# Capturing Expression Using Linguistic Information

**Özlem Uzuner and Boris Katz**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139

### Abstract

Recognizing similarities between literary works for copyright infringement detection requires evaluating similarity in the expression of content. Copyright law protects expression of content; similarities in content alone are not enough to indicate infringement. Expression refers to the way people convey particular information; it captures both the information and the manner of its presentation. In this paper, we present a novel set of linguistically informed features that provide a computational definition of expression and that enable accurate recognition of individual titles and their paraphrases more than 80% of the time. In comparison, baseline features, e.g., tfidf-weighted keywords, function words, etc., give an accuracy of at most 53%. Our computational definition of expression uses linguistic features that are extracted from POS-tagged text using context-free grammars, without incurring the computational cost of full parsers. The results indicate that informative linguistic features do not have to be computationally prohibitively expensive to extract.

## Introduction

Copyrights protect an author's expression of content;[1] in order to constitute potential infringement, two works need to present similar content and use a similar manner of expression.

For literary works, *content* refers to the story or the information and *expression* refers to the linguistic choices of authors in presenting this content, such as authors' choices of particular vocabulary items from a set of synonyms (e.g., "clever" vs. "smart" in sentences in (1)), whether they tend toward passive or active voice (e.g., sentences in (2)), or whether they prefer complex sentences with embedded clauses to simple sentences with independent clauses (e.g., sentences in (3)), as well as combinations of such choices.

1 (a) Jill is very clever.
(b) Jill is very smart.

2 (a) The pirates sank the boat.
(b) The boat was sunk by the pirates.

3 (a) The woman carrying the umbrella walked in the rain.
(b) The woman walked in the rain. She was carrying an umbrella.

Expression focuses on the linguistic choices of the authors and does not include layout or generic genre characteristics of documents because neither layout (such as use of titles, tables, and figures) nor genre characteristics (e.g., all poems consist of stanzas) represent linguistic choices of the authors. In this paper, we set out to create a computational definition of expression which can help evaluate similarities between literary works for copyright infringement detection. In particular, we study syntax and semantics to identify a novel set of linguistic elements that capture expression, and that provide a computational definition of expression, in the genre of narrative fiction.

Given a computational definition of expression, our goal is to generate fingerprints that help differentiate between two independently copyrighted works on the same content but also help recognize infringing copies of a work even when the infringement is not verbatim (i.e., paraphrases).

The ideal data set for this study would use examples of real-life infringement. Unfortunately, such a data set is not readily available. However, we have access to a corpus of parallel translations of *titles*; in this context, a title is an original work. Parallel translations, while not necessarily infringing, are derived from the same original title. During the translation process, translators add their own expression to the work and convey the same content in different ways, providing us with different *books* derived from the same *title*; we make this distinction between books and titles throughout this paper and rely on this distinction in our experiments.

Books derived from the same title can be treated as paraphrases of each other (and of the original title) and, in the absence of real-life infringement data, serve as our surrogate. Using this surrogate data, in this paper, we build models with a novel set of linguistic features and compare the performance of these features with baselines. The success of linguistic features in recognizing titles indicates that despite the differences in the way people phrase the same content, the essence of a literary work requires certain syntactic constructs to be present (either because of the content, or because people who derive content from the same original preserve some aspects of the original). We believe that our surrogate data and our findings will generalize to real-life

[1] United States Code, Title 17, Chapter 1, §102.

infringement cases: during infringement, despite efforts to paraphrase works, people use some similar constructs either to adequately convey content or because they are unwilling to rewrite the whole work—most infringers will make simple modifications to a work but are unwilling to put significant effort into re-creating it; if effort were not an issue, they would most likely create their own original rather than copying someone else's work.

## Related Work

Expression is defined as "the linguistic choices of authors in presenting content". Automatically evaluating expression similarity requires studying text similarity in terms of meaning and in terms of linguistic similarity (Uzuner, Davis, & Katz 2004; Uzuner & Davis 2003).

To classify documents based on their meaning, i.e., the story they present, most approaches use keywords. However, other linguistic information has also been used to represent the content of documents, e.g., subject–verb and verb–object relationships, noun phrases, WordNet synsets, and semantic classes of verbs (Hatzivassiloglou, Klavans, & Eskin 1999) from Levin's studies (Levin 1993). Linguistic similarity between works has been studied in stylometry for identifying the style of an author in terms of a variety of features, including distribution of word lengths (Williams 1975) and sentence lengths (Sichel 1974), distribution of function words (Mosteller & Wallace 1963; Peng & Hengartner 2002), and measures of richness of vocabulary (Holmes 1994; Thisted & Efron 1987). Overall, both linguistically uninformed features, e.g., sequences of letters (Kukushkina, Polikarpov, & Khmelev 2000), and linguistically more informed features, e.g., syntactic classes (parts of speech) of words (Glover & Hirst 1996; Koppel, Akiva, & Dagan 2003), have been successfully used for capturing an author's style.

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. Similarities in the expression of similar content in two different works signal potential copying and require further scrutiny under copyright.

Expression and style are both based on linguistic elements of authors' writings. Which linguistic features are more useful for identifying expression and which are more useful for style depends on the group of authors and works that are studied. But in general, different groups of features would be used to define an author's overall style and to define his unique expression in a work. For example, if an author always uses long sentences, his style can partly be described in terms of the length of his sentences; however, this information is not enough for capturing expression as it does not indicate which work is copied. On the other hand, the author may use predominantly left-embedded sentences in one work and predominantly right-embedded sentences in another. This information can be used to capture the different expressions of his works, but would not help define his style.

A fingerprint that can identify a work for copyright purposes has to capture the expression of content that is unique to that work, and that differentiates it from the expressions of other authors who write about similar content as well as the expression of other content by the same author. We hypothesize that syntax and semantics are useful for this purpose.

## Linguistic Elements of Expression

Authors of creative works rely on elements of language to create a particular expression. Translated literary works provide examples of linguistic choices that differ in expression but convey similar content. For example, consider the following semantically equivalent excerpts from three different translations of *Madame Bovary* by Gustave Flaubert.

> Excerpt 1: "The evening the Bovarys were expected at Yonville, Madame Lefrancois, the widow who owned this hotel, was so frantically busy with her saucepans that large beads of sweat stood out on her face. Tomorrow was market day, and she had to get every-thing ready in advance. Cut the meat, clean the chickens, make soup, roast and grind the coffee." (Translated by Unknown1.)
>
> Excerpt 2: "On the evening when the Bovarys were to arrive at Yonville, widow Lefrancois, the landlady of this inn, was so very busy that she sweated great drops as she moved her saucepans. To-morrow was market-day. The meat had to be cut beforehand, the fowls drawn, the soup and coffee made." (Translated by Aveling.)
>
> Excerpt 3: "The night the Bovarys were due to arrive at Yonville, widow Lefrancois, who kept the inn, was in such a fluster that the sweat fell from her in huge drops as she bustled about among her pots and pans. Tomorrow was market day; she had the joints to prepare, the fowls to draw, the soup to make and the coffee to brew." (Translated by Unknown2.)

Inspired by the syntactic differences observed in parallel translations, in this section we first present a novel set of syntactic features that relate to how people convey content (syntactic elements of expression); we then provide features that capture content itself (semantic elements of expression). All of the features presented in this section are extracted from part-of-speech tagged text (Brill 1992), using context-free grammars. This particular approach was taken in order to test the hypothesis that extraction of linguistic information for text classification purposes does not have to be computationally prohibitively expensive; that we can extract linguistically-informed features without full parsing; and that despite the tradeoff between accuracy and efficiency, the features extracted are informative.

### Syntactic Elements of Expression

We hypothesize that given particular content, authors choose from a set of semantically equivalent syntactic constructs to create their expression. Our observations of the expressive choices of authors in parallel translations led us to define syntactic elements of expression in terms of sentence-initial and -final phrase structures, semantic classes and argument structures of verb phrases, syntactic classes of verb phrases, and linguistic complexity of sentences.

**Sentence-initial and -final phrase structures** The order and the nature of phrases in a sentence can be an expressive tool: authors often shift the emphasis of a sentence by reordering particular facts and by employing relative clauses. The resulting expressive differences affect the distributions of sentence-initial and -final noun phrases, prepositional phrases, verb phrases, and adverb phrases, as well as sentence-final stranded prepositions, modals, and auxiliary verbs (indicating movement of constituents).

**Semantic Classes of Verbs** Levin (1993) observed that the syntax and semantics of verbs are related, and showed that verbs that exhibit similar syntactic behavior are also related semantically. Based on this observation, Levin sorted around 3000 verbs into 49 semantic classes. We use these classes to describe the expression of an author in a particular work in terms of the semantic classes of verbs she uses and the particular argument structures she prefers for them. For example, for the semantic class of "coil verbs", the base form and the causative alternation have the following formulae:

1. **Base Form**
   - Cora coiled the rope around the post.
   - NP + V + NP + PP.
2. **Causative Alternation**
   - The rope coiled around the post.
   - NP + V + PP.

Semantic classes of verbs provide useful information for many natural language processing applications: START was the first natural language system to successfully use such verb classes for question answering (Katz & Levin 1988).

Identification of semantic classes of verbs is not trivial because many verbs belong to multiple semantic classes. As word sense disambiguation is outside the scope of this paper, to capture the semantics of the verbs used in a document, we obtained the distribution of semantic classes of verbs by crediting all semantic classes of all verbs in a document. We combined this information with information about the argument structures in which verbs are observed (Levin 1993). We expressed the argument structures in terms of phrase structures, e.g., NP + V + PP, and extracted them from part-of-speech tagged text using context-free grammars.

**Syntactic Classes of Verbs** Levin's verb classes are limited to those that do not take clausal or verb phrase embeddings (a.k.a. "non-embedding verbs") and need to be supplemented by classes of "embedding verbs" that take complex arguments such as clauses and verb phrases. We study the syntax of embedding and non-embedding verbs in two different ways. For non-embedding verbs, we find the semantic class (from Levin) and the argument structures in which they are observed as described in the previous section. For embedding verbs, we identify their syntactic class and the structure of their observed embedded arguments.

Alexander and Kunz (1964) identified syntactic classes of embedding verbs and collected examples of verbs for each class. In their studies, they described verb classes with formulae written in terms of phrasal and clausal elements, such as verb phrases (VP), participial phrases (Particip.), infinitive phrases (Inf.), indicative clauses (IS), subjunctives (Subjunc.), and small clauses (SC). A similar set of embedding verbs was used for parsing and generation in START (Katz 1990). For our studies, we used 29 of the more frequent verb embedding classes from Alexander and Kunz, and identified the distributions of these embedding classes in different works. Examples of these verb classes are shown in Table 1.

| Syntactic Formula | Example |
|---|---|
| NP + Vh + IS | I assume she left. |
| NP + Vh + NP + IS | I will show him I'm right. |
| NP + Vh + SC | They saw John leave. |
| NP + Vh + NP + Partic. | I caught him stealing. |
| NP + Vh + NP + that + IS | She told him that she left. |
| NP + Vh + NP + to + Inf. | They asked him to help. |
| NP + Vh + NP + wh + IS | He asked me if she came. |
| NP + Vh + Particip. | I began singing. |
| NP + Vh + Subjunc. | I request she go alone. |

Table 1: Sample syntactic formulae and examples of embedding verb classes based on Alexander and Kunz (1964).

Syntactic and semantic characteristics of verb phrases can be combined to reveal further elements of expression, e.g., syntactic classes of embedding verbs and the classes of semantic non-embedding verbs they co-occur with.

**Linguistic Complexity** Sentence length distributions have been used in the literature to describe an author's style. We hypothesize that sentence length is a rough approximation of a deeper linguistic phenomenon, namely the level of linguistic complexity. To test this hypothesis, we studied linguistic complexity in terms of the mean and standard deviation of the depths of the top-level left and right branches in sentences in terms of phrase depth; the number of prepositional phrases in sentences; the mean and standard deviation of the number of prepositional phrases in sentences; the mean and standard deviation of the depths of the deepest prepositional phrases in sentences; the percentage of left-heavy, right-heavy, and equal-weight sentences (e.g., sentences where the top-level right branch of the syntax tree is deeper than the top-level left branch are considered right-heavy); the mean and standard deviation of the number of embedded clauses in the top-level left and right branches in sentences; the percentage of left-embedded, right-embedded, and equally-embedded sentences (e.g., sentences where the top-level right branch of the syntax tree embeds more clauses than the top-level left branch are considered right-embedded); the mean and standard deviation of the depths of sentence-initial subordinating clauses in sentences.

The extraction of these features (explained in detail by Uzuner (2005)) relies on a partial parse and identification of main constituents of sentences. Full parsing and identifying the dependencies between the constituents in a sentence is not essential for extracting these features, e.g, prepositional phrases can be identified and their depth can be reasonably estimated even when we do not have any information about the correct prepositional attachment.

**Validation** For each of these features, we used the chi-square (and/or likelihood ratio) test of independence to check whether these features are distributed differently in different works (alternate hypothesis). For each feature set, we tested the null hypothesis (that these features are used similarly by all authors and that the observed differences are due to chance) in three different settings: on translations of the same title (similar content but different expression), on different books by different authors (different content and different expression), and finally on disjoint sets of chapters from the same book (similar content and expression).

For almost all of the identified features, we rejected 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.

### Semantic Elements of Expression

To create a compact and effective representation of content, we used the General Inquirer (GI) dictionary[2] from which we gathered 62 high-level semantic categories (GI categories). The GI dictionary contains 11,000 words marked by semantic categories such as `Strong`, `Weak`, `Hostile`, `Place`, etc. As with many dictionaries, each word in the GI dictionary can have multiple senses and can belong to multiple GI categories. For example, in this dictionary, the word "make" has nine senses (see Table 2 for examples) and each of its senses belongs to one or more GI categories.

| Sense | GI Categories |
|---|---|
| Verb: create, construct | Strong, Active, Work, IAV |
| Verb: coerce, force to | Negativ, Strong, PowTOT, Active, SocRel, IAV |

Table 2: Two senses of "make" and their GI categories.

Given ambiguity in word senses, to map each word to its correct GI category in a given context, we created a representation (in terms of GI categories) for the global context of documents and used this global context to resolve ambiguities in favor of the categories that were dominant in the global context. Our representation of global context consisted of all the GI categories associated with all of the senses of all of the words in the document; each category was weighted appropriately for its contribution to the global distribution of GI categories (Uzuner 2005). Given this global context, for each sense of each polysemous word in the document, we calculated a score based on how well its GI categories aligned with the global GI categories of the document and resolved ambiguity in favor of highest-scoring sense, i.e., the sense whose GI categories had the highest overlap with the GI categories of the document.

**Validation** To validate the GI categories, we compared their performance with tfidf-weighted keywords on a 45-way classification task (see the section on Data for details on our corpus). This experiment showed that the GI categories correctly recognize paraphrased titles (3 *titles* out of 45 had paraphrases; these 3 titles were paraphrased in a total of 7 *books*.) 83% of the time whereas tfidf-weighted keywords recognize the paraphrased titles only 71% of the time. The random chance baseline on this task is 2.2%. The GI categories achieve their performance on a 62 dimensional space, reducing the feature space significantly from 11,000 keywords that appear in our corpus.

## Evaluation

In this paper, we present 5 general categories of linguistic features (4 syntactic and 1 semantic) that capture expression; however, these features and their combinations expand into around 1400 features which represent the linguistic elements of expression. For evaluation, we compared these features with the appropriate baselines.

### Data

As mentioned in the Introduction, the ideal data set for this study would use examples of real-life infringement. Unfortunately, such a data set is not readily available. In its absence, we used a corpus of parallel translations of *titles* as a surrogate for infringement data. This corpus included 45 titles and 49 books derived from these titles; for 3 of the titles, the corpus included multiple books (3 books paraphrased the title *Madame Bovary*, 2 books paraphrased *20000 Leagues*, and 2 books paraphrased *The Kreutzer Sonata*). The remaining titles included literary works from J. Austen (1775-1817), F. Dostoyevski (1821-1881), C. Dickens (1812-1870), A. Doyle (1859-1887), G. Eliot (1819-1880), G. Flaubert (1821-1880), T. Hardy (1840-1928), I. Turgenev (1818-1883), V. Hugo (1802-1885), W. Irving (1789-1859), J. London (1876-1916), W. M. Thackeray (1811-1863), L. Tolstoy (1828-1910), M. Twain (1835-1910), and J. Verne (1828-1905).

### Baseline Features

To evaluate the linguistic elements of expression, we used as baselines, features that capture content as well as features that capture the way works are written.

**Tfidf-weighted Keywords** Most content-based text classification tasks use unordered sets of stemmed keywords, i.e., content words, to classify documents. We use tfidf-weighted keywords as a baseline, but exclude from this set proper nouns. As mentioned in the Introduction, most infringers will make simple modifications to a work but are unwilling to put significant effort into re-creating a work. Modifying proper nouns is an example of a simple modification that makes it difficult to identify copies of a work. Therefore, while traditional paraphrase recognition approaches often rely on proper nouns to identify text components that paraphrase each other, we believe that it is more realistic to evaluate approaches to copyright infringement detection without relying on proper nouns.

**Function Words** In studies of authorship attribution, many researchers have taken advantage of the differences

[2] http://www.wjh.harvard.edu/ inquirer/homecat.htm

in the way authors use function words (Mosteller & Wallace 1963; Peng & Hengartner 2002). In our studies, we used the set of 363 function words from which Mosteller and Wallace's 70 function words were selected. We augmented this list with 143 function words, for a total of 506, that are frequently used in modern English. The list of function words can be found in (Uzuner 2005).

**Word Length Distribution** Distributions of word lengths have been used for authorship attribution. Mendenhall (1887) showed that the distribution of word lengths in the works of Shakespeare was different from the distribution representing the works of Bacon, but that word length distributions did not help separate the works of Shakespeare from those of Marlowe. Although Williams (1975) later provided a genre-related explanation for this result, for completeness, we included them in our study.

**Distribution of Sentence Lengths** Sentence length distributions, means, and standard deviations (Holmes 1994) have also been frequently used for authorship attribution; they are included in our experiments as a baseline.

**Baseline Linguistic Features** Sets of surface, syntactic, and semantic features have been used in the literature for authorship attribution and have been adopted here as baseline features. These features include surface features: number of words and the number of sentences in the document; type–token ratio, i.e., the ratio of the total number of unique words in the document to the number of words in the document; average and standard deviation of the lengths of words (in characters) and of the lengths of sentences (in words) in the document. Baseline syntactic features included: frequencies of declarative, interrogative, imperative, and fragmental sentences; frequencies of active voice sentences, be-passives, e.g., "I was robbed", and get-passives, e.g., "I got robbed"; frequencies of 's-genitives, e.g., "Jane's book", of-genitives, e.g., "pots of the flowers", and of phrases that lack genitives. Baseline semantic features were frequency of overt negations, e.g., "not", "no", etc., and frequency of uncertainty markers, e.g., "could", "maybe", etc.

## Classification Experiments

To evaluate our features, we compared the linguistic elements of expression with the baseline features on two separate experiments: recognizing titles even when some titles are paraphrased and recognizing books even when some of them paraphrase the same title. For these experiments, we split books into chapters and classified chapters (using boosted decision trees (Witten & Frank 2000)) into titles and books. We extracted all relevant features from each chapter and normalized them by the length of the chapter.

For both experiments, we created a balanced data set of relevant classes, using 60% of the chapters from each class for training and the remaining 40% for testing. Parameter tuning on the training set showed that the performance of classifiers (for all feature sets) stabilized at around 200 rounds of boosting. In addition, limiting our features to those that had non-zero information gain on the training set eliminated noisy features.

**Recognizing Titles** Copyrights are granted for a limited time period. During the copyright period of a title, its paraphrases are considered derivatives of the original; reproduction of a work and generation of its derivatives are exclusive rights of the copyright holder. For copyright infringement detection, paraphrases have to be recognized as such.

In other words, given a title, we need to recognize its paraphrases. For this, we randomly selected 40–50 chapters from each title in our corpus. For paraphrased titles, we selected training chapters from one of the paraphrases and testing chapters from the remaining paraphrases. We repeated this experiment three times; at each round, a different paraphrase was chosen for training and the rest were used for testing.

Our results show that, on average, linguistic elements of expression accurately recognize titles 81% of the time and significantly outperform all baselines (see Table 3).[3]

| **Feature Set** | **Avg. accuracy (complete corpus)** | **Avg. accuracy (paraphrases only)** |
|---|---|---|
| Linguistic elements | 81% | 95% |
| Linguistic elements w/o GI | 73% | 95% |
| Function words | 53% | 34% |
| Tfidf-weighted keywords | 47% | 38% |
| Baseline linguistic | 40% | 67% |
| Dist. of word length | 18% | 54% |
| Dist. of sentence length | 12% | 17% |

Table 3: Classification results (on the test set) for recognizing titles even when some titles are paraphrased (middle column) and for recognizing only the paraphrased titles included in the corpus (right column). Random chance would recognize a paraphrased title 2% of the time.

Further analysis of the results indicate (see right column in Table 3) that linguistic elements of expression accurately recognize on average 95% of the paraphrased titles and that removing the GI categories from this feature set does not change this result. This finding implies that some of our linguistic elements of expression are common to paraphrases of a title. This commonality could be due to the similarity of their content (which implies a dependency between semantics and syntax), or due to the underlying expression of the original author (which the translators consciously or subconsciously reflect in their paraphrases).

**Recognizing Books** Paraphrases of a title, if created after the copyright period, are eligible for their own copyright. In such cases, it is necessary to recognize each book (even when some books paraphrase the same title) from its unique expression of content.

To test different feature sets for recognizing books, we ran classification experiments on a collection that contained 40–50 chapters from each book (including each of the books that paraphrased the same title) in our corpus. We found

[3] For the corpora used in this paper, a difference of 4% or more is statistically significant with $\alpha = 0.05$.

that (see Table 4) linguistic elements of expression accurately recognized books 82% of the time and that removing GI categories from this feature set reduced accuracy to 76%; in either case, our features significantly outperformed all baseline features. Further analysis of the results showed that linguistic elements of expression accurately recognized each of the paraphrases 92% of the time and that removing GI categories from the feature set did not change the results significantly (see right column on Table 4). That linguistic features can differentiate between paraphrases of the same title indicates that translators add their own expression to works; the expressive elements chosen by each translator help differentiate between paraphrases of the same title.

| **Feature Set** | **Accuracy (complete corpus)** | **Accuracy (paraphrases only)** |
|---|---|---|
| Linguistic elements | 82% | 92% |
| Linguistic elements w/o GI | 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% |

Table 4: Classification results on the test set for recognizing books.

## Conclusion

In this paper, we presented a novel set of linguistic features that capture expression of content and demonstrated that these linguistic elements of expression recognize books (even when they paraphrase the same title) and titles (even when they are paraphrased), more accurately than any of the baseline features traditionally used in the literature. By capturing differences in expression of the same content, these features enable recognition of independently copyrighted paraphrases of the same title. By capturing similarities in the expression of paraphrases of a title, these features enable recognition of potentially infringing copies.

Despite being linguistically informed, this novel set of features has been extracted from POS-tagged text using context-free grammars, without imposing onto the system the computational cost of a full syntactic parser. Our results indicate that useful linguistic information does not have to be computationally prohibitively expensive to extract.

## Acknowledgements

The authors would like to thank Sue Felshin and the anonymous reviewers for their insightful comments. This work is supported in part by the Advanced Research and Development Activity as part of the AQUAINT research program.

## References

Alexander, D., and Kunz, W. J. 1964. Some classes of verbs in English. In *Linguistics Research Project*. Indiana University.

Brill, E. 1992. A simple rule-based part of speech tagger. In *3rd Conference on Applied Natural Language Processing*.

Glover, A., and Hirst, G. 1996. Detecting stylistic inconsistencies in collaborative writing. In Sharples, M., and van der Geest, T., eds., *The new writing environment: Writers at work in a world of technology*. Springer-Verlag.

Hatzivassiloglou, V.; Klavans, J.; and Eskin, E. 1999. Detecting similarity by applying learning over indicators. In *37th Annual Meeting of the ACL*.

Holmes, D. I. 1994. Authorship attribution. *Computers and the Humanities* 28.

Katz, B., and Levin, B. 1988. Exploiting lexical regularities in designing natural language systems. In *12th Int'l Conference on Computational Linguistics, COLING '88*.

Katz, B. 1990. Using english for indexing and retrieving. In Winston, P. H., and Shellard, S. A., eds., *Artificial Intelligence at MIT: Expanding Frontiers*. MIT Press.

Koppel, M.; Akiva, N.; and Dagan, I. 2003. A corpus-independent feature set for style-based text categorization. In *IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*.

Kukushkina, O. V.; Polikarpov, A. A.; and Khmelev, D. V. 2000. Using literal and grammatical statistics for authorship attribution. *Problemy Peredachi Informatsii* 37(2).

Levin, B. 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. University of Chicago Press.

Mendenhall, T. C. 1887. Characteristic curves of composition. *Science* 11.

Mosteller, F., and Wallace, D. L. 1963. Inference in an authorship problem. *Journal of the American Statistical Association* 58(302).

Peng, R. D., and Hengartner, H. 2002. Quantitative analysis of literary styles. *The American Statistician* 56(3).

Sichel, H. S. 1974. On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society (A)* 137.

Thisted, R., and Efron, B. 1987. Did Shakespeare write a newly-discovered poem? *Biometrika* 74.

Uzuner, Ö., and Davis, R. 2003. Content and epression-based copy recognition for intellectual property protection. In *Proceedings of the 3rd ACM Workshop on Digital Rights Management (DRM'03).*

Uzuner, Ö.; Davis, R.; and Katz, B. 2004. Using empirical methods for evaluating expression and content similarity. In *37th Hawaiian International Conference on System Sciences (HICSS-37). IEEE Computer Society.*

Uzuner, Ö. 2005. *Identifying Expression Fingerprints Using Linguistic Information*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Williams, C. B. 1975. Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. *Biometrika* 62(1).

Witten, I. H., and Frank, E. 2000. *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco.