A Comparative Study of Language Models for Book and Author Recognition

Özlem Uzuner and Boris Katz

MIT,Computer Science and Artificial Intelligence Laboratory,
Cambridge, MA 02139
{ozlem, boris}@csail.mit.edu

Abstract. Linguistic information can help improve evaluation of similarity between documents; however, the kind of linguistic information to be used depends on the task. In this paper, we show that distributions of syntactic structures capture the way works are written and accurately identify individual books more than 76% of the time. In comparison, baseline features, e.g., tfidf-weighted keywords, function words, etc., give an accuracy of at most 66%. However, testing the same features on authorship attribution shows that distributions of syntactic structures are less successful than function words on this task; syntactic structures vary even among the works of the same author whereas features such as function words are distributed more similarly among the works of an author and can more effectively capture authorship.

1 Introduction

Expression is an abstract concept that we define as "the way people convey particular content". Copyrights protect an author's expression of content where *content* refers to the information contained in a work and *expression* refers to the linguistic choices of authors in presenting this content. Therefore, capturing expression is important for copyright infringement detection.

In this paper, we evaluate *syntactic elements of expression* in two contexts: book recognition for copyright infringement detection and authorship attribution. Our first goal is to enable identification of individual books from their expression of content, even when they share content, and even when they are written by the same person. For this purpose, we use a corpus that includes translations of the same original work into English by different people. For the purposes of this study, we refer to the translations as *books* and an original work itself as a *title*.

Given the syntactic elements of expression, our second goal is to test them on authorship attribution, where the objective is to identify all works by a particular author. Our syntactic elements of expression capture differences in the way people express content and could be useful for authorship attribution. However, the experiments we present here indicate that syntactic elements of expression are more successful at identifying expression in individual books while function words are more successful at identifying authors.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 969–980, 2005.

2 Related Work

In text classification literature, similarity of works has been evaluated, for example, in terms of genre, e.g., novels vs. poems, in terms of the style of authors, e.g., Austen's novels vs. Kipling's novels, and in terms of topic, e.g., stories about earthquakes vs. stories about volcanoes. In this paper, we compare several different language models in two different classification tasks: book recognition based on similarity of expression, and authorship attribution. Authorship attribution has been studied in the literature; however, evaluation of similarity of expression, e.g., Verne's *20000 Leagues* vs. Flaubert's *Madame Bovary*, is a novel task that we endeavor to address as a first step towards copyright infringement detection.

We define expression as "the linguistic choices of authors in presenting content": content of works and the linguistic choices made while presenting it together constitute expression. Therefore, capturing expression requires measuring similarity of works in terms of both of these components.

To classify documents based on their content, most approaches focus on keywords. Keywords contain information regarding the ideas and facts presented in documents and, despite being ambiguous in many contexts, have been heavily exploited to represent content. In addition to keywords, subject–verb and verb–object relationships [12], noun phrases [12,13], synonym sets of words from WordNet [12], semantic classes of verbs [12] from Levin's studies [21], and proper nouns have all been used to capture content.

Linguistic choices of authors have been studied in stylometry for authorship attribution. Brinegar [7], Glover [9] and Mendenhall [22], among others, used distribution of word lengths to identify authors, e.g., Glover and Hirst studied distributions of two- and three-letter words [9]. Thisted et al. [33] and Holmes [14] studied the idea of richness of vocabulary and the rate at which new words are introduced to the text. Many others experimented with distributions of sentence lengths [9,18,24,30,31,32,38,40], sequences of letters [17,20], and syntactic classes (part of speech) of words [9,20,19].

Mosteller and Wallace [25] studied the distributions of function words to identify the authors of 12 unattributed Federalist papers. Using a subset of the function words from Mosteller and Wallace's work, Peng [26] showed that verbs (used as function words, e.g., be, been, was, had) are important for differentiating between authors. Koppel et al. [19] studied the "stability" of function words and showed that the features that are most useful for capturing the style of authors are "unstable", i.e., they can be replaced without changing the meaning of the text. Koppel et al.'s measure of stability identified function words, tensed verbs, and some part-of-speech tag trigrams as unstable.

Syntactically more-informed studies of the writings of authors came from diMarco and Wilkinson [39] who treated style as a means for achieving particular communicative goals and used parsed text to study the syntactic elements associated with each goal, e.g., clarity vs. obscurity. Adapting elements from Halliday and Hasan [10,11], diMarco et al. studied the use of cohesive elements of text, e.g., anaphora and ellipsis, and disconnective elements of text,

e.g., parenthetical constructions, as well as the patterns in the use of relative clauses, noun embeddings, and hypotaxis (marked by subordinating conjunctions) when authors write with different communicative goals.

Expression is related to both content and style. However, it is important to differentiate expression from style. Style refers to the *linguistic elements that, independently of content, persist over the works* of an author and has been widely studied in authorship attribution. Expression involves the *linguistic elements that relate to how an author phrases particular content* and can be used to identify potential copyright infringement.

3 Syntactic Elements of Expression

We hypothesize that, given particular content, authors choose from a set of semantically equivalent syntactic constructs to create their own expression of it. As a result, different authors may choose to express the same content in different ways. In this paper, we capture the differences in expression of authors by studying [34,35,36]:

- sentence-initial and -final phrase structures that capture the shift in focus and emphasis of a sentence due to reordered material,
- semantic classes and argument structures of verbs such as those used in START for question answering [16] and those presented by Levin [21],
- syntactic classes of embedding verbs, i.e., verbs that take clausal arguments, such as those studied by Alexander and Kunz [1] and those used in START for parsing and generation [15], and
- linguistic complexity of sentences, measured both in terms of depths of phrases and in terms of depths of clauses, examples of which are shown in Table 1.

Table 1. Sample sentences broken down into their clauses and the depth of the top-level subject (the number on the left) and predicate (the number on the right)

Sentence	**Depth of Clauses**
$[\text{I}]_a$ [would not think that $[\text{this}]_b$ [was possible$]_b]_a$	0, 2
$[\text{I}]_a$ [have found $[\text{it}]_b$ [difficult to say that $[\text{I}]_c$ [like it$]_c]_b]$ $_a$.	2, 2
[That $[\text{she}]_b$ [would give such a violent reaction$]_b]_a$ [was unexpected$]_a$.	1, 1
[For $[\text{her}]_b$ [to see this note$]_b]_a$ [is impossible$]_a$.	1, 1
[Wearing the blue shirt$]_a$ [was a good idea$]_a$.	1, 1
$[\text{It}]_a$ [is not known whether $[\text{he}]_b$ [actually libelled the queen$]_b]_a$.	0, 2
$[\text{He}]_a$ [was shown that [the plan$]_b$ [was impractical$]_b]_a$.	0, 2
[They$]_a$ [believed $[\text{him}]_b$ [to be their only hope$]_b]_a$.	0, 2
$[\text{I}]_a$ [suggest $[\text{he}]_b$ [go alone$]_b]_a$.	0, 2
$[\text{I}]_a$ [waited for $[\text{John}]_b$ [to come$]_b]_a$.	0, 2

We extracted all of these features from part-of-speech tagged text [5] and studied their distributions in different works. We also studied their correlations with each other, e.g., semantic verb classes and the syntactic structure of the alternation [21] in which they occur. The details of the relevant computations are discussed by Uzuner [34].

3.1 Validation

We validated the syntactic elements of expression using the chi-square (and/or likelihood ratio) test of independence. More specifically, for each of sentence-initial and -final phrase structures, and semantic and syntactic verb classes, we tested the null hypothesis that these features are used similarly by all authors and that the differences observed in different books are due to chance. We performed chi-square tests in three different settings: on different translations of the same title (similar content but different expression), on different books by different authors (different content and different expression), and on disjoint sets of chapters from the same book (similar content and similar expression).

For almost all of the identified features, we were able to reject the null hypothesis when comparing books that contain different expression, indicating that regardless of content, these features can capture expression. For all of the features, we were unable to reject the null hypothesis when we compared chapters from the same book, indicating a certain consistency in the distributions of these features throughout a work.

4 Evaluation

We used the syntactic elements of expression, i.e., sentence-initial and sentence-final phrase structures, semantic and syntactic classes of verbs, and measures of linguistic complexity [34,35,36], for book recognition and for authorship attribution.

4.1 Baseline Features

To evaluate the syntactic elements of expression, we compared the performance of these features to baseline features that capture content and baseline features that capture the way works are written. Our baseline features that capture content included tfidf-weighted keywords [27,28] excluding proper nouns, because for copyright infringement purposes, proper nouns can easily be changed without changing the content or expression of the documents and a classifier based on proper nouns would fail to recognize otherwise identical works. Baseline features that focus on the way people write included function words [25,26], distributions of word lengths [22,40], distributions of sentence lengths [14], and a basic set of linguistic features, extracted from tokenized, part-of-speech tagged, and/or syntactically parsed text. This basic set of linguistic features included the number of words and the number of sentences in the document; type–token ratio;

average and standard deviation of the lengths of words (in characters) and of the lengths of sentences (in words) in the document; frequencies of declarative sentences, interrogatives, imperatives, and fragmental sentences; frequencies of active voice sentences, be-passives, and get-passives; frequencies of 's-genitives, of-genitives, and of phrases that lack genitives; frequency of overt negations; and frequency of uncertainty markers [9,34].

4.2 Classification Experiments

We compared the syntactic elements of expression with the baseline features in two separate experiments: recognizing books even when some of them are derived from the same title (different translations) and recognizing authors. For these experiments, we split books into chapters, created balanced sets of relevant classes, and used boosted [29] decision trees [41] to classify chapters into books and authors. We tuned parameters on the training set: we determined that the performance of classifiers stabilized at around 200 rounds of boosting and we eliminated from each feature set the features with zero information gain [8,37].

Recognizing Books: Copyrights protect original expression of content for a limited time period. After the copyright period of a work, its derivatives by different people are eligible for their own copyright and need to be recognized from their unique expression of content. Our experiment on book recognition focused on and addressed this scenario.

Data: For this experiment, we used a corpus that included 49 *books* derived from 45 *titles*; for 3 of the titles, the corpus included multiple books (3 books for the title *Madame Bovary*, 2 books for *20000 Leagues*, and 2 books for *The Kreutzer Sonata*). The remaining titles included works from J. Austen, F. Dostoyevski, C. Dickens, A. Doyle, G. Eliot, G. Flaubert, T. Hardy, I. Turgenev, V. Hugo, W. Irving, J. London, W. M. Thackeray, L. Tolstoy, M. Twain, and J. Verne. We obtained 40–50 chapters from each book (including each of the books that are derived from the same title), and used 60% of the chapters from each book for training and the remaining 40% for testing.

Results: The results of this evaluation showed that the syntactic elements of expression accurately recognized books 76% of the time; they recognized each of the paraphrased books 89% of the time (see right column in Table 2). In either case, the syntactic elements of expression significantly outperformed all individual baseline features (see Table 2).

The syntactic elements of expression contain no semantic information; they recognize books from the way they are written. The fact that these features can differentiate between translations of the same title implies that translators add their own expression to works, even when their books are derived from the same title, and that the expressive elements chosen by each translator help differentiate between books derived from the same title.

Despite recognizing books more accurately than each of the individual baseline features, syntactic elements of expression on their own are less effective

Table 2. Classification results on the test set for recognizing books from their expression of content even when some books contain similar content

Feature Set	Accuracy on complete corpus	Accuracy on paraphrases only
Syntactic elements of expression	76%	89%
Tfidf-weighted keywords	66%	88%
Function words	61%	81%
Baseline linguistic	42%	53%
Dist. of word length	29%	72%
Dist. of sentence length	13%	14%

than the combined baseline features in recognizing books; the combined baseline features give an accuracy of 88% on recognizing books (compare this to 76% accuracy by the syntactic elements of expression alone). But the performance of the combined baseline features is further improved by the addition of syntactic elements of expression (see Table 3). This improvement is statistically significant at $\alpha = 0.05$.

Table 3. Classification results of combined feature sets on the test set for book recognition even when some books contain similar content

Feature Set	Accuracy on complete corpus	Accuracy on paraphrases only
All baseline features + syntactic elements of expression	92%	98%
All baseline features	88%	97%

Ranking the combined features based on information gain for recognizing books shows that the syntactic elements of expression indeed play a significant role in recognizing books accurately; of the top ten most useful features identified by information gain, seven are syntactic elements of expression (see rows in italics in Table 4).

In the absence of syntactic elements of expression, the top ten most useful features identified by information gain from the complete set of baseline features reveal that the keywords "captain" and "sister" are identified as highly discriminative features. Similarly, the function words "she", "her", and "'ll" are highly discriminative (see Table 5). Part of the predictive power of these features is due to the distinct contents of most of the books in this corpus; we expect that as the corpus grows, these words will lose predictive power.

Recognizing Authors: In Section 2, we described the difference between style and expression. These concepts, though different, both relate to the way people write. Then, an interesting question to answer is: Can the same set of features help recognize both books (from their unique expression) and authors (from their unique style)?

Table 4. Top ten features identified by information gain for recognizing books even when some books share content. Features which are syntactic elements of expression are in italics; baseline features are in roman.

Features
Std. dev. of the depths of the top-level left branches (measured in phrase depth)
Std. dev. of the depths of the top-level right branches (measured in phrase depth)
Std. dev. of the depths of the deepest prepositional phrases of sentences (measured in phrase depth)
% of words that are one character long
Average word length
% of sentences that contain unembedded verbs
% of sentences that contain an unembedded verb with noun phrase object (0-V-NP)
Frequency of the word "the" (normalized by chapter length)
Avg. depth of the subordinating clauses at the beginning of sentences (measured in phrase depth)
% of sentences that contain equal numbers of clauses in left and right branch
Type-token ratio

Table 5. Top ten baseline features identified by information gain that recognize books even when some books share content

Features
% words that are one character long
Average word length
Frequency of the word "the" (normalized by chapter length)
Type-token ratio
Frequency of the word "captain" (tfidf-weighted)
Probability of Negations
Frequency of the word "sister" (tfidf-weighted)
Frequency of the word "she" (normalized by chapter length)
Frequency of the word "her" (normalized by chapter length)
Frequency of the word "'ll" (normalized by chapter length)

Data: In order to answer this question, we experimented with a corpus of books that were written by native speakers of English. This corpus included works from eight authors: three titles by W. Irving, four titles by G. Eliot, five titles by J. Austen, six titles by each of C. Dickens and T. Hardy, eight titles by M. Twain, and nine titles by each of J. London and W. M. Thackeray.

Results: To evaluate the different sets of features on recognizing authors from their style, we trained models on a subset of the titles by each of these authors and tested on a different subset of titles by the same authors. We repeated this experiment five times so that several different sets of titles were trained and tested on. At each iteration, we used 150 chapters from each of the authors for training and 40 chapters from each of the authors for testing.

Table 6. Results for authorship attribution. Classifier is trained on 150 chapters from each author and tested on 40 chapters from each author. The chapters in the training and test sets come from different titles.

Feature Set	Accuracy Run 1	Accuracy Run 2	Accuracy Run 3	Accuracy Run 4	Accuracy Run 5
Function words	86%	89%	87%	90%	81%
Syntactic elements of expression	64%	63%	64%	55%	62%
Distribution of word length	33%	37%	44%	53%	35%
Baseline linguistic	39%	39%	41%	48%	28%
Distribution of sentence length	33%	41%	31%	41%	25%

Table 7. Average classification results on authorship attribution

Feature Set	Avg. Accuracy
Function words	87%
Syntactic elements of expression	62%
Distribution of word length	40%
Baseline linguistic	39%
Distribution of sentence length	34%

The results in Table 7 show that function words capture the style of authors better than any of the other features; syntactic elements of expression are not as effective as function words in capturing the style of authors. This finding is consistent with our intuition: we selected the syntactic elements of expression for their ability to differentiate between individual works, even when some titles are written by the same author and even when some books were derived from the same title. Recognizing the style of an author requires focus on the elements that are similar in the works written by the same author, instead of focus on elements that differentiate these works. However, the syntactic elements of expression are not completely devoid of any style information: they recognize authors accurately 62% of the time. In comparison, the function words recognize authors accurately 87% of the time. Top ten most predictive function words identified by information gain for authorship attribution are: `the`, `not`, `of`, `she`, `very`, `be`, `her`, `'s`, `and`, and `it`.

Combining the baseline features together does not improve the performance of function words on authorship attribution: function words give an accuracy of 87% by themselves whereas the combined baseline features give an accuracy of 86%.[1] Adding the syntactic elements of expression to the combination of baseline features hurts performance (see Table 8).

We believe that the size of the corpus is an important factor in this conclusion. More specifically, we expect that as more authors are added to the corpus, the contribution of syntactic elements of expression to authorship attribution will increase. To test this hypothesis, we repeated our experiments with up to thirteen authors. We observed that the syntactic elements of expression improved the

[1] This difference is not statistically significant.

Table 8. Average classification results of combined feature sets on authorship attribution

Feature Set	Average Accuracy for 8 Authors
All baseline features + syntactic elements of expression	81%
All baseline features	86%
Function words	87%
Syntactic elements of expression	62%

Table 9. Average classification results of combined feature sets on authorship attribution. For these experiments, the original corpus was supplemented with works from W. Ainsworth, L. M. Alcott, T. Arthur, M. Braddon, and H. James.

Feature Set	Average Accuracy for 8-13 Authors					
	8	9	10	11	12	13
All baseline features + syntactic elements of expression	81%	88%	88.4%	87.6%	88%	88%
All baseline features	86%	86%	87.8%	86.6%	86%	86.8%
Function words	87%	86.4%	85.4%	85.2%	84.8%	82.6%
Syntactic elements of expression	62%	65.6%	68.2%	67.4%	66%	64.4%

performance of the baseline features: as we added more authors to the corpus, the performance of function words degraded, the performance of syntactic elements of expression improved, and the performance of the combined feature set remained fairly consistent at around 88% (see Table 9).

4.3 Conclusion

In this paper, we compared several different language models on two classification tasks: book recognition and authorship attribution. In particular, we evaluated syntactic elements of expression consisting of sentence-initial and -final phrase structures, semantic and syntactic categories of verbs, and linguistic complexity measures, on recognizing books (even when they are derived from the same title) and on recognizing authors. Through experiments on a corpus of novels, we have shown that syntactic elements of expression outperform all individual baseline features in recognizing books and when combined with the baseline features, they improve recognition of books.

In our authorship attribution experiments, we have shown that the syntactic elements of expression are not as useful as function words in recognizing the style

of authors. This finding highlights the need for a task-dependent approach to engineering feature sets for text classification. In our experiments, feature sets that have been engineered for studying expression and the language models based on these feature sets outperform all others in identifying expression. Similarly, feature sets that have been engineered for studying style and the language models based on these feature sets outperform syntactic elements of expression in authorship attribution.

References

1. D. Alexander and W. J. Kunz. *Some Classes of Verbs in English.* Linguistics Research Project. Indiana University, 1964.
2. J. C. Baker. A Test of Authorship Based on the Rate at which New Words Enter an Author's Text. *Journal of the Association for Literary and Linguistic Computing*, 3(1), 36–39, 1988.
3. D. Biber. A Typology of English Texts. *Language*, 27, 3–43, 1989.
4. D. Biber, S. Conrad, and R. Reppen. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press, 1998.
5. E. Brill. A Simple Rule-Based Part of Speech Tagger. *Proceedings of the 3rd Conference on Applied Natural Language Processing*, 1992.
6. M. Diab, J. Schuster, and P. Bock. A Preliminary Statistical Investigation into the Impact of an N-Gram Analysis Approach based on Word Syntactic Categories toward Text Author Classification. In Proceedings of *Sixth International Conference on Artificial Intelligence Applications*, 1998.
7. C. S. Brinegar. Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship. *Journal of the American Statistical Association*, 58, 85–96, 1963.
8. G. Forman. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, 3, 1289–1305, 2003.
9. A. Glover and G. Hirst. Detecting stylistic inconsistencies in collaborative writing. In *Sharples, Mike and van der Geest, Thea (eds.), The new writing environment: Writers at work in a world of technology.* London: Springer-Verlag, 1996.
10. M. Halliday and R. Hasan. *Cohesion in English.* London: Longman, 1976.
11. M. Halliday. *An introduction to functional grammar.* London; Baltimore, Md., USA : Edward Arnold, 1985.
12. V. Hatzivassiloglou, J. Klavans, and E. Eskin. Detecting Similarity by Applying Learning over Indicators. *37th Annual Meeting of the ACL*, 1999.
13. V. Hatzivassiloglou, J. Klavans, M. Holcombe, R. Barzilay, M.Y. Kan, and K.R. McKeown. SimFinder: A Flexible Clustering Tool for Summarization. *NAACL'01 Automatic Summarization Workshop*, 2001.
14. D. I. Holmes. Authorship Attribution. *Computers and the Humanities*, 28, 87–106. Kluwer Academic Publishers, Netherlands, 1994.
15. B. Katz. Using English for Indexing and Retrieving. *Artificial Intelligence at MIT: Expanding Frontiers.* P. H. Winston and S. A. Shellard, eds. MIT Press. Cambridge, MA., 1990.
16. B. Katz and B. Levin. Exploiting Lexical Regularities in Designing Natural Language Systems. In *Proceedings of the 12th International Conference on Computational Linguistics*, COLING '88, 1988.
17. D. Khmelev and F. Tweedie. Using Markov Chains for Identification of Writers. *Literary and Linguistic Computing*, 16(4), 299–307, 2001.

18. G. Kjetsaa. *The Authorship of the Quiet Don.* ISBN 0391029487. International Specialized Book Service Inc., 1984.
19. M. Koppel, N. Akiva, and I. Dagan. A Corpus-Independent Feature Set for Style-Based Text Categorization. Proceedings of *IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.
20. O. V. Kukushkina, A. A. Polikarpov, and D. V. Khemelev. Using Literal and Grammatical Statistics for Authorship Attribution. Published in *Problemy Peredachi Informatsii*,37(2), April-June 2000, 96–108. Translated in "Problems of Information Transmission", 172–184.
21. B. Levin. *English Verb Classes and Alternations. A Preliminary Investigation.* ISBN 0-226-47533-6. University of Chicago Press. Chicago, 1993.
22. T. C. Mendenhall. Characteristic Curves of Composition. *Science*, 11, 237–249, 1887.
23. G. A. Miller, E. B. Newman, and E. A. Friedman.: Length-Frequency Statistics for Written English. *Information and Control*,1(4), 370–389, 1958.
24. A. Q. Morton. The Authorship of Greek Prose. *Journal of the Royal Statistical Society (A)*, 128, 169–233, 1965.
25. F. Mosteller and D. L. Wallace. Inference in an authorship Problem. *Journal of the American Statistical Association,* 58(302), 275–309, 1963.
26. R. D. Peng and H. Hengartner. Quantitative Analysis of Literary Styles. *The American Statistician*, 56(3), 175–185, 2002.
27. G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513–523, 1998.
28. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620, 1975.
29. R. E. Schapire. The Boosting Approach to Machine Learning. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.
30. H. S. Sichel. On a Distribution Representing Sentence-Length in Written Prose. *Journal of the Royal Statistical Society (A)*, 137, 25–34, 1974.
31. M. W. A. Smith. Recent Experience and New Developments of Methods for the Determination of Authorship. *Association for Literary and Linguistic Computing Bulletin*, 11, 73–82, 1983.
32. D. R. Tallentire. *An Appraisal of Methods and Models in Computational Stylistics, with Particular Reference to Author Attribution.* PhD Thesis. University of Cambridge, 1972.
33. R. Thisted and B. Efron. Did Shakespeare Write a Newly-discovered Poem? *Biometrika*, 74, 445–455, 1987.
34. Ö. Uzuner. *Identifying Expression Fingerprints using Linguistic Information.* Ph.D. Dissertation. Massachusetts Institute of Technology, 2005.
35. Ö. Uzuner and B. Katz. Capturing Expression Using Linguistic Information. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, 2005.
36. Ö. Uzuner, B. Katz and Thade Nahnsen. Using Syntactic Information to Identify Plagiarism. In *Proceedings of the Association for Computational Linguistics Workshop on Educational Applications (ACL 2005)*, 2005.
37. Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning.* 412–420, 1997.

38. G. U. Yule. On Sentence-Length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship. *Biometrika*, 30, 363–390, 1938.
39. J. Wilkinson and C. diMarco. Automated Multi-purpose Text Processing. In *Proceedings of IEEE Fifth Annual Dual-Use Technologies and Applications Conference*, 1995.
40. C. B. Williams. Mendenhall's Studies of Word-Length Distribution in the Works of Shakespeare and Bacon. *Biometrika*, 62(1), 207–212, 1975.
41. I. H. Witten and E. Frank. Data Mining: Practical machine Learning Tools with Java Implementations. Morgan Kaufmann, San Francisco, 2000.