of representation of sentences and documents showed promising results, which implies the utility of textual patterns extracted using syntactic graph features for AA tasks. The authors also demonstrate that a feature might capture aspects of only certain styles, and therefore that features might have limited predictive power when used across different sets of authors.

There are also studies that explore techniques to remove the effect of word frequencies due to topic-specificity. One such study is another investigation by Stamatatos, wherein he investigates the use of text distortion as a preprocessing step to improve AA performance (Stamatatos, 2018). Stamatatos hypothesizes that by masking information related to the topic of the writing, AA models based on character features will improve when attempting to attribute authorship to documents from either many topics, many genres, or both. Although this work is based in low-level feature sets only, Stamatatos provides a relatively novel look at the performance of models when topic and genre are not held constant through the set of documents to be processed. In particular, this research indicates that models will behave with varying performance depending on the genre of writing to which we apply them.

3 Methodology

We adopted a similar methodology to past AA works mentioned above. Given a corpus of documents by different authors, we illustrate the effectiveness of our experimental features (coreferences and voice) by formulating a variety of features to capture each concept. We compare their ability to predict authorship with other features in the literature using a neural network across all tasks.

3.1 Literature Derived Features

We extracted lexical and syntactical features used in Stamatatos's work (2009) including character, word, and part-of-speech n-grams, as well as stopword usage, word length, world length standard deviation, sentence length and sentence length standard deviation adapted from Uzuner and Katz (2005). We used a measure of vocabulary richness for each document using Yule's K (Tweedie and Baayen, 1998). Since we analyzed works across several topics, we also removed entities (e.g. name, location, time) from our word n-grams and kept only the top 100 most common n-grams[2]. We performed tokenization, part-of-speech tagging, parsing, and entity recognition with the spaCy python library[3] which implements

[2] The top 100 n-grams from the union of the sets of 1-grams and 2-grams.

[3] https://github.com/explosion/spaCy

Features	*L*	*C*	*V*	*LC*	*LV*	*CV*	*LCV*
character	✓			✓	✓		✓
word	✓			✓	✓		✓
POS	✓			✓	✓		✓
word length	✓			✓	✓		✓
word variance	✓			✓	✓		✓
sentence length	✓			✓	✓		✓
sentence variance	✓			✓	✓		✓
vocabulary richness	✓			✓	✓		✓
stopword	✓			✓	✓		✓
coreference		✓		✓		✓	✓
passive			✓		✓	✓	✓

Feature Sets: L: Lexico-syntactic; C: Coreference; V: Voice

Table 1: Overview of control and experimental feature sets.

many state of the art neural based methods (Honnibal and Montani, 2017). These features formed the lexico-syntactic feature set (*L*) which served both as a baseline of comparison for experimental feature sets and as a means of confirming the results from prior studies on new types of documents.

3.2 Coreference Resolution

The high-level characteristics of coreference are our motivation for using it as a means to capture authorial style. By measuring how authors refer to previously mentioned entities in a document, we hypothesize that coreference captures yet another aspect of writing style which can be used to improve authorship attribution. We found that Feng and Hirst's implementation of coreference, or what they call 'coherence', was concise and produced good results. Therefore, our implementation of coreference borrowed from their work and further expanded their concept to longer range dependencies (Feng and Hirst, 2013).

As described previously, their method computes transition probabilities for entity references. To illustrate this process, we provide the following example.

> *Bob went to the park. He walked the dog. The dog beat him in a race.*

In Table 2, we represent the role transitions of each entity in each sentence. We classify entities into either a subject (S), object (O), other, or no reference (_). For each document we calculate bigram probabilities for each transition type. If an entity is mentioned more than once in a sentence, we only use the first mention.

Bob	**the park**	**the dog**
S	O	_
S	_	O
O	_	S

Table 2: Coreference transition example.

We also extend to longer range dependencies by recording the frequency at which authors referred to original entities up to 10 sentences ahead. Finally, we included summary metrics such as references per sentence and standard deviation of references per sentence.

We leveraged the Neuralcoref python library[4] to perform coreference resolution on each document. Neuralcoref is an adaptation of the work by Clark and Manning that harnesses neural architectures for coreference resolution (Clark and Manning, 2016b,a). Neuralcoref trains parallel neural networks to score single references and reference pairings based on various word embedding, distance, and syntactic features. We utilized this method of extracting coreferences because of its ability to identify references across many sentences with reasonable accuracy.

3.3 Passive Voice

Similar to coreference, we chose the passivity of writing style as another metric along which we can capture and measure style. Since passive voice is not captured by simpler features such as n-gram counts or vocabulary richness, we hypothesize that the inclusion of such features can improve author-

[4] `https://github.com/huggingface/neuralcoref`

Feature Name	Explanation	Example Phrase (Italicized)
HatTrick	Passive subject, auxiliary verb, and an agent	I hate that *he was rejected by her.*
Agentless	Passive subject & auxiliary verb but no agent	I hate that *he was rejected.*
Description	Active Subject with agent dependency	The deal, *loved by Jim*, was bad.
NoActive	Sentence without any active subject	*The deal was rejected.*

Table 3: Descriptions of passive phrase frequency features.

ship attribution accuracy. We extracted two varieties of voicing features, one group computed at the sentence level and one computed at the individual word level. The first group consisted of four different features which we derived by analyzing the frequency with which an author uses several varieties of passive voice phrasings indicated by word dependencies. We did so by counting the number of sentences within a document which contained each of these phrasings, and dividing by the number of total sentences comprising that document. These phrasings are named and explained in Table 3. We decided upon these categories of passive phrase by examining the dependencies that SpaCy determined for example sentences gathered from The New York Times which we identified as having a passive voice construction.

To extract the second group of three features, we analyzed the choice of auxiliary verb that the authors used in their passive phrases. In particular, we calculated the proportion of auxiliary passive verbs which were forms of *to be* or *to get* and formed a third category for other verb instances.

3.4 Evaluation Procedure

We created seven feature set combinations detailed in Table 1 to analyze the effects of the experimental features (coreference resolution and author voice) individually and jointly. We evaluated the effect of these feature sets through three tasks of authorship attribution: multiclass, one-vs-all, and one-vs-one.

In each evaluation, we separated the data into five folds, performing cross validation to estimate accuracy. In each round of cross validation we split four folds into 80% and 20% training and validation, respectively, and used the remaining fold for evaluation. Training, validation, and evaluation sets were each randomly up-sampled so classes had equal representation resulting in a naive accuracy of 50% for one-vs-one and one-vs-all and a 6.25% (i.e., one in sixteen, where sixteen is the number of authors in our data set) naive accuracy for the multiclass task. This method of testing was adapted from Feng and Hirst (2013).

4 Data

We strove to find a modern data set with a broad set of genres and authors. Given the abundance of data across genres for each author, we used a data set similar to that of the Guardian10 corpus used by Stamatatos (2012) and Sundararajan and Woodard (2018). This data set was comprised of Guardian opinion articles and book reviews by a group of ten authors. Because the platform for accessing text has changed, we were unable to use the exact data set from these previous studies; however, we were able to aggregate 1665 Guardian opinion articles written by 16 authors between 2008 and 2016 across five different genres including Society, UK, World, Politics, and Book Reviews. This amounted to an average of 104 articles per author, 333 articles per genre, 5801 characters per article, and 935 words per article. We purposefully filtered for authors with at least 30 articles in this time frame as a means to ensure sufficient variety in training data for the classification model. This naturally results in document counts that varies across the 16 authors. We provide an overview of the corpus document by author and document in Table 4.

5 Results and Evaluation

We evaluated our hypothesis by measuring the effect of incorporating coreferences and passive voice features and assessing incremental improvements in authorship attribution. In particular, we fixed a neural network[5] and compared its performance on the authorship attribution task using different feature sets. Assessing the accuracy of authorship attribution among each of these different

[5] All neural network parameters were constant except the number of hidden layers which were scaled according to the number of features.

	Topics					
Author	Society	World	UK	Politics	Review	**Total**
A_1	3	13	14	6	4	40
A_2	26	19	59	17	22	143
A_3	8	10	49	105	9	181
A_4	8	46	5	2	10	71
A_5	7	3	8	5	15	38
A_6	14	29	31	17	17	108
A_7	6	8	17	33	18	82
A_8	3	120	16	42	2	183
A_9	6	3	70	10	1	90
A_{10}	5	36	25	22	1	89
A_{11}	6	21	1	1	8	37
A_{12}	5	22	2	11	1	41
A_{13}	4	96	49	59	6	214
A_{14}	31	28	17	33	6	115
A_{15}	7	16	32	51	5	111
A_{16}	32	23	28	13	29	125

Table 4: Document counts in corpus by author and topic.

feature sets allowed us to evaluate which experimental features, if any, capture a new dimension of authorial style that is not determined by previously explored lexico-syntactic features alone.

We start with accuracy of the experimental features compared to the baseline in the multiclass task (see Table 5) where accuracy is defined as the rate at which the model was able to correctly attribute authorship out of all the potential authors. As expected, the conventional lexico-syntactic features demonstrated accuracies that are significantly higher than the 6.25% naive guessing accuracy. The accuracy metrics for feature sets containing only coreference resolution or author voice features were significantly higher than naive guessing[6]. We also notice that combining the two experimental features did not significantly improve upon the accuracy of either feature alone ($p < .05$). Similarly, we observe that adding either experimental feature (or both together) to the lexico-syntactic feature set did not significantly improve accuracy over that of lexico-syntactic features alone. This suggests that coreference resolution and author voice features capture authorial style in isolation and are thus distinguishing elements of author writing style. However, our experimental features do not significantly improve AA model performance above the current state of the art.

Feature set	**Accuracy (%)**
L	83.5^a
C	17.3^a
V	17.5^a
CV	19.3^a
LC	83.5^a
LV	83.4^a
LCV	84.0^a

[a] Significantly better than naive guessing ($p < .05$)
[b] Significantly better than lexico-syntactic features ($p < .05$)

Table 5: Accuracy scores (%) of multiclass classification experiments.

Results from the one-vs-all task (Table 6) roughly mirror results from the multiclass task, but provide a more granular view of the performance at the individual author level. Overall, each author is distinguishable with a higher accuracy than uniformly random guessing for each of the seven feature sets, excluding seven of the 116 *author-feature set* pairs. We note that we find two authors (A_1 and A_{11}) whose texts are classified by their coreference and voice patterns with at comparable levels of accuracy to the level attained by using their lexico-syntactic patterns alone. Otherwise, each author is classified at a lower accuracy with coreference or voice features than with lexico-syntactic features.

Similarly, authors are distinguished more easily with baseline features in the pairwise classification task. We aggregate the results of the pairwise task (i.e. one entry for each of the seven feature sets across all 120 author-to-author matchups) in (Table 7), showing the number of experiments showing improvements over naive guessing and lexico-syntactic feature, and the average accuracy for each of the feature sets. We note that the number of experiments which result in statistically significant improvement over naive guessing when using the *CV* feature set is greater than the corre-

[6] Although we only show significance at the 0.05 level, all feature sets are significant to the 0.01 level

Author	*L*	*C*	*V*	*CV*	*LC*	*LV*	*LCV*
A_1	60.0^a	69.9^a	59.2	69.0^a	61.2^a	59.3^a	60.8^a
A_2	97.4^a	61.6^a	63.9^a	64.6^a	96.4^a	95.9^a	97.0^a
A_3	94.6^a	70.0^a	72.4^a	75.6^a	93.3^a	94.4^a	96.5^a
A_4	81.7^a	57.3^a	60.8	54.9	83.0^a	86.5^a	83.6^a
A_5	81.2^a	56.1	65.0^a	56.8	79.9^a	76.8^a	75.6^a
A_6	91.1^a	66.0^a	68.6^a	69.6^a	94.2^a	92.7^a	91.7^a
A_7	88.4^a	68.9^a	67.5^a	69.6^a	88.5^a	89.0^a	89.5^a
A_8	92.2^a	59.7^a	63.4^a	61.8^a	91.2^a	90.6^a	90.7^a
A_9	97.3^a	58.4^a	60.5^a	58.1^a	97.1^a	96.9^a	94.3^a
A_{10}	90.7^a	58.2	60.8^a	56.8	85.4^a	87.7^a	84.9^a
A_{11}	65.6^a	64.7^a	71.1^a	65.3^a	70.2^a	64.9^a	64.5^a
A_{12}	86.7^a	68.5^a	59.3^a	64.6^a	80.5^a	85.6^a	86.6^a
A_{13}	94.0^a	68.2^a	64.2^a	71.4^a	93.5^a	93.2^a	93.1^a
A_{14}	88.9^a	60.6^a	65.2^a	62.2^a	87.7^a	89.5^a	89.6^a
A_{15}	93.5^a	71.9^a	79.7^a	79.0^a	92.6^a	93.7^a	92.0^a
A_{16}	90.3^a	72.6^a	65.3^a	73.5^a	87.4^a	84.2^a	83.7^a

^{a}Significantly better than naive guessing ($p < .05$)
^{b}Significantly better than lexico-syntactic features ($p < .05$)

Table 6: Accuracy scores (%) of one-to-all classification experiments.

Feature set	**# of Significant Experiments $>$ Naive**	**# of Significant Experiments $> L$**	**Accuracy (%)**
L	120	0	95.8
C	90	0	67.7
V	83	0	67.6
CV	100	0	70.9
LC	120	2	96.1
LV	120	0	96.0
LCV	120	2	96.0

Table 7: Accuracy scores (%) of pairwise classification experiments aggregated across all authors for each feature set. See Appendix A for details.

sponding number for both *C* and *V* alone, suggesting that the interaction of the two feature sets may provide further predictive power. The full array of results for the classification accuracy of pairwise tasks are included in Appendix A.

6 Conclusion

Stylometry and the related task of AA has been a point of interest for many researchers given its forensic applications and its ability to demonstrate which features of an author's writing are fundamental to their unique voice. Recent studies explored how more complex features such as semantics, word functions, and references could be used to differentiate among candidate authors, hoping to provide more literary and human-understandable descriptions of what contributes to an author's personal style. Our study explored coreferences and passive voice usage by 16 authors over various genres of modern writing. The results from our experiments confirm the dominance of baseline lexico-syntactic features in distinguishing authors in modern domains like online opinion articles, notably different from the 19th century English novels that were often employed by previous works in this field. Our derived passive voice and coreference features, while comparably noisy, have some stylistic signal given that they provide statistically significant improvements from naive author attribution. This represents a relative breakthrough in the ability to articulate what makes an author's voice so distinct: whereas past explanations derived from AA results would depend on the opaque descriptions of word and character frequencies, our new features provide an intelligible means of explaining these distinctions through familiar concepts.

7 Future Work

Given that our data only included texts within certain domains (e.g. reviews, political), a potential future direction for additional research can encompass additional domains and genres. However, it is important to keep in mind that documents may need to be of reasonable lengths in order for complex semantic features such as coreferences to manifest itself appropriately. Another area to explore would be AA in non-English texts using references and passive voice. While English is a widely used language with approximately 171K words in a typical dictionary, other languages such as Korean, Japanese, and Italian are estimated to vastly outnumber English at 1M, 500K and 260K words each, suggesting the possibility that these languages offer a more varied means of expression and writing styles that can be captured by these features. Finally, the notion of expression and style as suggested by Uzuner and Katz (2005) in combination with our findings suggest that further research may be needed in refining and identifying more stylistic features that stay consistent across content and presumably genre.

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.

Kevin Clark and Christopher D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Kevin Clark and Christopher D. Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Vanessa Wei Feng and Graeme Hirst. 2013. Patterns of local discourse coherence as a feature for authorship attribution. *Literary and Linguistic Computing*, 29(2):191–198.

Helena Gomez-Adorno, Grigori Sidorov, David Pinto, Darnes Vilarino, and Alexander Gelbukh. 2016. Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. In *Sensors (Basel)*. Sensors.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

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

Efstathios Stamatatos. 2012. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21(2):421–439.

Efstathios Stamatatos. 2018. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473.

Kalaivani Sundararajan and Damon L. Woodard. 2018. What constitutes style in authorship attribution? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2814–2822. Association for Computational Linguistics.

Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352.

Ozlem Uzuner and Boris Katz. 2005. Style versus expression in literary narratives. In *SIGIR Workshop on Stylistic Analysis of Text for Information Access, Salvador, Bahia, Brazil*. Citeseer.

A Appendices

		A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}
A_1	L	88.5^a	95.6^a	86.2^a	92.6^a	94.7^a	86.8^a	89.9^a	95.0^a	85.8^a	87.6^a	96.0^a	93.7^a	80.6^a	95.9^a	83.9^a
	C	64.9^a	74.0^a	66.3	59.6	81.2^a	70.3^a	74.9^a	47.8	73.0^a	70.4^a	62.7	64.2^a	55.9	58.0	69.8^a
	V	66.2^a	63.8^a	71.0^a	51.1	76.2^a	75.1^a	61.3	54.3	71.6^a	76.7^a	67.0^a	66.0^a	63.4	54.9	66.6^a
	LC	90.2^a	94.6^a	88.7^a	97.5^a	92.8^a	90.0^a	85.1^a	92.7^a	87.4^a	94.1^a	96.4^a	91.0^a	84.2^a	98.5^a	86.6^a
	LV	93.1^a	92.6^a	86.8^a	91.4^a	94.3^a	87.8^a	85.4^a	96.5^a	82.9^a	88.1^a	96.0^a	91.8^a	87.1^a	95.3^a	86.1^a
	CV	67.2^a	65.8	70.4^a	68.2^a	83.6^a	70.7^a	66.6^a	59.4^a	77.2^a	76.6^a	71.8^a	64.2^a	63.3	58.7	72.1^a
	LCV	87.0^a	96.3^a	84.5^a	92.6^a	91.8^a	87.7^a	81.2^a	92.7^a	81.3^a	93.9^a	93.7^a	91.2^a	85.8^a	96.3^a	87.0^a
A_2	L		98.3^a	95.8^a	93.0^a	99.6^a	99.7^a	98.7^a	99.6^a	98.9^a	88.7^a	100.0^a	98.9^a	99.0^a	99.6^a	98.6^a
	C		68.0^a	59.1	46.8	72.4^a	68.8^a	60.0^a	64.8^a	62.0^a	68.3^a	76.6^a	75.8^a	61.1^a	73.2^a	73.7^a
	V		75.7^a	58.1	57.4	71.9^a	68.2	60.6^a	67.0^a	65.0^a	73.2^a	62.6	65.0^a	58.5^a	82.8^a	62.9^a
	LC		98.4^a	96.6^a	94.6^a	100.0^a	99.7^a	98.7^a	99.6^a	97.8^a	89.1^a	100.0^a	99.3^a	97.5^a	100.0^a	97.9^a
	LV		99.0^a	98.2^a	94.0^a	99.6^a	99.7^a	98.7^a	99.6^a	99.3^a	89.9^a	100.0^a	99.1^a	99.6^a	99.3^a	98.3^a
	CV		74.0^a	60.3^a	55.2	76.4^a	71.5^a	63.7	66.8^a	60.4	69.9^a	67.8	77.2^a	61.1^a	79.0^a	70.9^a
	LCV		98.6^a	97.4^a	94.6^a	99.6^a	98.2^a	98.7^a	97.1^a	98.9^a	92.2^a	100.0^a	99.3^a	100.0^{ab}	99.6^a	97.5^a
A_3	L			97.7^a	98.9^a	98.6^a	97.5^a	98.7^a	99.7^a	96.3^a	89.0^a	99.7^a	98.4^a	97.8^a	99.1^a	96.9^a
	C			58.5^a	54.9	73.8^a	57.7^a	73.4^a	77.0^a	74.9^a	67.8^a	80.0^a	81.3^a	72.4^a	80.2^a	74.8^a
	V			53.3	64.0^a	66.7^a	64.6^a	78.2^a	70.9^a	69.3^a	69.5^a	73.8^a	78.7^a	82.4^a	81.6^a	57.1
	LC			98.0^a	98.7^a	98.9^a	95.8^a	98.2^a	99.7^a	98.3^a	91.0^a	99.7^a	98.7^a	98.7^a	98.5^a	97.7^a
	LV			96.3^a	98.1^a	96.8^a	98.1^a	98.4^a	99.7^a	97.6^a	89.5^a	98.3^a	98.4^a	98.4^a	98.8^a	98.6^a
	CV			66.5^a	71.8^a	69.8^a	61.0^a	76.1^a	78.1^a	76.7^a	65.4^a	78.1^a	83.0^a	82.6^a	85.2^a	76.5^a
	LCV			97.4^a	95.1^a	97.9^a	98.7^a	98.2^a	100.0^a	97.3^a	91.5^a	96.7^a	98.9^a	98.8^a	97.9^a	98.6^a
A_4	L				96.0^a	94.4^a	93.8^a	96.9^a	98.8^a	93.2^a	90.1^a	97.8^a	97.3^a	96.2^a	98.2^a	91.7^a
	C				62.3	58.5	64.6	60.0	65.9^a	53.8	53.6	63.1^a	63.6^a	59.6	77.9^a	66.7^a
	V				61.9	66.7^a	55.2	54.6	65.8^a	54.0	57.5	59.1	69.9^a	70.4^a	84.7^a	51.6
	LC				97.7^a	91.3^a	92.8^a	96.7^a	97.2^a	94.1^a	90.6^a	98.2^a	97.3^a	96.6^a	97.8^a	90.6^a
	LV				96.4^a	95.9^a	92.8^a	97.0^a	96.6^a	92.5^a	91.1^a	96.9^a	96.6^a	94.1^a	99.6^a	90.0^a
	CV				55.7	60.6	67.2^a	59.1	62.0	55.8	57.7	61.3^a	71.8^a	61.7^a	82.4^a	66.6^a
	LCV				96.6^a	94.2^a	93.3^a	97.5^a	96.6^a	94.5^a	90.2^a	94.5^a	96.6^a	95.0^a	98.2^a	91.5^a

^{a}Significantly better than naive guessing ($p < .05$)

^{b}Significantly better than lexico-syntactic features ($p < .05$)

Table 8: Accuracy scores (%) of pairwise classification experiments. Authors $A_1 - A_4$.

		A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}
A_5	L					99.5^a	97.7^a	97.0^a	100.0^a	94.6^a	89.7^a	98.8^a	98.0^a	91.9^a	99.5^a	89.0^a
	C					73.5	67.1	56.9	66.0^a	61.8^a	58.9	72.3^a	71.0^a	63.4	60.7	62.2^a
	V					65.6^a	70.3	61.0	57.0	70.1^a	73.6^a	45.9	65.5^a	59.6	69.4^a	66.7
	LC					99.5^a	98.0^a	94.4^a	98.9^a	93.2^a	90.7^a	98.8^a	98.1^a	89.9^a	99.5^a	90.8^a
	LV					99.5^a	98.8^a	94.9^a	100.0^a	95.8^a	88.6^a	98.8^a	98.3^a	93.7^a	98.7^a	92.3^a
	CV					78.2^a	78.5^a	65.5^a	66.7^a	63.4^a	71.3^a	78.9^a	75.8^a	64.6^a	64.7^a	63.0^a
	LCV					99.5^a	96.5^a	94.5^a	100.0^a	94.9^a	89.7^a	98.8^a	98.3^a	89.4^a	97.2^a	92.7^a
A_6	L						92.1^a	98.0^a	96.4^a	95.9^a	98.1^a	96.2^a	98.2^a	99.2^a	99.1^a	96.1^a
	C						61.3^a	71.8^a	76.5^a	54.9	49.4	76.5^a	78.5^a	77.0^a	83.9^a	78.1^a
	V						53.0	66.6^a	76.1^a	50.1	65.5	66.8	75.0^a	78.1^a	87.0^a	65.2^a
	LC						91.8^a	98.1^a	95.9^a	98.2^a	95.6^a	96.7^a	98.4^a	99.6^a	98.7^a	97.6^a
	LV						93.9^a	97.4^a	97.7^a	98.2^a	95.7^a	95.2^a	97.3^a	98.7^a	98.7^a	97.3^a
	CV						62.3^a	77.3^a	77.8^a	60.7	54.0	71.6^a	81.6^a	80.3^a	87.9^a	73.8^a
	LCV						94.3^a	97.3^a	95.9^a	96.8^a	98.1^a	98.5^a	97.5^a	98.4^a	99.1^a	96.1^a
A_7	L							96.5^a	97.4^a	94.0^a	86.4^a	94.4^a	98.4^a	97.3^a	97.2^a	94.0^a
	C							78.4^a	77.1^a	68.4^a	45.4	79.4^a	79.2^a	78.0^a	82.9^a	71.3^a
	V							71.0^a	69.8^a	60.3	63.8	70.5^a	74.3^a	79.3^a	86.9^a	53.9
	LC							95.8^a	97.5^a	96.1^a	96.8^{ab}	94.8^a	96.5^a	97.4^a	98.1^a	94.8^a
	LV							96.4^a	98.4^a	95.1^a	87.6^a	91.9^a	97.9^a	98.7^a	97.6^a	94.8^a
	CV							80.7^a	81.2^a	69.9^a	53.2	81.1^a	81.7^a	82.9^a	86.7^a	73.9^a
	LCV							97.1^a	98.0^a	95.5^a	94.1^a	94.8^a	95.6^a	98.7^a	96.7^a	96.0^a
A_8	L								98.8^a	96.6^a	94.9^a	97.8^a	97.2^a	97.0^a	98.7^a	96.2^a
	C								67.8^a	60.0^a	72.1^a	67.8^a	68.6^a	58.5^a	74.2^a	74.8^a
	V								69.6^a	69.0^a	75.4^a	51.9	64.9^a	60.0^a	81.2^a	66.6^a
	LC								99.1^a	97.0^a	90.4^a	97.1^a	97.2^a	97.6^a	97.7^a	96.5^a
	LV								99.1^a	97.2^a	94.1^a	98.6^a	96.5^a	97.6^a	98.2^a	97.0^a
	CV								70.5^a	62.9^a	80.2^a	71.2^a	67.9^a	59.8^a	83.8^a	77.6^a
	LCV								99.2^a	98.1^a	95.2^a	100.0^a	96.8^a	97.1^a	98.4^a	96.7^a

[a] Significantly better than naive guessing ($p < .05$)

[b] Significantly better than lexico-syntactic features ($p < .05$)

Table 9: Accuracy scores (%) of pairwise classification experiments. Authors $A_5 - A_8$.

		A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}
A_9	L									98.5^a	100.0^a	100.0^a	99.5^a	100.0^a	99.2^a	99.6^a
	C									68.3^a	77.6^a	68.8^a	69.6^a	58.7	69.0^a	73.4^a
	V									69.8^a	77.1^a	60.5	62.6^a	62.6^a	69.6^a	69.4^a
	LC									97.0^a	100.0^a	100.0^a	99.1^a	100.0^a	99.2^a	100.0^a
	LV									98.5^a	100.0^a	100.0^a	99.3^a	99.7^a	99.6^a	100.0^a
	CV									70.4^a	77.6^a	66.6^a	69.3^a	64.7^a	71.2^a	74.8^a
	LCV									98.9^a	100.0^a	99.5^a	100.0^a	100.0^a	99.6^a	98.5^a
A_{10}	L										89.6^a	99.4^a	99.3^a	97.5^a	97.4^a	95.9^a
	C										55.4	56.7^a	66.9^a	66.4^a	76.7^a	70.2^a
	V										43.7	69.5^a	72.0^a	65.1^a	87.1^a	54.8
	LC										92.4^a	99.4^a	97.9^a	97.0^a	98.2^a	96.7^a
	LV										88.9^a	100.0^a	98.3^a	98.3^a	98.7^a	97.5^a
	CV										59.2^a	64.5	67.9^a	69.2^a	84.1^a	65.0^a
	LCV										94.3^a	99.4^a	98.8^a	98.2^a	99.5^a	97.1^a
A_{11}	L											95.4^a	94.4^a	98.3^a	95.3^a	89.4^a
	C											70.3^a	66.3^a	68.8^a	74.6^a	58.4
	V											78.9^a	83.2^a	84.1^a	92.4^a	66.1^a
	LC											95.3^a	93.6^a	96.4^a	95.7^a	82.1^a
	LV											97.8^a	94.4^a	97.9^a	96.2^a	88.6^a
	CV											78.8^a	77.2^a	78.6^a	85.6^a	61.1
	LCV											94.2^a	93.6^a	94.3^a	99.1^a	89.9^a
A_{12}	L												86.1^a	98.7^a	93.9^a	97.9^a
	C												62.4	49.5	69.8^a	74.1^a
	V												58.0^a	60.9	78.1^a	71.5^a
	LC												94.8^{ab}	98.3^a	94.5^a	97.9^a
	LV												91.3^a	99.1^a	95.5^a	97.9^a
	CV												62.7	62.0	77.3^a	76.0^a
	LCV												91.2^a	99.6^a	94.0^a	97.9^a

[a] Significantly better than naive guessing ($p < .05$)

[b] Significantly better than lexico-syntactic features ($p < .05$)

Table 10: Accuracy scores (%) of pairwise classification experiments. Authors $A_9 - A_{12}$.

		A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}
A_{13}	L													97.4^a	95.3^a	98.2^a
	C													56.6^a	77.4^a	76.5^a
	V													55.4	78.1^a	73.9^a
	LC													97.5^a	96.0^a	98.9^a
	LV													98.0^a	94.2^a	98.1^a
	CV													63.0^a	80.0^a	80.5^a
	LCV													97.3^a	95.9^a	98.0^a
A_{14}	L														98.8^a	93.6^a
	C														70.1^a	68.4^a
	V														79.9^a	66.6^a
	LC														99.6^a	97.2^a
	LV														99.6^a	96.0^a
	CV														74.5^a	73.6^a
	LCV														100.0^{ab}	95.3^a
A_{15}	L															97.5^a
	C															79.4^a
	V															83.5^a
	LC															99.2^a
	LV															99.2^a
	CV															86.3^a
	LCV															97.8^a
A_{16}	L	98.6^a	96.9^a	91.7^a	89.0^a	96.1^a	94.0^a	96.2^a	99.6^a	95.9^a	89.4^a	97.9^a	98.2^a	93.6^a	97.5^a	
	C	73.7^a	74.8^a	66.7^a	62.2^a	78.1^a	71.3^a	74.8^a	73.4^a	70.2^a	58.4	74.1^a	76.5^a	68.4^a	79.4^a	
	V	62.9^a	57.1	51.6	66.7	65.2^a	53.9	66.6^a	69.4^a	54.8	66.1^a	71.5^a	73.9^a	66.6^a	83.5^a	
	LC	97.9^a	97.7^a	90.6^a	90.8^a	97.6^a	94.8^a	96.5^a	100.0^a	96.7^a	82.1^a	97.9^a	98.9^a	97.2^a	99.2^a	
	LV	98.3^a	98.6^a	90.0^a	92.3^a	97.3^a	94.8^a	97.0^a	100.0^a	97.5^a	88.6^a	97.9^a	98.1^a	96.0^a	99.2^a	
	CV	70.9^a	76.5^a	66.6^a	63.0^a	73.8^a	73.9^a	77.6^a	74.8^a	65.0^a	61.1	76.0^a	80.5^a	73.6^a	86.3^a	
	LCV	97.5^a	98.6^a	91.5^a	92.7^a	96.1^a	96.0^a	96.7^a	98.5^a	97.1^a	89.9^a	97.9^a	98.0^a	95.3^a	97.8^a	

[a] Significantly better than naive guessing ($p < .05$)

[b] Significantly better than lexico-syntactic features ($p < .05$)

Table 11: Accuracy scores (%) of pairwise classification experiments. Authors $A_{13} - A_{16}$.

B Team Discussion

Yong Sik Cho: Background research, aggregate corpus statistics, added some lexical features consistent with previous research, experiment design, research paper write up, export and organize statistical analysis for paper, presentation drafting and design

Samuel Sharpe: Majority of feature extraction, neural network design, coreference resolution feature design, model training and tuning, experiment design, research paper write up, aggregate accuracy statistics

Harry Smith: Research on linguistics, voice feature design and extraction, experiment design, background research, research paper write up, export and analyze statistical analysis for paper