Investigating Lexical and Syntactic Differences in Written and Spoken English Corpora

Mina Rajaei Moghadam
Northern Illinois University
mina.rajaei.moghadam@niu.edu

Mosab Rezaei
Northern Illinois University
mosab.rezaei@niu.edu

Miguel Williams
Northern Illinois University
miguel.williams@niu.edu

Gülşat Aygen
Northern Illinois University
gaygen@niu.edu

Reva Freedman
Northern Illinois University
rfreedman@niu.edu

Abstract

This paper presents an analysis of the differences between written text and the transcription of spoken text using current Natural Language Processing (NLP) methods. The purpose of the study is to investigate the long and rich history of attempts to differentiate spoken and written text in fields such as linguistics, communication, and rhetoric, which date back to the early 20th century. Given the availability of large quantities of machine-readable data and machine learning algorithms that can handle them, it is possible to use a large number of derived features. The research focuses on syntactic and lexical differences in written books and transcriptions of speeches by United States presidents. The analysis investigates morphological, lexical, syntactical, and text-level aspects. In this process, multiple features have been considered including lexical diversity, syllable count, frequency of parts of speech, and features relating to the parse tree, like the average length of noun phrases, and the use of interrogative sentences, among others. This study will enhance our understanding of the difference between written text and the transcription of spoken text in various disciplines including computer science, applied linguistics, communication, and similar fields.

Introduction

From Aristotle to the current date, disparities between speaking and writing form a long and important narrative. The more this topic is delved into, the more its relevance becomes evident in various disciplines, including computer science, linguistics, psychology, cognitive science, and others. Deriving a precise algorithm for this topic is challenged by several factors: the volume of data, variations in data collection, preference for different methodologies in different fields, different audiences for this type of research, and a large variety of available features.

Given the widespread collection of data through various online platforms in the form of transcribed audio and visual files, such research could significantly enhance the accuracy and categorization of data annotation, thus improving the training processes and precision of data labeling. This approach can not only improve the performance of the transcription models but also facilitate the conversion of spoken text into written form across different genres. Moreover, engaging in this research supports the language learning and acquisition process for children affected by speech or writing impairments and non-native speakers.

A deeper understanding of this subject helps to explain the differences between these two styles from philosophical, cognitive, and other perspectives. It is also vital for identifying features that have been incorrectly addressed or overlooked before deep learning models were able to autonomously recognize and infer these characteristics without a detailed understanding of the field. One challenge in this area is ensuring data fairness, meaning the availability of data within the same category in both speech and written formats. Our available sources for this research were books and speeches of American presidents.

In this paper, the focus is on syntactic and lexical differences in the transcription of speeches and written books of United States presidents. The main goal is to analyze these differences in relation to morphological, lexical, syntactic, and text-level aspects. Considering the long history of non-computerized research on this topic, computable features from previous studies have been selected, employing CoreNLP (Manning et al., 2014) and BERT (Devlin et al., 2018) as state-of-the-art tools for text analysis. Here, text is analyzed in different chunks, including word, sentence, and full document levels. In machine learning classification models, Support Vector Machines (SVM) and Random Forest (RF) have been utilized. Additionally, the BERT model was used to investigate whether an algorithm could distinguish between spoken and written without text using any predefined features. The specific research questions that we attempted to answer in this paper are:

RQ1: Which syntactic features are most useful in differentiating written text and transcription of speech?

RQ2: Which full-text lexical features are most useful in differentiating written text and transcription of speech?

RQ3: Which type of algorithm performs better on this task, feature-based machine learning or a large language model like BERT?

In the literature review, the long history of attempts to solve this problem from early 20th-century linguists to NLP researchers will be explored. In the procedure section, the methodology will be described. Finally, results, discussion, conclusions, limitations of our work, and a short description of potential future work have been provided. Furthermore, the artifacts of this work are accessible in an online repository.[1]

Related Work

Early 20th-century researchers like Woolbert (1922) explored differences between speaking and writing, emphasizing the roles of thought, language, and typography in writing, along with voice and action in speaking. Olson (1996) challenges the traditional view that writing is simply a transcription of speech and rather believes writing serves as a model for speech.

With a shift towards quantitative differentiation, Fairbanks (1944) and Mann (1944) employ methods like type-token ratios and frequency of part of speech analysis. Other researchers like Blankenship (1962) and Drieman (1962) examined tense, mood, and voice as well as features like syllable count and vocabulary diversity.

Later, O'Donnell (1974) noted differences such as longer syntactic units and more frequent usage of dependent clauses in written language. Einhorn (1978) considered seven stylistic features and found that speaking uses more personal references, shorter thought units, more repetition, more monosyllables, and more familiar words, with no significant difference in the variety of thought unit length.

Akinnaso (1982) highlighted syntactico-semantic differences and discussed matters of word choice and lexical structure, including variations in vocabulary richness, word frequency, and types of words. However, Akinnasso concluded that the distinction ultimately is a result of the communicative situation, the goals of the speaker and writer, and the context.

Biber (1986a,b) employs a multi-feature, multi-dimensional analysis, examining 41 linguistic features across a broad range of spoken and written texts. This approach reveals four underlying textual dimensions: interactional vs. informational focus, situational vs. abstract content, reference to a distant vs. immediate context, and opinionated vs. objective style. Biber's findings indicate that no single textual dimension or feature can be the reason for the similarities and differences.

Chafe and Tannen (1987), while comparing linguistic features and measuring involvement or detachment in both speaking and writing, highlight how these elements vary depending on context and usage. Researchers like O'Donnell (1974) and Poole and Field (1976) believe written language is more complex in structure than spoken.

DeVito (1966, 1967) finds that speaking employs more verbs and adverbs. It also contains more self-reference terms and quantifiers. Connors (1979) provided insights into the psychological and educational aspects of speech and writing. Chafe (1979) and Redeker (1984) claimed that spoken language is more complex and written language is more organized. Redeker sees higher involvement in spoken language, because of features like self-references, colloquial expressions, and direct quotes.

Written language, conversely, demonstrates detachment through structures like passive voice, indirect quotes, and literary expressions. In the perspective of DeVito (1966, 1967) spoken language tends to be more informal, personal, and involving immediate feedback, whereas written language is more structured and formal. Yet the exploration of these differences continues in contemporary studies like that of Liu (2023) that focus on formality in spoken versus written English, considering differences in lexical richness, grammatical structures, and so forth.

Biber (2020) notes that transcribed spoken language can be analyzed using corpus-based methods similar to those for written texts and points to the importance of developing spoken corpora with detailed annotations for exploring and comparing lexico-grammatical patterns in both spoken and written forms. Gray and Biber (2013) analyzed lexical frames in academic prose and conversation and showed that there are more recurrent lexical sequences in conversation than in academic writing. They illuminate how academic writing contains more function words while conversation relies more on verb-based frames. However, in their more recent research, Biber and Grey (2023) challenge the traditional view and suggest that the complexity of academic writing is due not only to longer sentences with embedded clauses but rather to its dense use of phrasal modification. They showed that both speaking and writing exhibit grammatical complexities. Speaking often incorporates a higher frequency of dependent clauses (complement, adverbial, and relative clauses, both finite and non-finite), and academic writing relies on phrasal embedding for complexity, including attributive adjectives, nouns as pre-modifiers, and prepositional phrases as post-modifiers.

Pangtay-Chang (2009) investigated text-based computer-mediated communication (CMC) and showed the blurry border between speaking and writing, suggesting the tendency of such written texts to move towards the speaking modality. Her study shows the use of pragmatic markers, like interjections and emoticons, to express emotions and structural elements such as greetings, topic shifts, and adjacency pairs, similar to face-to-face interaction on MSN Instant Messenger (IM) conversations.

Meanwhile, studies like Cleland and Pickering (2006) state that the same cognitive process of syntax construction occurs in both speaking and writing. Their study on syntactic priming across modalities supports the notion of a common syntactic representation and highlights the influence of verb repetition on priming effects, as syntactic structures are reused. Rezaii (2022) finds the balance between lexical complexity and syntactic simplicity as a fundamental property of language. Therefore, she believes the syntax-lexicon tradeoff observed in spoken language also exists in written language. The study also notes how familiarity with terms and topics is an influential factor as reducing the cognitive load allows for the retrieval of more complex syntax and indicates a shared cognitive process in language production.

[1] https://github.com/mosabrezaei/lexical-and-syntactic-analysis

In NLP research, Freedman and Krieghbaum (2014) focus on using machine learning techniques to analyze educational dialogues and student responses, considering features like percent of nouns, adjectives, adverbs, and prepositional phrases. In later work, Freedman and Krieghbaum (2015) used syntactic features like frequency of parts of speech to help distinguish writing styles. Wright and Freedman (2017) utilized the Stanford Parser (Klein and Manning, 2003) to differentiate quotations from narration, analyzing sentence length in addition to parse tree height and frequency of conjunctions. They found shorter sentences and less subordination in dialogue. In Freedman (2017), syntactic and bag of words approaches are used to differentiate sections of the book of Isaiah.

Procedure

This study aims to revive the inquiries in this field using newer computational methods, leveraging a larger dataset and applying recent work in machine learning to identify the most significant and computable syntactic and lexical features to find the differences between speaking and writing. This will help to determine whether these features can effectively lead to differentiating between spoken and written language.

Dataset

The dataset used in this investigation consists of transcriptions of spoken language from United States presidents, derived from the Miller Center of Public Affairs University of Virginia (2022), which include data ranging from George Washington's time to the contemporary presidency. The dataset also includes three complete books written by presidents obtained from Project Gutenberg (nd), and 10% of each of 10 additional books.

Presidents' books and speeches were selected as our dataset to minimize the influence of variables such as subject, content, and degree of formality. This approach reduces potential biases arising from personal background including education, age, and other social constructs that affect people's oral and written production. Therefore, the dataset is limited to the texts belonging to similar subjects across different modalities. Given that our subjects are presidents, their public speeches and written texts are produced with a common purpose and for the same audience, resulting in similar content and genre. Uniformity in subject matter is thus assured by focusing on materials either authored by the individuals themselves or officially endorsed as their personal statement.

All numbers, currency values, commas, and hyphens are removed to ensure they are not counted as words in the text complexity calculations. Any instance of more than one white space is replaced with a single white space. Also, the nltk library (Bird et al., 2009) is utilized to chunk the data into groups of 512 tokens, to remove the effect of length disparities in the metrics. The dataset comprises 13,600 spoken samples and 13,600 written samples.

Features

The features listed in Fig. 1 were identified in prior studies and have shown their relevance and effectiveness across various text levels. Their computability is crucial as it ensures they can be quantitatively assessed and consistently applied across different datasets which enables a systematic analysis that ultimately leads to the development of scalable and robust computational models. These features also cover multiple levels of text analysis, from units smaller than a word to larger units such as sentences. This is to ensure a holistic understanding of the textual data which acknowledges the complex and layered nature of language. These extracted features are categorized as morphological, lexical, syntactic, and text-level features.

Morphological aspects:
- Average syllables per word
- Average words per sentence
- Average characters per word

Lexical aspects of text:
- Lexical diversity
- Readability

Lexical aspects of sentences:
- Number of words in a sentence
- Percentage of POS (verb, adjective, noun, adverb, coordinators)
- Percentage of personal pronouns (first, second, and third)

Syntactical aspects:
- Frequency and percentage of subordinate clauses
- Depth of parse tree
- Frequency and percentage of noun phrases
- Average length of noun phrases
- Yes/no questions
- Direct wh-questions

Text-level aspects:
- Sentences

Figure 1: Lexical and Syntactic Features Used

Morphological aspects here include the average number of syllables per sentence, which is used in various text complexity and lexical variation formulas. Additionally, the average number of syllables per word, which is more precisely counted as a morphophonological feature, is defined as the ratio of the total number of syllables in a text to the total number of words. This measure gives insight into the word length and complexity within a given text since texts with higher average syllables per word are generally considered to be more complex. Moreover, the average number of characters per word is calculated as the total number of characters (excluding spaces) divided by the total number of words. This gives insight into the relationship between lexical complexity and word length since texts with higher average characters per word use longer and potentially more complex words.

With regard to the lexical aspects, text complexity and lexical diversity were examined. Text complexity refers to the level of difficulty encountered when reading a piece of text. Numerous metrics are used to calculate text complexity. Lexical diversity is the ratio of unique words to the total number of words. In general, higher lexical diversity suggests a broader vocabulary and less repetition of words, which can enrich the content and style. The grandfather of lexical diversity scores is the type-token ratio (TTR) (Johnson, 1939) which divides the number of unique words in a text by the total number of words in the text. TTR is a simple metric to calculate but is affected by the length of the text, i.e., the longer the text, the lower score you will obtain. Two lexical diversity metrics, root type-token ratio (Guiraud, 1959) and corrected root type-token ratio (Carroll, 1964), that were previously thought not to be affected by length in fact are (Torruella and Capsada, 2013). In an attempt to combat the length issue, more sophisticated metrics from Covington and McFall (2010) and Yule (1944) were added. The Python package LexicalRichness (Shen, 2022) was used to obtain lexical diversity measures.

Readability measures are quantitative measures that use a plethora of linguistic features to calculate the complexity of text. These features include word count, syllable count, and the total number of sentences. A large number of readability measures were used, including Flesch Kincaid Grade Level (Kincaid et al., 1975), SMOG Index (McLaughlin, 1969), and Gunning Fog Index (Gunning et al., 1952). To generate readability measures we use the Python library Textstat (Bansal and Aggarwal, 2022).

To parse the sentences we used CoreNLP. This parsing model employs a context-free grammar, along with associated probabilities for each rule, to generate a parse tree for each sentence. The token and sentence boundaries and other features provided by CoreNLP help in the analysis process. Figure 2 presents an example of a parse tree generated by CoreNLP for the sentence "I walk slowly, but I never walk backward."

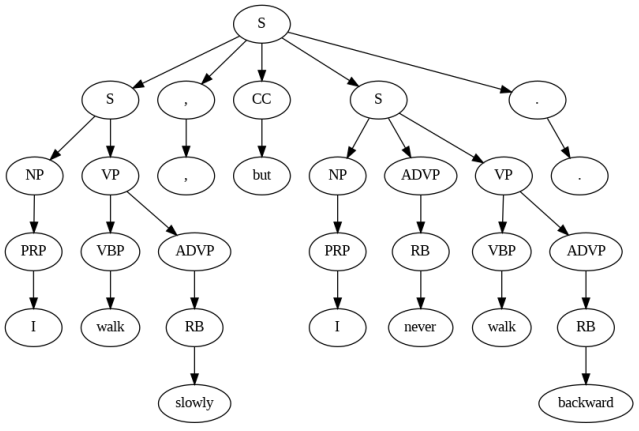

Figure 2: Parse Tree of a Quote Attributed to Abraham Lincoln

Regarding the lexical aspect of sentences, the number of words in a sentence was used as an aid to understanding syntactic complexity, since longer sentences often indicate more complex ideas or more detailed information. The analysis includes part-of-speech tags such as verbs, adjectives, adverbs, nouns, and coordinators. Analyzing the POS distribution can reveal structural, stylistic, and functional aspects of the text. We roll up the multiple types of nouns and verbs provided by CoreNLP into one type for each. We calculate the percentage of personal pronouns per sentence, including both third-person (she/he/it) and first-person (I/we) pronouns because they shed light on the author's voice and the level of formality and informality of texts. We also need to investigate syntactic properties at the sentence level, especially dependent clauses. The most used subordinating conjunctions include "when", "while", "after", "before", "because", "if", "unless", and "so that" (Aygen, 2016).

Identifying subordinators enables the calculation of the percentage of subordinate clauses. This provides insight into sentence structure as well as complexity. Additionally, the parse tree provides the depth of the syntax tree directly. The depth of the parse tree represents the complexity of sentence construction, with deeper trees indicating more complex syntactic structures. The combination of these features offers a clearer picture of the complexity of a sentence. We also compute the percentage of noun phrases and the average length of noun phrases in a sentence. These two help to better understand the level of detail in descriptions and other information in the text.

We compute the frequency of interrogatives, including yes/no questions and wh-questions. The presence of yes/no questions can influence the text's interactive quality or the directness of the information presented. This question type was specifically chosen to prevent confusion with other kinds of questions, such as rhetorical ones. Additionally, counting direct wh-questions (who, what, when, where, why, how) helps to avoid including subordinators. We also counted the number of sentences to see if that could provide meaningful information separating written and transcribed spoken text.

While the first two research questions employ sets of features, the BERT algorithm does not use features as it uses the written and transcribed spoken text directly.

Experiments

In this work, three experiments were conducted. In the first experiment, we used machine learning to derive significant features from the parse trees of the individual sentences. In the second experiment, we used machine learning to identify significant features from a long list of calculations of lexical diversity and readability. In the third experiment, we used the large language model BERT to differentiate between writing and speaking.

In the first experiment, we started by extracting a wide variety of features from the original sentences. The core of the feature extraction is based on the parse tree. For the classification task, we used two machine learning models, SVM and Random Forest, due to their effectiveness in text classification tasks and their assistance in showing the importance

of each feature (Gasparetto et al., 2022). We dropped features that have a high correlation with other features, including character count, verb count, adverb count, noun count, coordinator count, and subordinator count.

In the second experiment, after calculating all of the text complexity metrics, we discarded the outliers for each feature by removing any data element more than three standard deviations from the mean. We applied Random Forest models to the set of text complexity metrics. These metrics included ari, coleman liau, dale chall, difficult words, flesch, fleschkincaid, gunning, hdd, herdanvm, linsearwrite, simpsond, smog, text standard, yulek, dugast, herdan, mattr, msttr, mtld, rttr, summer, ttr, and yulei from the textstat package.

Then we added average sentence length and average word length. We checked the correlation between each pair of features and removed any that had an absolute correlation greater than 0.5. Of the correlated features we decided to keep average sentence length and average word length as they had the greatest impact on the models. Additionally, we decided to keep maas (Maas, 1972) because it is one of the lexical diversity metrics that is not affected by text length.

For the third experiment, the BERT model, one of the earliest and most well-known deep neural network models for text classification, was used. The BERT model was trained with the sentences from the original data. This is our initial attempt to train a deep neural network model to distinguish between transcribed speeches and written books by US presidents. The accurate performance of the model can demonstrate that this task is feasible, suggesting that there may be significant features to extract that could enhance the results of other machine learning models.

Results and Discussion

The first experiment demonstrated that the Support Vector Machine (SVM) and Random Forest (RF) models achieved 0.54 and 0.61 accuracy respectively on the syntactic features. Furthermore, as shown in Table 1, which reports the other evaluation metrics, the RF model demonstrates considerably better performance than SVM in both classes. Although there were other models like Decision Tree that performed better than SVM, the RF model exhibited significantly higher accuracy in all metrics.

Furthermore, the amount of influence of each syntactic feature is shown in Fig. 3. The four most important features are the length, percentage of nouns, percentage of verbs, and depth of parse tree. On the other hand, the four least important features are the wh-questions, yes/no questions, first-person pronouns, and second-person pronouns.

In the second experiment, Table 2 shows the performance of the Random Forest model on the lexical features. Using the random forest method on just the set of complexity metrics achieved 72.22% accuracy. Adding average sentence length and average word length improved the accuracy of the model to 79.20%. We removed the highly correlated features and retrained the model, obtaining a significant improvement in performance of 87.40%. Figure 4 shows that the most important features are word length, average sentence length, and maas. This shows that simple metrics such as word length and sentence length are better at distinguishing speeches from written text than complex methods.

Table 1: Evaluation of Syntactic Models and BERT

	Labels	Precision	Recall	F1
SVM	Spoken	58.6%	24.3%	34.4%
	Written	52.2%	82.7%	64.0%
RF	Spoken	60.9%	61.0%	61.0%
	Written	61.0%	60.9%	60.9%
BERT	Spoken	89.9%	90.4%	**90.1%**
	Written	90.6%	90.1%	**90.3%**

Table 2: Evaluation of Random Forest Lexical Models

Labels	Precision	Recall	F1
Spoken	94.2%	92.9%	93.6%
Written	91.2%	92.8%	92.0%

	Accuracy
Random Forest (RF)	72.2%
RF with Avg Sent & Word Len	79.2%
RF without Correlated Features	87.4%
RF with Only Avg Sent & Avg Word Len	**92.9**%

We performed a t-test using the Bonferroni correction on each of the features. The significance of 0.05 becomes 0.0166 to reject the null hypothesis that the means of the features are equal. Table 3 shows that every feature still has a p-value significant enough to reject the null hypothesis.

The last study we performed on the chunked dataset was to remove all lexical diversity and text complexity metrics, leaving only average word length and average sentence length. This study produced the highest accuracy of 92.91%, due to the fact that maas is a function of the remaining two features.

Table 3: Hypothesis Testing for Lexical Features

Feature	p-value
maas	1.95e-9
Average Sentence Length	1.31e-5
Average Word Length	1.78e-4

The third experiment illustrates that BERT achieved an accuracy of 90% using the ktrain library (Maiya, 2020). The batch size and max features hyperparameters were set to 6 and 35000, respectively. BERT, unlike SVM and RF, performed accurately on both spoken and written text, as shown in Table 1.

To test if the text complexity metrics are a viable method for distinguishing speeches from written text, we utilized random under-sampling to obtain an equal quantity of each type. We used an 80/20 split of training and test data. We trained a random forest model on each dataset to check the feature importance of each model. The only hyperparameter we set is a max tree depth of 15.

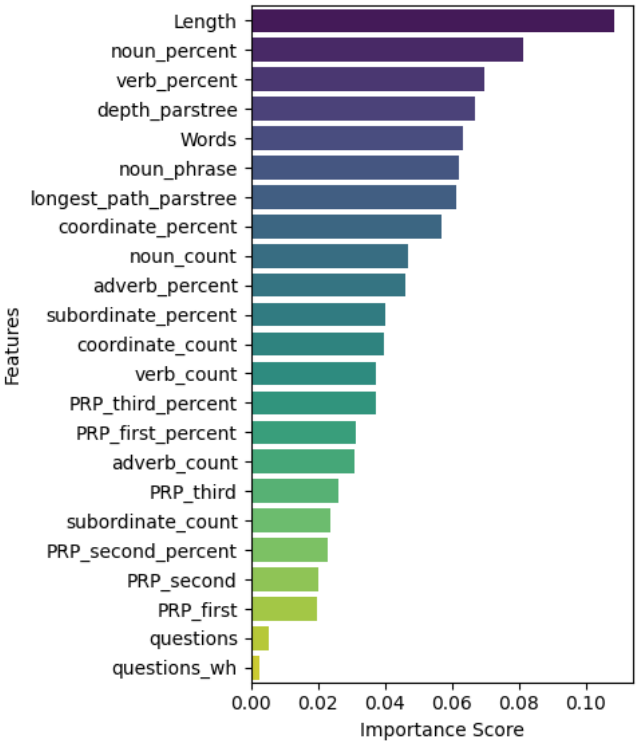

Figure 3: Feature Importance for the Syntactic Features

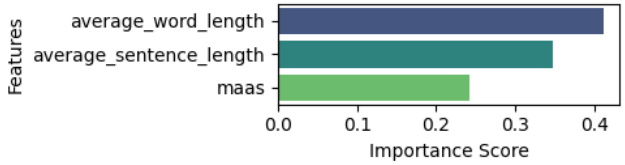

Figure 4: Feature Importance for Lexical Features

Conclusions

In this paper, we analyzed the difference between written books by US presidents and transcription of some of their speeches.

In our first experiment, we used syntactic features derived from parsing sentences. We analyzed the data using SVM and Random Forest. We explored a wide variety of features, including morphological, lexical, syntactic, and text-level features, to evaluate which ones exert the most influence on machine learning models for this task. The most relevant features were sentence length, percentage of nouns, percentage of verbs, and depth of parse tree. These results show that it may be possible to distinguish spoken from written text using syntactic features.

In the second experiment, we examined lexical metrics such as text complexity and lexical richness using the full text. The results show that average word length and average sentence length were more useful than any of the many lexical richness and text complexity metrics for differentiating these two types of texts since the random forest model performs better when we exclude all text complexity metrics.

The third experiment showed that BERT had notably higher accuracy than any other models in this study. This indicates that such differentiation is feasible without the use of features. However, machine learning models might still be helpful in identifying specific features relevant to distinguishing between written text and transcribed spoken text.

Limitations

One of the limitations of this study was accessibility to primary sources, specifically presidential books. Therefore, the goal is to actively expand samples to incorporate a wider array of books to enrich the dataset and do further analysis. By leveraging more extensive data alongside state-of-the-art analytical techniques, it is anticipated to have more precise identification and understanding of the patterns distinguishing these two forms of discourse.

Another limitation of this research process was coding some of the syntactic and lexical features. Some features require complex coding and suffer from a lack of consensus on their definitions, including time and aspect, prepositional phrases, and particles. Future research efforts will aim to refine these methodologies, seeking a balance between comprehensive linguistic analysis and practical feasibility in coding and interpretation.

Future Work

The reported results open new areas for the next steps in future work. First and foremost is increasing the number of extracted sentences. Also, we would like to add more features known by linguists to be important, especially more types of coordinating and subordinating conjunctions. Adding these features will not only help to increase the accuracy of the models but also provide a human-understandable explanation of the differences between speaking and writing.

Due to fast-growing studies in the deep neural network area, especially in explainable artificial intelligence (XAI), it would be worthwhile to utilize and develop these concepts to better understand the features that will enhance the performance of the deep neural model.

References

Akinnaso, F. N. (1982). On the differences between spoken and written language. *Language and Speech*, 25(2):97–125.

Aygen, G. (2016). *English Grammar: A Descriptive Linguistic Approach*. Kendall Hunt, third edition.

Bansal, S. and Aggarwal, C. (2022). Textstat. Accessed: March 29, 2024.

Biber, D. (1986a). On the investigation of spoken/written differences 1. *Studia Linguistica*, 40(1):1–21.

Biber, D. (1986b). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, 62(2):384–414.

Biber, D. (2020). Corpus analysis of spoken discourse. *Pronunciation in Second Language Learning and Teaching Proceedings*, 11(1).

Biber, D. and Grey, B. (2023). Is conversation more grammatically complex than academic writing? In *Grammatik und Korpora 2009: Dritte Internationale Konferenz*, pages 47–61.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly.

Blankenship, J. (1962). A linguistic analysis of oral and written style. *Quarterly Journal of Speech*, 48(4):419–422.

Carroll, J. B. (1964). Language and thought. *Reading Improvement*, 2(1):80.

Chafe, W. and Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology*, 16(1):383–407.

Chafe, W. L. (1979). Integration and involvement in spoken and written language. In *2nd Congress of the International Association for Semiotic Studies.*, pages 195–215. Vienna.

Cleland, A. A. and Pickering, M. J. (2006). Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language*, 54(2):185–198.

Connors, R. J. (1979). The differences between speech and writing: Ethos, pathos, and logos. *College Composition and Communication*, 30(3):285–290.

Covington, M. A. and McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

DeVito, J. A. (1966). The encoding of speech and writing. *Communication Education*, 15(1):55–60.

DeVito, J. A. (1967). A linguistic analysis of spoken and written language. *Communication Studies*, 18(1):81–85.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Drieman, G. H. (1962). Differences between written and spoken language: An exploratory study. *Acta Psychologica*, 20:36–57.

Einhorn, L. (1978). Oral and written style: An examination of differences. *Southern Journal of Communication*, 43(3):302–311.

Fairbanks, H. (1944). II. The quantitative differentiation of samples of spoken language. *Psychological Monographs*, 56(2):17–38.

Freedman, R. (2017). Can natural language processing help identify the author(s) of the book of Isaiah? In *30th International FLAIRS Conference*, pages 297–300.

Freedman, R. and Krieghbaum, D. (2014). Effects of rewriting essays on linguistic measures of complexity. In *25th Annual Meeting of the Society for Text and Discourse*.

Freedman, R. and Krieghbaum, D. (2015). Comparison of expert tutors through syntactic analysis of transcripts. In *Artificial Intelligence in Education*, pages 574–577. Springer.

Gasparetto, A., Marcuzzo, M., Zangari, A., and Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. Information, 13 (2), 83.

Gray, B. and Biber, D. (2013). Lexical frames in academic prose and conversation. *International journal of Corpus Linguistics*, 18(1):109–136.

Guiraud, P. (1959). *Problèmes et méthodes de la statistique linguistique*. Presses Universitaires de France.

Gunning, R. et al. (1952). *The Technique of Clear Writing*. McGraw-Hill.

Johnson, W. (1939). *Language and Speech Hygiene*. Institute of General Semantics.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training, University of Central Florida*.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pages 423–430.

Liu, Y. (2023). Differences between spoken and written English. *Communications in Humanities Research*, 3:757–761.

Maas, H.-D. (1972). Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 2(8):73.

Maiya, A. S. (2020). ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*.

Mann, M. B. (1944). III. The quantitative differentiation of samples of written language. *Psychological Monographs*, 56(2):39–74.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

McLaughlin, G. H. (1969). SMOG Grading–A new readability formula. *Journal of Reading*, 12(8):639–646.

Miller Center of Public Affairs University of Virginia (2022). Presidential speeches: Downloadable Data. Accessed: 2022-03-17, Available at https://data.millercenter.org.

O'Donnell, R. C. (1974). Syntactic differences between speech and writing. *American Speech*, 49(1/2):102–110.

Olson, D. R. (1996). Towards a psychology of literacy: On the relations between speech and writing. *Cognition*, 60(1):83–104.

Pangtay-Chang, Y. (2009). IM conversations in Spanish: Written or oral discourse? *Illinois Language and Linguistics Society 1 (ILLS)*.

Poole, M. E. and Field, T. (1976). A comparison of oral and written code elaboration. *Language and Speech*, 19(4):305–312.

Project Gutenberg (n.d.). Project Gutenberg. Retrieved February 21, 2016, from https://www.gutenberg.org.

Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, 7(1):43–55.

Rezaii, N. (2022). The syntax-lexicon tradeoff in writing. *arXiv preprint arXiv:2206.12485*.

Shen, L. (2022). Lexicalrichness: A small module to compute textual lexical richness. https://github.com/LSYS/lexicalrichness.

Torruella, J. and Capsada, R. (2013). Lexical statistics and tipological structures: A measure of lexical richness. *Procedia–Social and Behavioral Sciences*, 95:447–454.

Woolbert, C. H. (1922). Speaking and writing—A study of differences. *Quarterly Journal of Speech*, 8(3):271–285.

Wright, M. and Freedman, R. (2017). Syntactic differentiation in Oscar Wilde's "Dorian Gray". In *Proceedings of the 28th Modern Artificial Intelligence and Cognitive Science Conference (MAICS)*, pages 204–206.

Yule, C. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.