

Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution

SESSION

ORGANIZER: Harald Baayen

*Max Planck Institute for Psycholinguistics, P.O.
Box 310, 6500 AH, Nijmegen, The Netherlands.*

AUTHORS:

Hans van Halteren
Fiona Tweedie
Harald Baayen

KEYWORDS: authorship attribution, syntactic annotation, principal components analysis, hapax legomena, function words

AFFILIATION: Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

E-MAIL: baayen@mpi.nl
FAX NUMBER: +31-24-3521213
PHONE NUMBER: +31-24-3521323

Abstract

This session describes an experiment in authorship attribution in which statistical measures and methods that have been widely applied to words and their frequencies of use are applied to rewrite rules as they appear in a syntactically annotated corpus. The outcome of this experiment suggests that the frequencies with which syntactic rewrite rules are put to use provide at least as good a cue to authorship as word usage. Moreover, one method, which focuses on the use of the lowest-frequency syntactic rules, has a higher resolution than traditional word-based analyses, and promises to be a useful new technique for authorship attribution.

Introduction

A number of recent contributions to authorship attribution are based on words and their frequencies of occurrence (see, e.g., Burrows 1992, 1993; Holmes, 1994; Holmes and Forsyth 1995). This comes as no surprise, as the statistical analysis of word frequencies requires minimal textual preprocessing. Nevertheless, precisely those words which have proved to have a high discriminatory resolution in the seminal work by Burrows (1992, 1993), the so-called function words (a, the, that, and, but, ..., etc.), appear to tap into the use of syntax. This suggests it might be profitable to

study the use of syntax directly by analyzing the use of rewrite rules in texts.

We have designed a statistical experiment using syntactically annotated corpus material to investigate the discriminatory potential of syntactic rewrite rules for authorship attribution. The corpus, its syntactic annotation, and the details of the design of our statistical experiment, are discussed in section 1 by van Halteren. In section 2, Tweedie discusses the accuracy of methods based on measures for vocabulary richness and of methods based on the highest-frequency elements, applied both to words and rewrite rules. In section 3, Baayen investigates the discriminatory potential of the way in which authors make use of the lowest-frequency rewrite rules.

Before going into further detail, we need to make explicit three crucial details of our methodology. First, traditionally, as in the study by Mosteller and Wallace (1964), a text of unknown authorship is compared with texts of which authorship is beyond doubt. In our experiment, the authorship of all texts is known (be it only to the experiment leader, van Halteren, and not to Tweedie and Baayen, who carried out the analyses). This allows us to straightforwardly evaluate the accuracy of the methods we have used. Second, a preliminary pilot study shows that texts written by one author in different genres can differ more than texts written by different authors in the same genre. We have therefore selected our texts from one particular text type, crime fiction. Third, to ensure the accuracy of assignment is independent of our particular split in labeled and unlabeled text fragments, we also required that a successful method should group all text fragments of different authors into clearly distinguishable clusters.

References

- Burrows, J. F., (1992). Computers and the Study of Literature. In: C.S. Butler (Ed.), *Computers and Written Texts*. Oxford: Blackwell. (pp. 167–204).
- Burrows, J. F., (1993). Tiptoeing into the infinite: testing for evidence of national differences in the language of English narrative. In: S. Hockey and N. Ide (Eds.), *Research in Humanities Computing '92*. London: Oxford University Press.
- Holmes, D. I., (1994). Authorship Attribution. *Computers and the Humanities* 28(2):87–106.
- Holmes, D. I. and Forsyth, R. S., (1995). The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing* 10(2):111–127.
- Mosteller, F. and Wallace, D. L., (1964). *Applied Bayesian and Classical Inference. The case of the Federalist Papers*. New York: Springer.

1. Experimental design: Syntactic annotation as words

Hans van Halteren

Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD, Nijmegen, The Netherlands.

AFFILIATION: University of Nijmegen, The Netherlands.

E-MAIL: cor_hvh@vms.uci.kun.nl

The first step in the setup of our experiment was the selection of suitable data. Because our focus was on the difference between authors rather than the difference between genres, we decided to take two samples from the Nijmegen corpus (Keulen 1986) which are of the same genre, crime fiction. The two samples each consisted of 20,000 words of running text and were taken from M. Innes' *The Bloody Wood* (henceforth Sample A) and M. Allingham's *The Mind Readers* (henceforth Sample B). The Nijmegen corpus has been syntactically annotated with two different analysis systems, the CCPP system (cf. Keulen 1986) and the TOSCA system (cf. Oostdijk 1991). The TOSCA analysis is the more detailed one and uses a more consistent description model. We therefore selected the TOSCA analysis as the one to be used in our experiment.

Figure 1 exemplifies the syntactic annotation assigned to the sentence "He walks his dog in the park". On each node in the analysis tree, we find labels for syntactic function, for syntactic category, and for additional attributes. Consider, for instance, the node immediately to the left of the word "park". For this node, the function is 'Noun Phrase

Head' (NPHD), the category is 'Noun' (N), and the attributes are 'Common' (com) and 'Singular' (sing).

There are many different aspects of syntactic analysis trees that might be exploited for purposes of authorship attribution. We opt to translate part of the information present in each analysis tree into a pseudo-word sequence. The crucial question is which part. We have used two criteria to decide which information to include: a) Focus on the most important information and b) try to keep the resulting pseudo-words as similar to normal words as possible, so that a greater accuracy for the syntax-based methods can only be attributed to a higher information content (with regard to the problem at hand) of the pseudo-words.

Criterion b) led us to exploit the individual rewrites (combinations of a node and its immediate constituents), since these are the building blocks of the tree, just as words are building blocks of the sentence. Criterion a) led us to focus first on the category label (e.g., NP, 'Noun Phrase'), then on the function label (e.g., SU, 'Subject') and only last on the attribute labels (e.g., sing) on the nodes. For an exact choice of the information to use we counted the number of pseudo-word tokens and types. The total number of rewrite tokens in the two samples is 46402. Using only the category labels, e.g.,

NP -> DTP + N

(where NP, N, and DTP denote 'Noun Phrase', 'Noun, and 'Determiner Phrase' respectively), leads to 2318 types. Adding the function labels at the right hand side, e.g.,

NP -> DET:DTP + NPHD:N

(where DET denotes 'Determiner' and NPHD 'Noun Phrase Head'), increases this number to

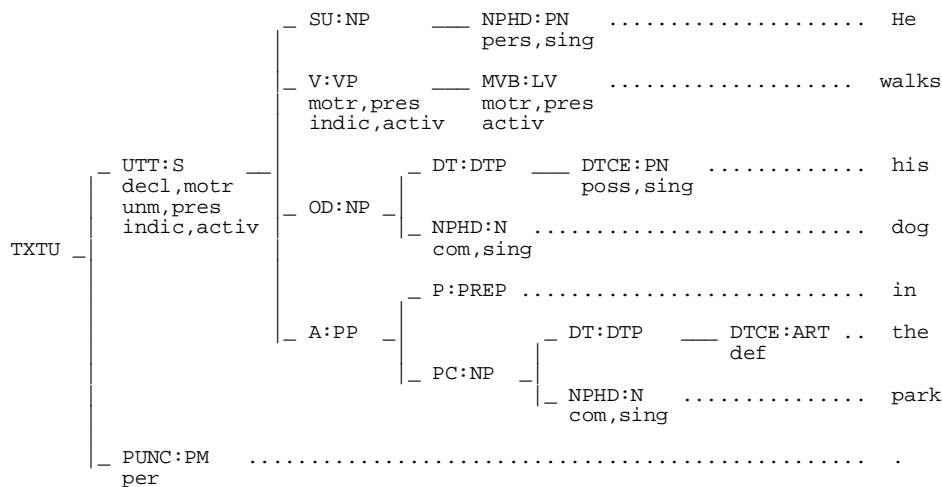


Figure 1: A Sample Analysis Tree.

2732. Addition of the function label on the left hand side as well,

PC:NP -> DET:DTP + NPHD:N

(where PC denotes ‘Prepositional Complement’), brings this number up to 4194. As the resulting type-token ratio is fairly close to that for the normal words of our samples, this is the labeling we have decided to use.

The most frequent rewrites are present in both samples. The first one missing in sample A is the 59th most frequent one,

UTT:COORD -> CJ:S + COOR:CONJN + CJ:S

(UTT: ‘Utterance’; COORD: ‘Coordination’; CJ: ‘Conjoin’; S: ‘Sentence’; COOR: ‘Coordinator’; CONJN: ‘Conjunction’) which occurs 85 times in sample B. The first one missing in sample B is the 231st most frequent one,

RPDU:CLOID -> DIFU:REACT + PUNC:PM + DIFU:REACT

(RPDU: ‘Reported Utterance’; CLOID: ‘Clausoid’; DIFU: ‘Discourse Function’; REACT: ‘Reaction Signal’; PUNC: ‘Punctuation’; PM: ‘Punctuation Mark’) which occurs 14 times in sample A. These simple numbers already suggest that there are marked differences in the way A and B make use of syntactic rewrite rules.

We have translated the syntactic rewrite information from the samples into pseudo-words. The main reason for this is that the existing software is likely to expect words rather than the complex (and long) expressions the rewrites are. For the translation, we have sorted the rewrites according to their frequency (cumulative over both samples) and we have named them accordingly. Thus, the most frequent rewrite becomes W0001, the second most frequent one W0002, etc.:

Translated Rewrite	Frequency	Rewrite
W0001	4670	V:VP -> MVB:LV
W0002	3566	SU:NP -> NPHD:PN
W0003	2674	DT:DTP -> DTCE:ART
W0004	1948	A:AVP -> AVHD:ADV
W0005	1729	A:PP -> P:PREP + PC:NP
W0006	1435	V:VP -> OP:AUX + MVB:LV
W0007	1395	NPPR:AJP -> AJHD:ADJ
W0008	1172	DT:DTP -> DTCE:PN
W0009	1017	PC:NP -> DT:DTP + NPHD:N
W0010	1016	-.TXTU -> UTT:S + PUNC:PM

[V: ‘Verb’; VP: ‘Verb Phrase’; MVB: ‘Main Verb’; LV: ‘Lexical Verb’; PN: ‘Pronoun’; DTCE: ‘Central Determiner’; ART: ‘Article’; A:

‘Adverbial’; AVP: ‘Adverb Phrase’; AVHD: ‘Adverb Phrase Head’; ADV: ‘Adverb’; PP: ‘Prepositional Phrase’; P: ‘Preposition’; PREP: ‘Preposition’; OP: ‘Operator’; AUX: ‘Auxiliary’; NPPR: ‘Noun Phrase Premodifier’; AJP: ‘Adjective Phrase’; AJHD: ‘Adjective Phrase Head’; ADJ: ‘Adjective’; TXTU: ‘Textual Unit’.]

The translated rewrites were presented in the original order in which they appear in the samples. In addition, text unit separators were inserted to indicate which pseudo-words together formed a pseudo-sentence (i.e., which rewrites jointly form an analysis tree). As a result, the experimenters received the following kind of data:

S W0084 W3165 W0048 S W0021 W0061 W0002 W0001 W0031 W0019 S W0010 ...

The unlabeled samples were provided to the experimenters in two different forms. For general statistical operations, the two complete pseudo-texts were available. For authorship attribution techniques, the two texts were available in the form of fourteen labeled samples and six unlabeled samples.

Unknown to the experimenters, the two pseudo-texts were both divided into ten parts, such that a new part was initiated at the first text unit separator after 2500 pseudo-words (including separators). All parts were about the same size, except for the tenth part of sample B, which contained only 2254 pseudo-words. The first seven parts of each pseudo-text were provided as labeled samples (SA1-SA7, and SB1-SB7). The remaining 6 parts were provided as unlabeled samples (SQ1 (=SA10), SQ2 (=SB10), SQ3 (=SB8), SQ4 (=SA8), SQ5 (=SB9) and SQ6 (=SA9)). All correspondence information was withheld from the experimenters. Without a-priori knowledge of how many of the six unlabeled samples Q1–6 should be attributed to A or B – this number can vary between zero and six – the probability of finding the correct assignment by chance equals $(1/2)^6 = 0.016$. The probability of correctly assigning at least five samples equals $7/64 = 0.109$. These probabilities show that our experiment is statistically non-trivial: the experimenters are not likely to arrive at the correct solution by chance.

References

- Keulen, F. (1986). The Dutch computer corpus pilot project. In: J. Aarts and W. Meijjs (Eds.), *Corpus Linguistics II. New studies in the analysis and exploitation of computer corpora*. Amsterdam: Rodopi.
- Oostdijk, N. (1991). *Corpus Linguistics and the Automatic Analysis of English*. Amsterdam: Rodopi.

2. Comparison of word-based and syntax-based methods: Vocabulary richness measures and the highest frequency elements

Fiona Tweedie

Department of Mathematical Sciences, University of the West of England, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, Great Britain

AFFILIATION: University of the West of England, Bristol, Great Britain

E-MAIL: fj-tweed@csm.uwe.ac.uk

To evaluate the possible advantages of using rewrite rules instead of words for authorship attribution, we carried out two kinds of comparisons. First, we compared the accuracy of methods based on statistics for vocabulary richness as applied to word counts on the one hand, and to frequency counts of rewrites on the other. Second, we also compared the accuracy of methods based on the counts of the highest-frequency elements, the 50 highest-frequency function words for the word-based analyses, and the 50 highest-frequency rewrites for the syntax-based approach. Various measures have been proposed throughout the history of stylometry. We have used a selection of them in our multivariate analyses, following Holmes and Forsyth (1995). Our first measure was proposed by Yule (1944). It is defined as:

$$K = 10^4 x \frac{\sum_{i=1}^v i^2 V(i, N) - N}{N^2}$$

with N the number of tokens, $V(i, N)$ the number of types which occur i times in a sample of N tokens, and v the highest frequency of occurrence. A related measure was proposed by Simpson (1949), who focused on the probability that two words randomly selected from the text are the same. His measure is defined as

$$D = \frac{\sum_{i=1}^v i(i-1)V(i, N)}{N(N-1)}$$

The values of both D and K are primarily determined by the high end of the frequency distribution structure. They quantify the repeat rate of the samples.

In order to consider the low frequency end of the distribution, we also include measures proposed

by Honoré (1979) and Sichel (1975). Honoré's measure,

$$R = \frac{100 \log N}{1 - V(1, N)/V(N)}$$

where $V(N)$ denotes the number of different rewrite types, was used initially to examine the vocabulary of Latin judicial authors and has subsequently been used by others including Holmes and Forsyth (1995). R takes into account the probability that the author will re-use a given type in the text rather than using a new one. It's dependence on $V(1, N)$, the number of hapax legomena, may add useful. Another measure that is sensitive to the low end of the frequency distribution was proposed by Sichel (1975):

$$S = V(2, N)/V(N)$$

By means of this measure we take the number of dis-legomena, the words which appear twice in the text, into account.

Finally, we included a variable which has measured vocabulary richness with success in various fields. Proposed by Brunet (1978), it is defined as:

$$W = N^{V(N)^a}$$

where a is a parameter, usually fixed at 0.17, such that W is approximately constant and independent of N .

Values for R , D , S , K and W were calculated for each of the twenty samples of our experiment, for both words and rewrites. In this way we obtained two (20,5) data matrices. A Principal Components Analysis of the word matrix revealed a misclassification rate of 2/6 for the unlabeled samples and a misclassification rate of 1/14 for the labeled samples. Considerably improved results were obtained on the basis of the rewrite matrix. All unlabeled samples were correctly assigned to their authors, and except for one labeled sample, the samples by A were clearly distinguishable from those by B. We conclude that methods based on measures of vocabulary richness are more accurate when applied to rewrites than when applied to words.

Following Burrows (1992), we also investigated the discriminatory potential of the 50 highest-frequency function words, and compared the result with an analysis based on the 50 most frequent rewrites. The two (20,50) data matrices were subjected to Principal Components Analysis. For the words, the labeled samples of A and B were well-separated into two distinct clusters. Five of the six unlabeled samples appeared in the correct clusters. One unlabeled sample, however, appeared exactly

half-way between the two clusters, and was equally likely in the analysis to be by A or B. The rewrite-based analysis, by contrast, correctly separated all samples of A and B. The unlabeled sample that could not be assigned with confidence in the word-based analysis to A or B now clearly sided with the cluster of samples by B. Again we find that a rewrite-based analysis leads to an improved classification.

Finally, note that methods based on the 50 most frequent elements appear to have a higher discriminatory potential than methods based on statistics of vocabulary richness. In the word-based analyses, changing from data on vocabulary richness to the data on the 50 most frequent function words led to a decrease in the misclassification rate from 3/20 to 1/20. Similarly, the misclassification rate dropped from 1/20 to 0/20 in the rewrite-based analyses. We conclude that optimal results may be expected for analyses based on the highest-frequency rewrite rules.

References

- Brunet, E., (1978). *Vocabulaire de Jean Giraudoux: Structure et Évolution*, Slatkine.
- Burrows, J. F., (1992). Computers and the Study of Literature. In: C.S. Butler (Ed.), *Computers and Written Texts*. Oxford: Blackwell. (pp. 167–204).
- Holmes, D. I. and Forsyth, R. S., (1995). The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing* 10(2):111–127.
- Honoré, A., (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 172–177.
- Juillard, M. (1990). Proper nouns as proper style markers of poetry and prose. *Literary and Linguistic Computing*, 5(1):1–8.
- Sichel, H. S., (1975). On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association*, 70:542–547.
- Simpson, E. H., (1949). Measurement of Diversity. *Nature* 163:168.
- Yule, G. U., (1944) *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

3. The discriminatory potential of the lowest frequency rewrite rules

Harald Baayen

*Max Planck Institute for Psycholinguistics,
Nijmegen, The Netherlands*

E-MAIL: baayen@mpi.nl

We have also pursued the hypothesis that further reliable and robust clues to authorship identity might be found among the hapax legomena, the rewrite rules with the lowest possible frequency of use. This hypothesis is grounded in two considerations.

First, units in the highest frequency ranges often have properties that are atypical for the population as a whole (see Baayen and Sproat, 1996). Second, since the likelihood of storage in memory increases with frequency of use, and since awareness builds on memory, it is in the highest frequency ranges that conscious and deliberate wording and syntactic phrasing may be expected. Taken jointly, these considerations suggest that the lowest frequency ranges might provide a clue to authorship that is less contaminated by conscious rhetorical manipulation and thematic structuring that we think may affect the higher-frequency units of analysis.

Among the low-frequency units, the hapax legomena, the units which occur once only, are of special interest. Good (1953) has shown that the likelihood of observing an unseen type is estimated by the ratio of hapax legomena to the total number of tokens: $V(1,N)/N$. In other words, $P(N) = V(1,N)/N$ estimates the rate at which new units appear, the rate at which the vocabulary of units increases. With respect to distributions of syntactic rewrite rules, this growth rate $P(N)$ estimates the probability that an author will produce a new rewrite rule that she/he has not yet used before. In other words, $P(N)$ taps into an author's syntactic creativity, and can be used to gauge how well an author has mastered the possibilities offered by the grammar.

Does $P(N)$ have a good discriminatory resolution for authorship attribution for our experiment? Text A appears to make a more productive use of syntax than text B, as both $V(N)$, the total number of different construction types, and $P(N)$ are significantly higher for A (2114, 0.090) than for B (1883, 0.074) (in both cases, $p < .001$, proportions test). Not surprisingly, this difference in construction richness carries over to the seven labeled samples

of A and B. After correcting for the differences in size of the twenty text samples, a classification tree analysis (Breiman, Friedman, Olshen, and Stone, 1984) on the basis of P(N) correctly assigns all unlabeled text samples. This positive result is counterbalanced by a rather imperfect classification of the labeled fragments. The same classification tree reveals a misclassification rate of 2/14 for the labeled samples. Interestingly, using V(N) instead of P(N), again corrected for differences in sample size, a misclassification rate for the labeled samples of 1/14 is obtained, and again all unlabeled samples are assigned to the correct authors.) Although P(N) and V(N) clearly capture important differences between our two authors, they are by themselves unable to satisfy the criteria we have set ourselves, namely, to obtain a classification with a misclassification rate of 0/20. To increase our sensitivity to author-specific differences in the use of the lowest-frequency rewrite

($j = 1, 2, \dots, 20$) denote the number of rewrite rules in text sample j belonging to set L_i that occur once only in sample j and that do not occur in any of the other text samples (a hapax legomenon occurring in sample j). Furthermore, let

$$rh_{i,j} = \frac{h_{i,j}}{\sum_{i=1}^{49} h_{i,j}}$$

be the relative frequency of unique hapax legomena in text j falling in category L_i with respect to the total number of unique hapax legomena in j summed over all 49 categories. The relative frequency $rh_{i,j}$ measures the extent to which the syntactic creativity unique to a particular author (or text sample) manifests itself in the i -th set of rewrite rules.

A Principal Components Analysis on the 20×49 matrix of relative frequencies $rh_{i,j}$ revealed the

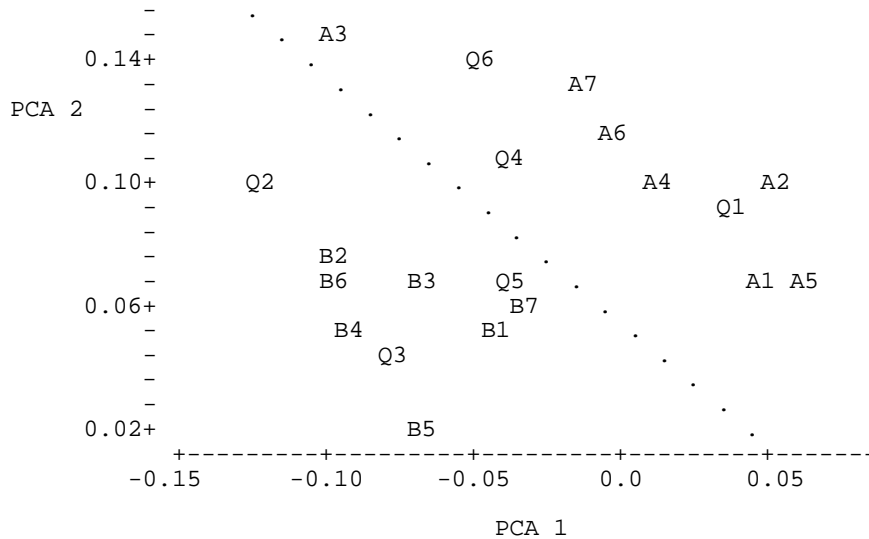


Figure 2. Results of a Principal Components Analysis on the relative frequencies $rh_{i,h}$ for the labeled samples of A (A1-7) and B (B1-7) and the unlabeled samples Q1-6. The division indicated by the dotted line is supported by a Discriminant Analysis on the space defined by the significant principal components.

rules, a subclassification of the hapax legomena is required. To do so, we sorted all rewrite rules, irrespective of their frequency, according to their left hand side, the information appearing to the left of the arrow in the rewrite rule. Some left hand sides L appear in a great many different rewrite rules, others appear in just a few rules. We selected the left-hand sides with more than 10 different right-hand sides for further analysis. There were 49 such left hand sides in the pooled twenty text fragments. Let L_i ($i = 1, 2, \dots, 49$) denote the set of rewrite rules with the i -th left hand side, and let $h_{i,j}$

pattern shown in Figure 2.

The first principal component is highly correlated with the left hand side UTT:S ('Utterance: Sentence', $r = 0.96$), the second principal component with the left hand sides CJ:CL ('Conjoin: Clause', $r = -0.66$), RPDU:S ('Reported Utterance: Sentence', $r = 0.63$), RPGT:S ('Reporting Tail: Sentence', $r = -0.63$) and V:VP ('Verb: Verb Phrase', $r = 0.63$). All unlabeled samples are correctly classified, and the samples by A and those by B also appear well-separated in Figure 2, a visual impression that is supported by a Discriminant Analysis.

This analysis again shows that syntactic annotation provides excellent clues for authorship attribution. In addition, a detailed comparison of word usage on the one hand with the use of syntax on the other (not reported here for lack of space) reveals that there is less variability in the use of syntax than in word usage. This suggests that syntax-based analyses are less likely to be foiled by idiosyncracies of individual samples.

Interestingly, the differences between the authors A and B reveal a consistent pattern. A has a greater vocabulary size than B, both with respect to words and with respect to syntactic constructions. Moreover, analyses that for lack of space have not been discussed here reveal that A makes more use of morphologically complex words than B (the adverbial suffix *-ly* also appears to be a reasonable classifier), and that narrative development in A is more complex than in B. Across the board, A reveals a more creative use of the possibilities of English. Since A, Innes, is a literary critic as well as a writer of crime fiction, this difference comes as no surprise.

The lowest-frequency rewrite rules provide a window to this difference in creativity. An analysis of their use has enabled us to tease apart the text samples written by Innes from text samples written by B, Allingham.

References

- Baayen, R. H. and Sproat, R., (1996). Estimating lexical priors for low-frequency morphologically ambiguous forms. To appear in *Computational Linguistics*.
- Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J., (1984). *Classification and Regression Trees*. Belmont, California: Wadsworth International Group.
- Good, I. J., (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* 40, 237–264.