Literary and Linguistic Computing Advance Access published April 9, 2013

Patterns of local discourse coherence as a feature for authorship attribution

Vanessa Wei Feng and Graeme Hirst
University of Toronto, Canada

Correspondence: Graeme Hirst, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 3G4.
E-mail:
gh@cs.toronto.edu

Abstract

We define a model of discourse coherence based on Barzilay and Lapata's entity grids as a stylometric feature for authorship attribution. Unlike standard lexical and character-level features, it operates at a discourse (cross-sentence) level. We test it against and in combination with standard features on nineteen book-length texts by nine nineteenth-century authors. We find that coherence alone performs often as well as and sometimes better than standard features, though a combination of the two has the highest performance overall. We observe that despite the difference in levels, there is a correlation in performance of the two kinds of features.

1 Introduction

Contemporary methods of authorship attribution and discrimination that are based on text classification algorithms invariably use low-level within-sentence aspects of the text as stylometric features. In his recent survey, for example, Stamatatos (2009) lists twenty types of stylometric features that mostly involve character and word unigrams and *n*-grams, part-of-speech tags, and syntactic chunks and parse structures. Koppel, Schler, and Argamon (2009) adduce a similar set. In this article, we experiment with a stylometric feature that, by contrast, is drawn from the discourse level, across sentences: patterns of local coherence. We look at the choice that a writer makes when referring to an entity as to which grammatical role in the sentence the reference will appear in, and how this choice changes over a sequence of sentences as the writer repeatedly refers to the same entity. We show that differences in the resulting patterns of reference and grammatical role are a powerful stylometric feature for identifying an author.

2 Barzilay and Lapata's discourse entity grid

The basis of our method is the discourse entity grid, introduced by Barzilay and Lapata (2008) ('B&L' hereafter) as a model of local discourse coherence. B&L's model is based on the assumption that a text naturally makes repeated reference to the elements of a set of entities that are central to its topic. It represents local coherence as a sequence of transitions, from one sentence to the next, in the grammatical role of these references.

Consider, for example, this text:

(1) *(a)* Thus encouraged, Oliver tapped at the study door. *(b)* On Mr. Brownlow calling to him to come in, he found himself in a little back room, quite full of books, with a window, looking into some pleasant little gardens. *(c)* There was a table drawn up before the window, at which Mr. Brownlow was seated reading. *(d)* When he saw Oliver, he pushed the book away from him, and told him to come near the table, and sit down.[1]

doi:10.1093/llc/fqt021

The entity *Oliver* makes the following sequence of grammatical roles: it is the subject of *tapped* in *(a)*; it is both the subject and object of *found* and the object of the prepositional phrase *to him* in *(b)*; it does not appear in *(c)*; and it is an object (twice, of *saw* and of *told*) in *(d)*. As this example demonstrates, it is the referent entity itself that matters, not the form of the reference as, for example, *Oliver*, *him*, *himself*, or (not in this example) *the boy* are all the same entity, whereas the *books* of sentence *(b)* are not the same entity as the *book* of sentence *(d)*. When an entity appears in more than one role in a sentence, it is assigned only the 'highest ranking' role (Barzilay and Lapata, 2008); subject outranks object, which in turn outranks all others. Thus in sentence *(b)*, we assign (only) the subject role to *Oliver*. The sequence for the entity *Oliver* in this four-sentence text is thus [**S S – O**], where '–' indicates that the entity does not occur in the sentence; hence the sentence-to-sentence transitions made by this entity in the text are subject to subject ([**S S**]), subject to not-mentioned ([**S –**]), and not-mentioned to object ([**– O**]).

In the case of passive verbs, the surface-form subject is assumed to take the grammatical object role. In the example text, the *table*, which does not appear in the first two sentences, is construed as the grammatical object of the passive verb *drawn up* rather than its subject in *(c)*,[2] and it then appears in a prepositional phrase in *(d)*. Thus its sequence is [**– – O X**], where '**X**' indicates a role that is neither subject nor object, and the transitions (in addition to the empty transition [**– –**]) are [**– O**] and [**O X**].

Because a complete sentence, possibly containing several clauses, is considered at a single time in B&L's model, a sentence may have several entities in subject and object roles in the different clauses, and even a simple clause may refer to more than one entity within a single grammatical role. It is also possible, of course, that an entity will be mentioned only once in the entire text (such as *the gardens* in example (1) above) and thus not contribute to local coherence at all. B&L consider an entity to be **salient** if it is mentioned in the text at least twice (or some other threshold), and they describe models in which transitions for salient and non-salient entities are treated separately.

Table 1 The grid for the salient entities of example (1)

	Oliver	Mr. Brownlow	Window	Table
(a)	S	–	–	–
(b)	S	S	X	–
(c)	–	S	X	O
(d)	O	S	–	X

More formally, we can represent a document d as an **entity grid** in which the columns represent the entities referred to in d, and rows represent the sentences.[3] Each cell corresponds to the grammatical role of an entity in the corresponding sentence: subject (**S**), object (**O**), neither (**X**), or nothing (**–**). Each column is a complete sequence of roles. An example of an entity grid for the salient entities of example (1) is shown in Table 1.

B&L define a **local transition** as a sequence $\{\mathbf{S}, \mathbf{O}, \mathbf{X}, -\}^n$, representing the occurrence and grammatical roles of an entity in n adjacent sentences. In our example earlier, we took $n = 2$ for transitions, but we could use higher values; for example, for $n = 3$, *Oliver* has the local transition [**S – O**] from sentences *(b)* to *(d)*. Clearly, these transition sequences can be extracted from the entity grid as continuous subsequences in each column. For example, the entity *Mr. Brownlow* in Table 1 has a bigram transition [**S S**] from sentence *(b)* to *(c)*. For a given value of n, 4^n different local transitions are possible. We can count the number of times that each one occurs in a given text and thus compute the proportion of transitions that are of each type.

We can interpret these proportions as probabilities—the probability that any randomly chosen transition of length n in the text is of the given type. This lets us encode the entity grid as a feature vector $\mathbf{\Phi}(d) = (p_1(d), p_2(d), \ldots, p_m(d))$, where $p_t(d)$ is the probability of transition type t in the entity grid, computed as the number of occurrences of t in the entity grid of text d divided by the total number of transitions of length n in the entity grid, and m is the total number of transition types considered. For example, if we take transitions of length $n = 2$, then $m = 16$, and if $n = 3$, then $m = 64$. We have $\sum_{t=1}^{m} p_t(d) = 1$.

B&L evaluated the use of entity grids as a model of local coherence by showing that they can be used to discriminate original news texts from random permutations of the same sentences by learning a pairwise ranking preference between alternative renderings of a document based on the probability distribution of the entity-grid transitions. The model can also improve the performance of a text readability assessment system. Here, we will use entity grids in quite a different manner—not to measure degree of local coherence but rather, assuming the presence of coherence, to look at the different ways in which it is achieved.

3 Local transitions as features for authorship attribution

Because the set of transition probabilities forms a feature vector $\mathbf{\Phi}$ for a text, we can investigate it as a possible stylometric feature for authorship attribution. That is, we hypothesize that authors differ in the patterns of local transitions that they use—their tendency to use some types of local transitions rather than others—and that this forms part of an author's unconscious stylistic signature. We can test this hypothesis by seeing whether we can use these feature vectors to build a classifier to identify authorship.

Of course, we are not suggesting that local transitions by themselves are sufficient for high-accuracy authorship attribution. Rather, we are hypothesizing only that they carry information about authorship, which, if correct, is an interesting new stylometric observation. And they could thus be a useful complement to lexical and syntactic features for attribution. In our experiments below, therefore, we put our emphasis on examining how much authorship information these features carry by themselves in comparison with a simple baseline as well as with traditional lexical and lexico-syntactic features and in combination with such features.

Any implementation of entity grids must deal with the question of how entities are identified and tracked through the text. Entity recognition and coreference resolution remain incompletely solved problems in computational linguistics. B&L experimented with two approximations: the use of an imperfect coreference resolution tool (Ng and Cardie, 2002) and a very simple string-matching algorithm. Unsurprisingly, the former gave better results (which was also our experience in our own earlier work with entity grids (Feng and Hirst, 2012)). Accordingly, we use a coreference tool here too (see Section 4.1 below).

4 Experiments

4.1 Data

We gathered nineteen works of nine nineteenth-century British and American novelists and essayists from Project Gutenberg; they are listed in Table 2. We split the texts at sentence boundaries into chunks of approximately 1,000 words, regarding each chunk as a separate document by that author; leftovers of fewer than 1,000 words were discarded. The imbalance between different authors in the number of documents for each is corrected by the sampling methods in our experiments (see Section 4.2 below).

We applied coreference resolution to each document, using Reconcile 1.1 (Ng and Cardie, 2002). Reconcile is a learning-based end-to-end coreference resolution system that outputs the entities (noun phrases) and the coreference chains formed by these entities. It achieves F_1 scores of 60 to 70 on several coreference benchmark datasets. We then obtained a dependency parse of each sentence of each document to extract the grammatical role of each entity in the text, using the Stanford dependency parser (de Marneffe, MacCartney, and Manning, 2006). We could then construct an entity grid and the corresponding coherence feature vector for each document. We took $n = 2$, that is only transition bigrams, so there are $4^2 = 16$ transition types. But we counted transitions separately for salient entities—those entities with at least two occurrences in a document—and for non-salient entities. In the latter case, only seven of the sixteen transition types—those in which the entity appears in at most one sentence—can occur.

For each document, we also extracted a set of 208 low-level lexico-syntactic stylistic features: the

Table 2 The data used in our experiments

Text	Chunks
Anne Brontë	
Agnes Grey	78
The Tenant of Wildfell Hall	183
Jane Austen	
Emma	183
Mansfield Park	173
Sense and Sensibility	140
Charlotte Brontë	
Jane Eyre	167
The Professor	92
James Fenimore Cooper	
The Last of the Mohicans	156
The Spy	103
Water Witch	164
Charles Dickens	
Bleak House	383
Dombey and Son	377
Great Expectations	203
Ralph Waldo Emerson	
The Conduct of Life	67
English Traits	68
Emily Brontë	
Wuthering Heights	126
Nathaniel Hawthorne	
The House of the Seven Gables	106
Herman Melville	
Moby Dick	261
Redburn	27

frequencies of the 100 most frequent letter bigrams, of the 100 most frequent letter trigrams, and of the following eight types of function words: prepositions, pronouns, determiners, conjunctions, modal auxiliaries, auxiliary verbs (*be, have, do*), adverbs, and *to*.

4.2 Method

We conducted two sets of authorship attribution experiments: pairwise and one-versus-others. In the former, we select two authors and build a classifier that attempts to discriminate them, using either the coherence feature set, the lexico-syntactic feature set, or a combination of both. In the latter, we select one author and build a classifier that attempts to discriminate that author from all others in the dataset, again using one or both of the feature sets. The classifier in all experiments was a neural network with one hidden layer. We chose this classifier because it is able to handle non-linear relations among features, and it outperformed some other classifiers, such as decision trees and support vector machines, in our development experiments.

Each experiment used five-fold cross-validation, in which one-fifth of the data are chosen as test data and the training data are derived from the other four-fifths. The process is repeated for each one-fifth of the data in turn. To prevent class imbalance, the training data partitions are resampled (specifically, oversampled) in each iteration (Estabrooks, Jo, and Japkowicz, 2004) to obtain a balanced distribution in the training set between the two classes—that is, between the two selected authors in the pairwise experiments and between the selected author and all the others in the one-versus-others experiments. If the number of datapoints in one class is markedly fewer than that of the other, this procedure (implemented here by the Resample module of the Weka 3.6.8 toolkit (Hall *et al.*, 2009)) replicates datapoints at random until the two classes are approximately equal in size. For pairwise classification, we also oversample the test set in the same way in order to set an appropriate baseline.

4.3 Results

4.3.1 *Pairwise classification*

The results of pairwise classification are shown in Table 3. For each pair of authors, we show macro-averaged classification accuracy for four conditions: using only coherence features, using only traditional lexico-syntactic features, using all features, and, as a baseline, always guessing the author that has the greater representation in the training data. Because of our resampling procedure, the baseline is always close to, but not exactly, 50%, and it will be less than 50% when the more frequent author in the training data is not the more frequent one in the test data. Significant differences between the conditions for each pair of authors are indicated by superscripts ($p < .05$ in all cases).

As expected, the established lexico-syntactic stylometric features give accuracies that are significantly above the baseline (with one exception: Hawthorne versus Melville, where these features

Table 3 Accuracy scores (%) of pairwise classification experiments

		Austen	Charlotte	Cooper	Dickens	Emerson	Emily	Hawthorne	Melville
Anne	Coherence	78.3[d]	73.8[d]	88.4[d]	83.7[a,d]	85.5[d]	81.9[a,d]	81.0[d]	83.9[d]
	Lexico-syntactic	79.4[d]	77.7[d]	98.1[b,d]	78.4[d]	90.6[b,d]	71.5[d]	85.3[d]	90.7[b,d]
	Combined	86.7[a,b,d]	83.1[a,b,d]	99.1[b,d]	84.4[a,d]	90.9[b,d]	82.6[a,d]	90.1[a,b,d]	91.6[b,d]
	Baseline	53.3	57.7	53.6	52.5	47.1	48.0	46.9	44.8
Austen	Coherence		78.9[d]	82.3[d]	75.5[d]	74.9[d]	83.5[d]	71.9[d]	78.6[d]
	Lexico-syntactic		85.3[b,d]	97.5[b,d]	84.1[b,d]	97.4[b,d]	86.2[d]	90.3[b,c,d]	91.5[b,d]
	Combined		91.7[a,b,d]	97.5[b,d]	86.4[a,b,d]	98.3[b,d]	93.9[a,b,d]	85.3[b,d]	90.5[b,d]
	Baseline		47.2	49.3	53.6	45.4	45.2	46.1	46.4
Charlotte	Coherence			93.2[d]	80.6[a,c,d]	84.0[d]	78.8[a,c,d]	84.1[a,c,d]	82.9[a,d]
	Lexico-syntactic			92.8[d]	58.1[d]	93.3[b,d]	62.3[d]	73.7[d]	76.1[d]
	Combined			97.2[a,b,d]	77.3[a,d]	93.0[b,d]	73.4[a,d]	77.2[d]	83.8[a,d]
	Baseline			53.1	52.8	44.7	47.1	48.1	45.6
Cooper	Coherence				81.3[d]	79.6[d]	92.9[d]	73.5[d]	74.2[d]
	Lexico-syntactic				92.8[b,d]	87.1[b,d]	93.3[d]	81.0[b,d]	84.9[b,d]
	Combined				95.3[a,b,d]	88.5[b,d]	96.6[a,b,d]	83.4[b,d]	87.7[a,b,d]
	Baseline				53.7	44.6	46.0	46.1	46.0
Dickens	Coherence					86.5[d]	91.1[a,c,d]	75.2[d]	79.9[d]
	Lexico-syntactic					87.3[d]	77.8[d]	74.3[d]	83.2[b,d]
	Combined					94.3[a,b,d]	87.3[a,d]	77.6[a,d]	85.7[a,b,d]
	Baseline					48.8	49.1	49.6	49.0
Emerson	Coherence						97.5[d]	70.9[d]	75.6[d]
	Lexico-syntactic						97.5[d]	80.6[b,d]	89.9[b,d]
	Combined						95.1[d]	88.8[a,b,d]	94.2[a,b,d]
	Baseline						51.5	53.6	51.6
Emily	Coherence							78.9[d]	94.6[a,d]
	Lexico-syntactic							88.3[b,d]	86.0[d]
	Combined							91.1[b,d]	92.1[a,d]
	Baseline							58.3	52.5
Hawthorne	Coherence								67.9[a,d]
	Lexico-syntactic								58.5
	Combined								72.2[a,d]
	Baseline								51.3

[a] Significantly better than lexico-syntactic features ($p < .05$).
[b] Significantly better than coherence features ($p < .05$).
[c] Significantly better than combined features ($p < .05$).
[d] Significantly better than baseline ($p < .05$).

perform only seven percentage points above baseline, a difference that is not significant because we have relatively little data for Hawthorne). And, as we hypothesized, our coherence features also significantly exceed the baseline in all cases, showing that these features contain a considerable amount of information about authorship. The combined feature set also significantly exceeds the baseline in all cases. However, there is no consistency or pattern as to the relative performance of the three feature sets. In some cases (denoted by [a] in the table), such as Dickens versus Anne Brontë, the coherence features outperform the lexico-syntactic features, whereas in others (denoted by [b]), such as Dickens versus

Table 4 Accuracy scores (%) of pairwise classification experiments, aggregated across all authors for each feature set

Feature set	Acc. (%)
Coherence	81.3^{d}
Lexico-syntactic	$83.7^{b,d}$
Combined	$88.4^{a,b,d}$
Baseline	49.8

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

Austen, the reverse is true. In a few cases (denoted by [c]), such as Charlotte Brontë versus Hawthorne, the coherence features also outperform the combined feature set, although the converse is more usual. It is perhaps notable that, apart from Hawthorne versus Melville, all the pairs in which coherence features are superior to lexico-syntactic features involve a Brontë sister versus Dickens, Hawthorne, Melville, or another Brontë sister.

Generally speaking, however, Table 3 suggests that coherence features perform well but usually not quite as well as lexico-syntactic features, and that the combined feature set usually performs best. We confirm this generalization by aggregating the results for all pairs of authors by taking all predictions for all author pairs as a single set, and reporting accuracy for each set of features. The results, in Table 4, show this stated ordering for accuracy, with significant differences at each step.

4.3.2 *One-versus-others classification*

Unlike pairwise classification, where we are interested in the performance of both classes, in one-versus-others classification we are interested only in a single author class, and hence we can regard the problem as retrieval and report the results using the F_1 score of each class under each condition.

Table 5 F_1 scores of one-class classification experiments

Author	Features	F_1
Anne	Coherence	34.5^{d}
	Lexico-syntactic	$40.9^{b,d}$
	Combined	$48.0^{a,b,d}$
	Baseline	15.1
Austen	Coherence	39.1^{d}
	Lexico-syntactic	$60.3^{b,d}$
	Combined	$65.9^{a,b,d}$
	Baseline	26.3
Charlotte	Coherence	$37.2^{a,d}$
	Lexico-syntactic	25.2^{d}
	Combined	$38.0^{a,b,d}$
	Baseline	16.2
Cooper	Coherence	53.8^{d}
	Lexico-syntactic	$73.3^{b,d}$
	Combined	$78.0^{a,b,d}$
	Baseline	25.8
Dickens	Coherence	$63.8^{a,d}$
	Lexico-syntactic	61.3^{d}
	Combined	$70.6^{a,b,d}$
	Baseline	50.7
Emerson	Coherence	26.1^{d}
	Lexico-syntactic	$51.9^{b,d}$
	Combined	$61.6^{a,b,d}$
	Baseline	9.1
Emily	Coherence	29.5^{d}
	Lexico-syntactic	25.6^{d}
	Combined	$32.7^{a,b,d}$
	Baseline	7.9
Hawthorne	Coherence	17.3^{d}
	Lexico-syntactic	14.0^{d}
	Combined	$21.0^{a,b,d}$
	Baseline	7.2
Melville	Coherence	25.9^{d}
	Lexico-syntactic	$38.2^{b,d}$
	Combined	$39.6^{b,d}$
	Baseline	12.1

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

The results for each author are shown in Table 5, and aggregated results are shown in Table 6. A pattern similar to that of pairwise classification is observed. All feature sets perform significantly better than baseline, and the combined feature set is always significantly better than both others. The coherence features perform well, and significantly better than the lexico-syntactic features for Dickens and Charlotte Brontë; in addition, they perform notably better, albeit not significantly so, for Hawthorne and Emily Brontë. However, in aggregation, the lexico-syntactic features are significantly

Table 6 F_1 scores of one-class classification experiments aggregated across all authors for each feature set

Feature set	Acc. (%)
Coherence	40.5[d]
Lexico-syntactic	47.9[b,d]
Combined	56.6[a,b,d]
Baseline	20.0

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

better than coherence features, and the combined set is significantly better again.

5 Discussion

Our experiments show that entity-based local coherence, by itself, is informative enough to be able to classify texts by authorship almost as well as conventional lexico-syntactic information, even though it uses markedly fewer features. And the two types of information together perform better than either alone. This shows that local coherence does not just represent a subset of the same information as lexico-syntactic features, which is not a surprise, given that they focus on different aspects of the text at different levels. On the other hand, given this point, we might expect that the performance of the two feature sets would be independent, and that there would be authorship discriminations that are difficult for one set of features but easy for the other. However, we did not find this; while each had cases in which it was significantly better than the other, the scores for the two feature sets were correlated significantly (pairwise task, $r = .3657$, $df = 34$, $p < .05$; one-versus-others task, $r = .7388$, $df = 7$, $p < .05$).

Although in aggregation the combination of lexico-syntactic and coherence feature sets outperformed both feature sets individually, in a handful of cases, such as Hawthorne versus Austen in Table 3, the combined features obtained lower accuracy than using either lexico-syntactic or coherence features alone. We speculate that this is due to potential overfitting on the training data when using the combined feature set, which has a higher dimensionality than the other two.

6 Conclusion

We have shown that an author's presumably unconscious choices of transition types in entity-based local coherence is a stylometric feature. It forms part of an author's stylistic 'signature', and is informative in authorship attribution and discrimination. As a stylometric feature, it differs markedly from the syntactic, lexical, and even character-level features that typify contemporary approaches to the problem: It operates at the discourse level, above that of sentences. Nonetheless, the correlation that we found between the performance of this feature set and the lower-level feature set suggests that there is some latent relationship between the two, and this requires further investigation.

We took Barzilay and Lapata's model very much as they originally specified it. However, there are many possible variations to the model that bear investigation. Apart from the obvious parameter settings—the value of n, the threshold for salience—the ranking of grammatical roles when an entity occurs more than once in a sentence may be varied. In particular, we experimented with a variation that we called 'multiple-transition mode' in which each occurrence of an entity in a sentence is paired individually in all combinations. For example, if a specific entity occurs three times, in the roles **S**, **O**, and **X**, in one sentence and then twice, in the roles **S** and **O** in the next, then we extract six transition bigrams ([**S S**], [**O S**], [**X S**], [**S O**], [**O O**], and [**X O**]) rather than just the one with the highest priority, [**S S**]. However, in some of our early experiments, this variation showed little difference in performance from Barzilay and Lapata's single-transition model, so we abandoned it, as it adds significant complexity to the model. But we suspect that it might be more useful for characterizing authors, such as Dickens, who tend to write very long sentences involving complicated interactions among entities within a single sentence.

We see entity-based local coherence as possibly the first of many potential stylometric features at this level that may be investigated. These could include other aspects of repeated reference to entities. The present features consider only identical referents; but an author's use of other referential relationships such as various forms of meronymy (*the car ... the wheels; the government ... the minister*) could also be investigated. More generally, the set of various kinds of relationships used in lexical chains (Morris and Hirst, 1991) could be tried, but a too-promiscuous set would simply result in almost everything being related to almost everything else. Rather, instead of holding the relationships constant, as in the present approach, one would look for inter-author differences in the patterns of variation in the relationships. Such differences could also be sought in the surface forms of entity coreference across sentences (the author's choice of definite description, name, or pronoun); these forms are conflated in the present approach.

Understanding these kinds of discourse-level features better may bring new insight to this level of the stylistic analysis of text and the individuation of authors' style.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: an entity-based approach. *Computational Linguistics*, **34**(1): 1–34.

De Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, 449–54.

Feng, V. W. and Hirst, G. (2012). Extending the entity-based coherence model with multiple ranks. *Proceedings, 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012)*, Avignon, France, 315–24.

Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, **20**(1): 18–36.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, **11**(1): 10–18.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**(1): 9–26.

Morris, J. and Hirst, G. (1991). Lexical cohesion, the thesaurus, and the structure of text. *Computational Linguistics*, **17**(1): 21–48.

Ng, V. and Cardie, C. (2002). Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, 104–11.

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

Notes

[1] Charles Dickens, *Oliver Twist*, chapter XIV.

[2] To be precise: *table* is the elided surface-form grammatical subject of the reduced relative clause of which the passive verb *drawn up* is the head.

[3] The remainder of this section is based in part on material from Feng and Hirst (2012).