Identifying Authorship from Linguistic Text Patterns

Joshua Madden
Georgia State University
jmadden4@gsu.edu

Veda Storey
Georgia State University
vstorey@gsu.edu

Richard Baskerville
Georgia State University, USA,
and Curtin University, Western
Australia
baskerville@gsu.edu

Abstract

Research that deals with linguistic text patterns is challenging because of the unstructured nature of text. This research presents a methodology to compare texts to identify whether two texts are written by the same or different authors. The methodology includes an algorithm to analyze the proximity of text, which is based upon Zipf's Law [47][48]. The results have implications for text mining with applications to areas such as forensics, natural language processing, and information retrieval.

1. Introduction

Identifying the true authorship of a text based upon linguistic components has a long history in a variety of fields, including for authors as famous as William Shakespeare [3][42][44]. This tradition of identifying accurate authorship has applications beyond mere curiosity, with impacts within the national security and criminal justice system, where identifying authorship can be a key aspect of identifying suspects, with a famous example being the Unabomber (an American criminal who used the U.S. Postal Service to send explosives to victims), who was identified based upon the linguistic patterns in his manifesto [28].

Although prior work suggests that authors can be identified based upon the linguistic patterns they employ [40], it is unclear how much similarity is dependent upon repeated structural patterns. Relying on only structural patterns could lead to misidentification of authorship, when one considers inherent common cultural structures amongst authors from similar geographical areas or ideologies. For example, cooperation between terrorist groups can impact their longevity [36]. Since terrorist organizations originate from similar cultural, religious and ideological backgrounds, this presents a potentially large problem for forensic linguistics because considering only structural patterns may not be sufficient. At the same time, identifying terrorists and others based upon their online presence is needed [6]. Incorrectly, unnecessarily, or too quickly identifying a group responsible for an attack can cause problems [20][10].

The applications for looking beyond structural components within forensic linguistics are even clearer. Carr [6] emphasizes that the internet is serves as "an all-purpose communications network, surveillance medium, propaganda channel and recruiting tool." Researchers have retrieved audio messages, images of attack targets, covert terrorist websites and videos, highlighting the need for linguistic analysis from a forensic perspective [6]. The government has also funded research to identify authors of online text messages based upon the users' diction and syntax [6]. Besides law enforcement, there are many applications of big data analytics that could make use of an improved ability to identify a common author of multiple texts.

This research attempts to isolate linguistic components of texts with a similar structure to comparatively test one against another, even when the authors are different. It is intended to identify the presence of specific authors based on analysis of their writings, even when the linguistic components are held constant.

Big Data analysis techniques, including text analysis and text mining, have grown and enable faster and more precise understanding of large volumes of text than previously possible. Linguistic factors for analyzing text have also progressively become more important, although less adopted, in information systems research, than in other fields such as computational linguistics and human-computer interaction [34]. New methods for analyzing text have resulted in increasing quantification of large bodies of text (e.g., counts of numbers of terms) [21][22][39].

Traditional text analysis emphasizes the actual text, but often disregards the underlying linguistic factors or lack some key linguistic aspect needed for in-depth analysis [38][5]. In addition to conducting sentiment and other forms of text analysis that consider the

URI: https://hdl.handle.net/10125/60010
ISBN: 978-0-9981331-2-6

meanings of words, the actual patterns of language and word usage can provide useful data that is often ignored [5]. By including these linguistic components, we can expand the scope of text analysis.

The objective of this research is to develop a linguistic approach to generating useful information from text. To do so, we develop a methodology to quantify the similarities between texts as a digital innovation in the sense of Fichman et al. [18]. Specifically, this research takes a design science approach to creating a method for identifying commonalities in separate texts (even from different languages). The method is based upon an algorithm and implemented in a prototype for testing, thus serving as an instantiated artifact [23]. It also incorporates previous findings of extensibility, linguistic component theory and Zipf's law. The underlying logic of Zipf's law is used to create a new algorithm. The contribution of the research is to provide a methodology that enables text analysis and applications for a larger number and variety of writings in order to determine authorship than has previously been possible.

This paper proceeds as follows. Section 2 reviews related research on design science and linguistics within information systems. Section 3 presents the new algorithm, outlines the method and its implementation. Section 4 applies the method to identify common authorship amongst texts. Section 5 discusses the results and suggests areas for future research. Section 6 concludes the paper.

2. Theoretical background

Research dealing with text in unstructured forms is important in areas such as big data analytics, sentiment analysis, and social media analytics. Approaches to dealing with corpus of texts usually include natural language parsing techniques [8]. Although this research is primarily built upon the information systems literature, it also builds upon work in computer science, literary criticism and computational linguistics [34]. This is largely due to the influence of Zipf's law, an algorithm explaining frequency patterns within a group of phenomenon as diverse as word frequency, the distribution of city size and the distribution of income [48][26][25][45]. The principles behind Zipf's law, particularly those related to the exponentially increasing rarity of less commonly used words, serves as the foundation of part of the algorithm in our method.

2.1. Digital innovation

Fichman et al. [18] define digital innovation as an expansion of traditional information systems or technology innovation. Within information systems research, digital innovation has been given an increasing focus from 2009 – 2015 [17]. Yoo et al. [46] analyze the translation of physical products into digitalized forms. Crossan and Apaydin [13] define digital innovation as "both a process and an outcome" that occurs within organizational contexts. Information systems research has focused on digital innovation within organizational contexts [17].

2.2. Extensibility

Mastora et al. [34] argue that "Natural language is both fundamental and complicated as a communication system; therefore, it has been the subject of many disciplines" and that it has "rules, norms and patterns concerning its morphology and syntax" (pg. 496). They quote Portner [34][37] who argues that "the theory of [meaning] holism claims that the meaning of a word or phrase or sentence depends on its relationships with other words, phrases, and sentences" (pg. 496). In other words, the full meaning of a word cannot be determined without considering the context within which it is used.

Human language is dynamic and constantly changing. Subsequently, any method designed to analyze human language must feature *extensibility,* the ability to indefinitely expand without any barriers, in its design. Human language is anchored in culture, and cultures comprising a potentially infinite variety of combinations. Therefore, any artifact that is designed to analyze text in a meaningful way must accommodate a wide variety of *linguistic components.*

In this research, a theory known as linguistic component theory is presented, which proposes that authors will exhibit regularities in their language use, and that these regularities will be comparable both with language usage in general, and with the author's language usage, in particular. Therefore, our proposed method will operate within the context of linguistic regularities, of which *Zipf's Law* [47][48] is a well-known example due to the patterns it identifies across languages.

Natural language is indefinitely extensible [11][12][41][33], so it can be continually extended, and changed, existing in a state of impermanence. No true form of permanent modeling for language studies can ever really exist [33]. A similar concept, relative indefinite extensibility, can be explained through several examples (e.g., [33]), including, most notably, the fact that there is no complete, written set of all possible existing numbers (due to the infinite number of possible and valid combinations). Therefore, any information system artifact that attempts to model language must also be indefinitely extensible. No system can be pre-programmed to include an infinite

number of possible (and valid) numeric combinations, but there are still contexts within which these terms can be used. Because of this, systems should be applied within many contexts to adapt to the changing circumstances surrounding the language being studied.

2.3. Linguistic component theory

Linguistic component theory is a set of assumptions proposing that models can be improved by factoring in linguistic components (such as the analysis of text-based data). In a global economy, understanding "new signals" from other cultures is important, particularly where data can be taken from countries all over the world and integrated into one project [30]. A deeper integration of linguistics, which can only result from a deep understanding of the linguistic components inherent in the data, will facilitate the understanding of these signals. Senior executives now strive to run their companies on data-driven insights [30]. However, this approach cannot be effective if the insights from this data do not accurately reflect the linguistic context within which it exists. To understand the data that drives the insights, one must consider the larger linguistic context.

Previous deep structural work within information systems shows that information systems can be viewed and modeled as independent artifacts that reflect the real-world context it is intending to model [43]. These contexts include a linguistic component inherent in all informational transactions due to the universal usage of language by human beings. The inclusion of linguistic-based data [38][5] can help to represent this real-world context accurately. Although surface-level structure, such as the actual content of the text being analyzed, can change with social context, the underlying deep-structure is more consistent and can, potentially, provide more useful data, even across different genres of works or languages [43].

2.3. Zipf's Law

Zipf's Law [47][48] is a well-known linguistic algorithm which predicts that the frequency with which a word is used is inversely proportional to its ranking overall within the corpus. Zipf's law shows that the frequency in a word's usage decays at an exponential rate, based on its ranking against other words within the language as a whole [16]. This means that the word used second most in a language is used half as much as the first, the third most used is used one-third as often as the first, etc. Zipf's Law, as well as modified forms of the algorithm, have been used within the field of computational linguistics for some time [2].

Table 1. Design Science Research

Component	Task
Problem identification and motivation	Show how the lack of linguistic sensitive analysis within text analysis prevents some analyses from being sufficient
Objectives of a solution	Create a method for addressing linguistic components.
Design and development	Create a method to analyze linguistic factors within differing bodies of text by adapting and extending an algorithm.
Demonstration	Implement the method in a prototype.
Evaluation	Evaluate whether the prototype answers potential research questions and/or tests appropriate hypotheses.
Communication	Document the development of the method and the resulting calculations in proof-of-concept applications.

3. Methodology

This research uses the design science approach of Peffers et al. [35], as summarized in Table 1.

Problem identification and motivation. Because Big Data methods have allowed for an increased ability to quantify text ([21][22][39]), this research attempts to identify issues and challenges that could benefit from emphasizing the linguistic components of text. By identifying potential areas where this could be helpful, such as authorship identification a solution can be developed.

Objectives of a solution. We focus on authorship identification and error detection in automated translation software and present hypotheses and research questions that can be tested and/or answered by analyzing the linguistic structure of bodies of texts. For the authorship identification application, we present several hypotheses centered around a central notion. This is, given the choice between three pairs of works (e.g., book), where one pair represents a pair of works by the same author, and the other two pairs represent works by two different authors, a linguistic sensitive

analysis should be able to identify which pair of works were written by the same author more successfully than by random chance.

Design and development. We develop a method based on an algorithm that can analyzes the underlying linguistic structure of differing bodies of text. Based upon Zipf's Law [47][48], we develop a new algorithm focused on word frequencies within texts and show what this can reveal about authorship.

Demonstration. This algorithm is incorporated into a program that can take as input bodies of text (placed in .txt files, and ranging from short poems to entire novels) and can run the algorithm using the words provided within these .txt files. The program calculates relative measures of commonalities across the bodies of text, showing the similarities between different works.

Evaluation. The relative comparison values are used to test hypotheses and/or to answer research questions. Two applications are used in the evaluation: authorship identification and automated translation.

Communication. The results of this process are presented in this paper.

3.1. Zipf's-law based Method

The method developed to compare text is comprised of a set of steps that generate the data needed to make the comparisons. The steps of the method are as follows.

Step 1: Generate a set of corpus values for the entire data set.

Calculate the total number of words in the corpus. For the purposes of this paper, we use three works in each dataset. This can be a set of any three works (for example, three separate novels) that are tested together. Then, calculates the total number of words, which need not be unique:

$$\sum_{i} N(w_i \in Corpus) = \sum_{i,j} N(w_i \in T_j) \quad (1)$$

For each unique word, a value based upon the number of times a unique word occurs is calculated. Less frequently used words are valued more highly than more frequently used ones, a principle borrowed from the underlying logic of Zipf's law.

For $w_i \in \bigcup_{j \in J} T_j$, we have,

$$N(w_i \in Corpus) = \sum_{j} N(w_i \in T_j) \quad (2)$$

Step 2: Perform individual word analysis and values.

Once this value is created for every word in the corpus, it will be converted to a proportional value that shows the frequency of the usage within the context of the data set, and which can be adapted based on the structure of the text being analyzed. The analysis is primarily based on word counts and frequencies, rather than the structure of the actual work.

$$F(w_i \in Corpus) = N(w_i \in Corpus) / \sum_{i} N(w_i \in Corpus) \quad (3)$$

$F(w_i \in Corpus)$ represents the relative frequency for each unique word in the corpus (datasets containing a wide variety of texts).

$C(w_i \in Corpus)$ is the complementary value of the relative frequency for each unique word in corpus.

$$C(w_i \in Corpus) = 1 - F(w_i \in Corpus) \quad (4)$$

This value is generated for every word in the corpus. The number of times each word is used within two texts being compared (versus the corpus overall) is expressed as follows.

For each word $w_i \in T_1 \cap T_2$:

$$N(w_i \in T_1 \cap T_2) = N(w_i \in T_1) + N(w_i \in T_2) \quad (5)$$

For the word $w_i \notin T_1 \cap T_2$, $N(w_i) = 0$.
The commonalities between the texts are expressed using a unique word value.

$$I(w_i \in T_1 \cap T_2) = N(w_i \in T_1 \cap T_2) \times C(w_i \in Corpus) \quad (6)$$

This is generated for each word present in the two texts. It is the total word count from each of the two texts from each genre selected for comparison and is totaled to obtain what is referred to as the "comparison value."

Step 3: Generate comparable values.

Comparison of words is performed by:

$$Comp(T_1, T_2) = \sum_{i} I(w_i \in T_1 \cap T_2) \quad (7)$$

where $Comp(T_1, T_2)$ is comparable value of Text 1 and Text 2.

The total number of words in the comparison is:

$$\sum_i N(w_i \in T_1 \cap T_2) = \sum_i N(w_i \in T_1) + \sum_i N(w_i \in T_2) \quad (8)$$

However, this "comparison value" does not yet take into account the total number of words, so a "relative value" must be generated using the following formula:

$$R(T_1, T_2) = Comp(T_1, T_2) / \sum_i N(w_i \in T_1 \cap T_2) \quad (9)$$

The process is repeated to obtain $R(T_2, T_3)$ and $R(T_1, T_3)$ as relative comparison values. This process of using the combined inputs of the bodies of text themselves as well as previous results from within the algorithm is reflected in Figure 1 below.

Step 4: Create relative comparison values.

All of the steps are repeated for all possible combinations, to obtain the following (final) values:
RelativeComparisonValue(1) = The relative comparison value between text 1 and text 2.
RelativeComparisonValue(2) = The relative comparison value between text 2 and text 3.
RelativeComparisonValue(3) = The relative comparison value between text 1 and text 3.

3.2. Implementation

An overview of the implementation is shown in Figure 1. This Zipf's Law-based algorithm was designed to analyze large bodies of text. A program was then built using PHP to run these computations outlined in the above algorithm. This software analyzes three bodies of text and generates a value measuring the degree of similarity between all possible pairings, meaning that we are given a value for the degree of similarity between texts 1 and 2, texts 2 and 3, and texts 1 and 3. A higher value indicates a higher degree of similarity.

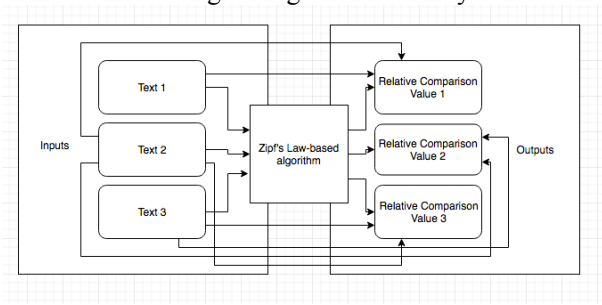

Figure 1 Zipf's law-based algorithm

3.3. Application of method

Zipf's Law suggests that while commonly used words (such as "the," "and," "or," etc.) will appear frequently in bodies of text, regardless of authorship, other words will appear significantly less often (such as proper nouns or other less commonly used words). Because of this, less frequently used words have more value in identifying patterns, because, by definition, these words appear less often than common ones. For example, if the works of two authors are being analyzed, seeing the word "the" in their work tells very little that is specific to one of the authors, because we would expect that both authors to use the word frequently. However, if one of the authors tends to use a much less common word (for example, xylophone) more frequently than the other, the appearance of that word could suggest a great deal about the authorship.

The evaluation is comparative across different genres with different degrees of linguistic components. Our method is used on a variety of genres, including haikus. Since haiku poems have linguistic components that are narrowly defined with a smaller number of words, we expect author identities to be more difficult to detect via the linguistic components in haiku poems. For comparison, we analyzed songs, which have a higher degree of structure, but less than haiku poems. Third, we considered online reviews, which have a much lower degree of structure. Finally, we analyzed poems, which have a lower degree of structure as well.

3.4. Selection of texts

To highlight extensibility and to isolate the structures present within text-based writings, texts were extracted from songs, haikus, online reviews and books. The individuals who extracted the text were not involved in the actual analysis and instructed to select works randomly. Although some degree of non-randomness occurs, due to limitations on the data (such as the need for writings by the same authors and their availability) the intent is that the data set represents an accurate reflection of the real-world context within which this analysis takes place.

3.5. Hypotheses

A goal is to determine whether pre-existing knowledge within linguistics can be confirmed using our method. One linguistic principle is whether the writings of authors are more similar to one another than to different authors. Each dataset of 3 separate works of texts generates 3 unique comparison values (one for each pair of works), so there is a 1/3, or 33%, chance that random chance would accurately identify which

two works were created by the same author. Findings that show this method's ability to correctly identify joint authorship across multiple genres, not just in books, would further highlight the extensibility of the method itself.

Because of this, we present the following hypotheses in order to appropriately test the method:

Hypothesis 1A: The songs written by the same author/artist should be correctly identified more than 33% of the time.
Hypothesis 1B: The reviews written by the same author should be correctly identified more than 33% of the time.
Hypothesis 1C: The haikus written by the same author should be correctly identified more than 33% of the time.
Hypothesis 1D: The books written by the same author should be correctly identified more than 33% of the time.

4. Results

4.1. Songs

Fifty datasets of three songs each were extracted by an individual instructed to select songs randomly from online sources. Perfect randomness was not possible due to the availability of data and the requirement that at least two of the songs be written by the same author/artist. However, the data is intended to be representative of the real-world context in which this type of analysis might take place. Within each dataset, texts A and B are works by the same author whereas text C is always by a different author.

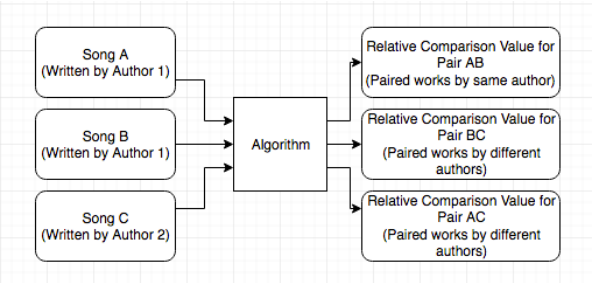

Figure 2 Examples of inputs and outputs

The highest value for the comparison indicates which two works the algorithm identifies as being the most similar. Thus, the comparison value for AB should the highest if the joint-authorship is properly identified, as a high value for AB and would indicate that texts A and B are the most similar. A result of BC or AC being the highest would be incorrect since C was written by a different author than A and B. The first five results of the analysis are given in Table 2.

Table 2 Comparison of analyzed tables

Dataset	AB (Same Author)	BC (Different Author)	AC (Different Author)
1	0.361044137	**0.437684826**	0.378155221
2	0.346523022	0.355811223	**0.369746338**
3	0.286866632	**0.407574696**	0.227307246
4	**0.421945449**	0.313599338	0.280545375
5	0.338712968	0.348575537	**0.477111064**

In total, 22 out of 50 pairs were correctly identified as being the work of the same author/artist, resulting in a probability of 0.44 or 44%, which is indeed higher than the probability that a correct result would have occurred through random chance. The results of the matched-pair analysis are displayed in Table 3.

Table 3 Comparison of matched pairs for songs

	AB (Same Author)	BC (Different Author)	AC (Different Author)
Amount	**22**	15	13
Probability	**0.44**	0.3	0.26

Dataset	AB (Same Author)	BC (Different Author)	AC (Different Author)
1	0.541827597	0.570669104	0.47316592
2	0.476357447	0.384955598	0.38377702
3	0.369754309	0.351143506	0.383070977
4	0.297540945	0.389610949	0.335790336
5	0.241538866	0.269408117	0.356709767

Since 0.44 is greater than the 0.33 probability that a correct result would have occurred through random chance, Hypothesis 1A is supported.

4.2. Reviews

Similar to the selection process for songs, fifty datasets of three reviews each were retrieved. The first five results of the analysis of each data set are shown in Table 4.

Table 4 Comparison of analyzed reviews

In total, 16 out of 50 pairs were correctly identified as being the work of the same author/artist, resulting in a probability of 0.32 or 32%, surprisingly lower than the

probability that a correct result would have occurred through random chance. The results of the matched-pair analysis are summarized in Table 5.

Table 5 Comparison of matched pairs for reviews

	AB (Same Author)	BC (Different Author)	AC (Different Author)
Amount	16	17	17
Probability	0.32	0.34	0.34

The results are nearly identical to what one would find by selecting the datasets randomly. The pair of reviews written by the same author was correctly identified only approximately one-third of the time and incorrectly identified approximately two-thirds of the time, suggesting that this method provided no support beyond that of random chance. The reasons for these results are unclear. Perhaps a larger sample would yield more conclusive trends, or this can be explained by the relatively small number of words commonly used in reviews. Hypothesis 1B is not supported.

4.3. Haikus

The Haiku Society of America (HSA) defines the structure of the Japanese haiku as either "an unrhymed Japanese poem recording the essence of a moment keenly perceived, in which Nature is linked to human nature. It consists of seventeen onji (Japanese sound-symbols)" or "a foreign adaptation of [the above]. It is usually written in three lines of five, seven, and five syllables" [24]. Others have defined "haiku" similarly, highlighting the consistency of the structure [27]. Since haikus have a brief and highly structured form, they are useful bodies of text with a consistent structure that can be used for comparisons.

Matsuo Bashō is a well-known haiku writer [27] whose haiku titled "Old Pond" is presented in Figure 2 in its original Japanese form, the romaji transliteration, and an English translation.

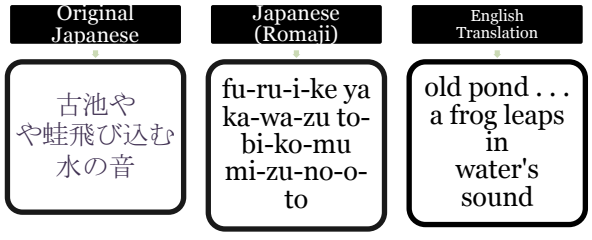

Figure 2 "Old Pond" [4]

These unique structural (and, to some extent, content-centric) characteristics provide an opportunity to eliminate the variance resulting from structure within forensic, and other, linguistic-type analysis. Since the structure of a haiku is rigidly defined, any author writing a haiku must produce a structure similar to that produced by all other authors who have ever written a haiku. Thus, this presents an opportunity for a technical analysis of the linguistic structural components within haikus while isolating other components. Haikus have been discussed for their unique structure and potential interplay with technology in speculative fictional works [29]. Haikus are one of the most rigidly defined forms of text. Since multiple haikus (at least those within the standard format) have the same structure and very similar word counts (due to the limitations on the number of syllables), identifying authorship of haikus is a unique challenge because one cannot rely only on the structural patterns that might be present, which further highlights the extensibility of the method.

Fifty datasets of three reviews each were selected. A limitation is that the number of haikus available in English is much more limited than the number of available songs or reviews. The first five results from analyzing the haikus data set are shown in Table 6.

Table 6 Comparison of analyzed haikus

Dataset	AB (Same Author)	BC (Different Author)	AC (Different Author)
1	**0.076576577**	0.149189189	0.149189189
2	0.1	**0.108**	0.064285714
3	0.107638889	**0.178888889**	0.149758454
4	0.141025641	0.141025641	**0.271634615**
5	0	**0.089093702**	0

Twenty out of 50 pairs were correctly identified as being the work of the same author/artist, resulting in a probability of 0.40 or 40%, which is higher than the probability that a correct result would have occurred through random chance. Interestingly, this group of datasets yielded a tie, likely due to the fact that, since the structure of haikus is so rigid and word usage is relatively limited, it is much more likely for three haikus to have no words in common than it is for three books, songs or reviews. This may partially be because the haiku structure allows the author to use more uncommon grammatical patterns, but it is unclear why this impact is so strong. The results are given in Table 7.

Table 7 Comparison of matched pairs for haikus

	AB (Same Author)	BC (Different Author)	AC (Different Author)	TIE
Amount	**20**	11	13	6
Probability	**0.4**	0.22	0.26	0.12
Probability (without ties)	**0.454545455**	0.25	0.295454545	N/A

Since 0.40 is greater than the 0.33 probability that a correct result would have occurred through random chance, Hypothesis 1C is supported. If ties are considered to be an "unable to identify"-type result rather than an "incorrectly identified"-result, they are excluded from the total and the probabilities recalculated. When this is done the probability for all other categories rises, resulting in an even higher probability of 0.4545, lending more support to Hypothesis 1C.

4.3. Books

Results of the analysis from the first five data sets, each comprised of three books, out of the fifty total datasets are shown in Table 8, with the largest value (meaning the two bodies of text are found to be most similar) listed in bold.

Table 8 Comparison of analyzed books

Data Set	AB (Same Author)	BC (Different Author)	AC (Different Author)
1	**0.880278**	0.857968	0.841821
2	0.778082	**0.890425**	0.751863
3	**0.883169**	0.728292	0.71815
4	**0.825125**	0.789261	0.753744
5	**0.890047**	0.780765	0.77339

The highest (bolded) value marks the comparison found to be most similar. To test the results, the number of times AB had the highest value was calculated with the results summarized in Table 9.

Table 9 Comparison of matched-pairs

	AB (Same Author)	BC (Different Author)	AC (Different Author)
Amount	**43**	5	2
Probability	**0.86**	0.1	0.04

Texts A and B are written by the same author, so our hypothesis would suggest that the value for AB should be the largest in the majority of cases. The value for AB was the highest 86% of the time, supporting Hypothesis 1. When combined, BC and AC were the highest only 14% of the time when random chance would have suggested around 66%. Hypothesis 1D is supported.

5. Discussion

The method is intended to provide a new form of analysis that could be designed and implemented to add useful surface-level data, contributing to modeling and comparing unstructured text. The research was motivated by work in computational linguistics and text analysis that recognizes the potential of massive amounts of text data for customer relations and other applications. The values generated represent structural data that is difficult to measure, thus, providing a comparison value that provides useful information beyond existing methods. With books, for example, the algorithm was tested against simple random chance and provided an accurate determination of authorship 53% more often than random chance. Application of the method identifies similarities between texts without necessarily having to read the content directly. This might be useful for linguistic forensics or translation software, if a big data-style sample of works, translated between two languages, were compiled and analyzed to assess the extent of the similarity.

Hypotheses 1A, 1C and 1D being supported supports the claim that this method is extensible in its application. In addition to being able to correctly identify joint authorship of books more often than random chance, it appears this method is also more accurate in terms of correctly identifying joint authorship of songs or haikus. The authorship issues resulting from reviews are unclear, requiring more research. The support for extensibility goes beyond the fact that the program can confirm well-known linguistic patterns. One of the challenges is whether it is possible to avoid the limitations on authorship identification based upon structural patterns. Whereas traditional authorship identification techniques rely on syntax and

other such patterns, this analysis focuses only on word usage and frequencies.

Some of the predictions are only slightly better than random selection. Hypothesis 1B, concerning reviews, was not supported. One possible reason for this lack of support is that the algorithm increases in accuracy alongside an increase in word count on the bodies of text being analyzed. This increase in accuracy means that it is easier to identify potential authorship on longer bodies of work and more difficult on shorter works, such as reviews. Further research should explore this correlation (or lack thereof) between word count and accuracy of predictions.

This is one such example beyond that of identifying common authorship and forensic linguistics in which this method may be useful. Having this additional data about the word patterns within bodies of text may be useful to integrate into a variety of models. Thus far, the data generated has primarily been presented as sufficient in its own right, but there is sufficient reason to believe that it could work well as supplementary data that serves not as a replacement to existing methods, but as a compliment to it. Other scholars may have the opportunity to adapt the method to other contexts beyond that which is described here or, as we have done, to new genres.

6. Conclusion

This research has proposed a method for comparing texts to identify those created by the same author. The method was implemented and tested. It is intended to be extensible and created for underrepresented applications such as haikus, arias, and foreign languages. The contribution is to successfully identify authorship without relying on traditional structural analysis. In contexts where the structure is uniform (e.g., homogenous groups) or not well-understood by outsiders (e.g., less frequently spoken languages), this could present opportunities for new forms of analysis and accuracy not previously possible if relying on structural patterns. The method has the potential to be effective in applications such as forensic linguistics. Additional genres (such as arias and blogs) will be analyzed in future research.

7. References

[1] Allen, J.F. 2003. Natural language processing.

[2] Baayen, H. 1992. Statistical Models for Word Frequency Distributions: A Linguistic Evaluation. Computers and the Humanities 26, 347-363.

[3] Barber, R. 2009. “Shakespeare Authorship Doubt in 1593, *Critical Survey,* (*21*:2, Questioning Shakespeare), pp. 83-110.

[4] Bashō, Matsuo. N.d. Old Pond.

[5] Binali, H., Wu, C., & Potdar, V. 2010. Computational approaches for emotion detection in text. In *Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on* (pp. 172-177). IEEE.

[6] Carr, N. G. 2008. *The big switch: Rewiring the world, from Edison to Google*. New York: W.W. Norton &. Co.

[7] Chowdhury, G.G. 2003. "Natural language processing." *Annual review of information science and technology* 37, no. 1: 51-89.

[8] Collins, M., & Duffy, N. 2002. Convolution kernels for natural language. In *Advances in neural information processing systems* (pp. 625-632).

[9] Collobert, R. and Weston, J. 2008. July. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). ACM.

[10] Conrad, J. 2011. "Interstate Rivalry and Terrorism: An Unprobed Link." *The Journal of Conflict Resolution,* (55:4), pp. 529-55.

[11] Cook, R.T. 2007. "Embracing the Revenge: On the Indefinite Extensibility of Language," in J.C. Bealt, ed., Revenge of the Liar, Oxford: Oxford University Press, pp. 31-52.

[12] Cook, R.T. 2009. "What is a Truth Value and How Many Are There?" *Studia Lógica*, (92), 183-201.

[13] Crossan, M. M., & Apaydin, M. (2010). A Multi-Dimensional Framework of Organizational Innovation: A Systematic Review of the Literature. Journal of management studies, 47(6), 1154-1191.

[14] Dickens, C. 1859. *A Tale of Two Cities*. Chapman & Hall.

[15] Dickens, C. 2011. *Eine Geschichte Aus Zwei Städten*. Insel Verlag.

[16] Ferrer i Cancho, R. and Solé, R.V. 2003. Least Effort and the Origins of Scaling in Human Language. Proceedings of the National Academy of Science of the United State of America 100, 788-791.

[17] Fielt E, Gregor S (2016) What’s new about digital innovation?. Information Systems Foundation Workshop, Canberra.

[18] Fichman, R. G., Dos Santos, B. L., & Zheng, Z. E. (2014). Digital Innovation as a Fundamental and Powerful Concept in

the Information Systems Curriculum. MIS Quarterly, 38(2), 329.

[19] "Free Ebooks by Project Gutenberg." Project Gutenberg. Accessed May 06, 2017. https://www.gutenberg.org/.

[20] Frey, B.S. 1987. "Fighting Political Terrorism by Refusing Recognition." *Journal of Public Policy* (7:2), pp. 179-88.

[21] Gao, L., Beling, P. A. 2003. "Machine quantification of text-based economic reports for use in predictive modeling," *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, pp. 3536-3541 vol.4.

[22] Ghosh, S., Samanta, D., and Sarma, M. 2012. "Cost of error correction quantification with Bengali text transcription," *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, Kharagpur, pp. 1-6.

[23] Gregor, S., & Henver, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly*, (27:2), pp. 337-355.

[24] Haiku Society of America, The Definitions Committee. 2004. *Official Definitions of Haiku and Related Terms* [Press release].

[25] Hill, B. 1970. Zipf's law and prior distributions for the composition of a population. Journal of American Statistical Association 65, 1220-1232.

[26] Hill, B. and Woodroofe, M. 1949. Stronger forms of Zip's law. Journal of American Statistical Association 70, 212-219.

[27] Hirshfield, J. 2011. *The Heart of Haiku.* Amazon Digital Services.

[28] Houtchens, B. C. 2001. "English in the News," *The English Journal*, (91:1), pp. 98-102.

[29] Howey, H. 2011. *The Plagiarist*. Amazon Digital Services.

[30] LaValle, S., Lesser, E., Shockley, R., Hopkins, M., & Kruschwitz, N. 2011. "Big data, analytics and the path from insights to value," *MIT Sloan Management Review*, (52:2), pp. 21-3.

[31] Lewis, D.D. and Jones, K.S. 1996. Natural language processing for information retrieval. *Communications of the ACM*, *39*(1), 92-101.

[32] Liddy, E. D. 2001. Natural language processing.

[33] Luna, L. 2013. "Indefinite Extensibility in Natural Language." *The Monist,* (96:2 Formal and Intentional Semantics), pp. 295-308.

[34] Mastora, A., Peponakis, M., & Kapidakis, Sarantos. 2017. *Journal of Information Science.* 43(4), pp. 492-508.

[35] Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. 2007. A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.

[36] Phillips, B.J. 2014. "Terrorist Group Cooperation and Longevity," *International Studies Quarterly,* (58:2), pp. 336-47.

[37] Portner, P.H. 2005. *What is Meaning? Fundamentals of Formal Semantics*. Malden, MA: Blackwell.

[38] Rau, L. F., Jacobs, P. S., & Zernik, U. 1989. Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*, *25*(4), 419-428.

[39] Rockwell, G., Sinclair, S. 2016. "The Measured Words: How Computers Analyze Text," in *Hermeneutica:Computer-Assisted Interpretation in the Humanities* , 1, MIT Press, pp.256-.

[40] Sarndal, C. 1967. "On Deciding Cases of Disputed Authorship," *Journal of the Royal Statistical Society*, (16:3), pp. 251-268.

[41] Schlenker, P. 2010. "Super Liars," The Review of Symbolic Logic, (3), pp. 374-414.

[42] Thomas, H. 1932. "The Shakespeare Authorship Controversy," *The British Museum Quarterly,* (*7*:2), pp. 40-41.

[43] Wand, Y., & Weber, R. 1990. Toward a theory of the deep structure of information systems. In J. DeGross, M. Alavi & H. Oppelland (Eds.), Proceedings of the Eleventh International Conference on Information Systems (pp. 61-72). Baltimore: ACM Press.

[44] Wells, S. 2014. *Why Shakespeare WAS Shakespeare*. Amazon Digital Services.

[45] Woodruffe, M. and Hill, B. 1975. On Zipf's Law. Journal of Applied Probablity 12, 425-434.

[46] Yoo, Y., Boland, R. J., Lyytinen, K., & Majchrzak, A. (2012). Organizing for Innovation in the Digitized World. Organization Science, 23(5), 1398-1408. doi:10.1287/orsc.1120.0771

[47] Zipf, G.K. 1935. *The Psychobiology of Language*. Houghton-Mifflin.

[48] Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.