Alec Wall

Alexus Lopez

Divya Mohan

Info 190: Computational Humanities

December 16, 2020

Exploring Spoken and Written Language through Computational Analysis

TABLE OF CONTENTS

Introduction

American historian Walter J. Ong wrote of orality and literacy: “Our understanding of the differences between orality and literacy developed only in the electronic age, not earlier. Contrasts between electronic media and print have sensitized us to the earlier contrast between writing and orality” [1]. There is a large body of research investigating the distinctions between spoken and written language. Yet we found none that explored a quantitative approach in the 21st century, finding much of this research expresses a context of racist, classist, and imperialist ideas without acknowledging such. We may direct our attention to the work of Jennifer Lynn Stoever, particularly *The Sonic Color Line*. Speech has developed under the context of a sonic color line,

"race's audible contour" [2]. "The listening ear enables the key dichotomies of the sonic color line… it normalizes the aural tastes and standards of white elite masculinity as the singular way to interpret sonic information" [2]. From the representation of dialect to the weight of aural imagery, a sonic color line has permeated literature and other arts of communication throughout history. With this in mind, we take a critical look at distinguishing oral and written language. Instead of attempting solely to distinguish differences between written and spoken language, we will explore the degree of orality of texts from different time periods. As Ong wrote in Orality and Literacy, "Diachronic study of orality and literacy... sets up a frame of reference in which it is possible to understand better not only pristine oral culture and subsequent writing culture, but also the print culture that brings writing to a new peak and the electronic culture which builds on both writing and print" [1]. We seek not to weigh the importance of one over the other. We instead seek to explore the validity of features that have historically been argued to distinguish between oral and written language and quantitatively explore the evolution of language over time. Through this work, we seek to introduce quantitative evidence of the evolution of the relationship between oral and written language, and to encourage others to analyze the evolution of this relationship through similar means.

Throughout our analysis, we aimed to answer the following questions: what features separate oral and written language, and can we use these features to accurately classify them? What are the implications of these differences? Could we observe a change over time in the relationship between spoken and written language, and the features that differentiate the two? What does its development over time imply about oral and literacy culture? With particular attention to historical research in the field, we assessed the validity of historical conclusions to a contemporary corpus by employing methods of computational linguistics. We analyzed features

that differentiate spoken language from written language by building a classifier, and compare our results with the work of previous research. We also explored the efficacy of each feature in determining spoken language in our contemporary corpus and our historical corpus, and the implications of their relationships. By assessing the accuracy and output of our classifier on historical literature, we intend to emphasize the evolution of spoken and written language as well as their relationship.

Literature Review

Our methods drew largely from the work of Gisela Redeker, who published a research paper on a small experiment observing the differences between the spoken and written language used by eight female undergraduates from one house of the Berkeley University Students Cooperative Association in 1984. Redeker explores the validity of earlier research, writing, “According to Chafe (1979), [a greater level of engagement in spoken language] is evidenced (in English) in... speakers' self-references and references to their mental processes, use of direct quotes and historical present,... monitoring devices..., evidentials..., vagueness, and hedges” [3], delineating the prevalence of these features in her observations. We ultimately chose to analyze the validity of Redeker’s arguments that evidentials, hedging, and certain parts of speech are more common in spoken language. We did not, however, choose to analyze the importance of any part of speech in particular. Along with assessing the conclusions made by Redeker, we constructed a feature to analyze syntactical sentence length, as proposed as a significant indicator by the work of Roy C. O’Donnell in 1974 [4].

Data Source

We chose to focus on the Corpus of Contemporary American English (COCA) for our research to align with the historical body of literature we have explored that wholly regards

American English. We narrowed our scope with the added intention of analyzing the validity of the conclusions of the foundational texts mentioned in our literature review to contemporary sources. The Corpus of Contemporary American English (COCA) contains more than 385 million words from 1990–2008 (20 million words each year), balanced between spoken, fiction, popular magazines, newspapers, and academic journals [5]. Using the free sample of 8.9 million words, we focused on the "spoken" and the other scripted dialogue samples.

Texts from Project Gutenberg served as our representation of historical literature [6]. We encountered much difficulty in searching for a thoroughly balanced, publicly available corpus that was compatible with our data from COCA. Project Gutenberg provided texts in a raw UTF-8 format, which proved to be the most compatible format we found for the data collected from COCA for classifier training. We have selected a sample in part supplied by David Bamman's publicly available LitBank, an annotated dataset of 100 works of English-language fiction from Project Gutenberg [7]. LitBank provides a temporal range of texts from 1719 to 1922. It is also important to note that this source contains some instances where multiple works were written by the same author, which we have sustained. To extend our temporal range, we specifically searched Project Gutenberg for its more recent texts. Our added texts increased our sample of historical pieces from 100 texts to 112 texts, and increased the temporal range of our Gutenberg texts to include pieces from 1719 to 1970.

To promote reproducibility, we have included more information on accessing the COCA corpus and what data we used for our analysis. Of the three main directories, we have chosen to use the data from `coca-samples-text`. An overview can be found in Table 1:

```
(chF20) .../coca-samples-text$ ls
text_acad.txt  text_fic.txt  text_news.txt  text_tvm.txt
text_blog.txt  text_mag.txt  text_spok.txt  text_web.txt
```

Table 1.

There is one file for each category, and each of these files is broken down into smaller articles, separated by a new line character and starting with a unique @@[ARTICLE_ID] described in Table 2:

```
@@17141 ERIC @!BURNS , FOX NEWS HOST : On this week 's " FOX New...
@@21741 qwq @ ! DOUGLAS-FORD-ARSO : I set the fire at the fire s...
...
```

Table 2.

For our analysis, we decided to solely classify against texts from COCA's fiction subcorpus, as our sample texts from Project Gutenberg are all classified as fiction. Quick statistics on these files, spoken and fiction, can be found in Table 3.

Filename	# of Articles	Average # of words per article
`text_spok.txt`	`263`	`4626.6`
`text_fic.txt`	`274`	`5284.1`

Table 3.

Methods

Our approach to the exploration of spoken and written language was largely centered around the use of a linear, logistical regression classifier made with Scikit-Learn. Before feeding the data into the classifier, we pre-processed the texts to represent our four main features of analysis. After reading related papers in the literature field, the features we focused on were sentence length, parts of speech, the presence of hedges, and the presence of evidentials. We were able to gather information about sentence length and parts of speech using the pre-trained `en_core_web_sm` model from the `spacy` library. Our feature analyzing hedging presence was largely based off of the work of Ulinski, Benjamin, and Hirschberg in "Using Hedge Detection to Improve Committed Belief Tagging" [8]. We included the list of hedges they provided in addition to a list of hedges provided by Titus (wooorm) on Github [9]. Our evidentials feature

drew from the definitions found in Erika Berglind Söderqvist's "Evidential Marking in Spoken English : Linguistic Functions and Gender Variation", however, we had to generate a list of evidentials on our own [10].

Evaluation Strategy

Using Bamman's Jupyter Notebooks as reference, we used the `train_test_split` function from `sklearn.model_selection` in order to split our COCA data into 80% training data and 20% testing data. We also seeded all random functions for the sake of reproducibility. Using the `linear_model.LogisticRegression` from `sklearn` to generate a model which is trained using the 80% training set, we calculated the accuracy of this model using the model's corresponding `score(test_X, test_Y)` function. We then applied this model to our Gutenberg texts with the intention of measuring the "orality" of texts from 1719 to 1970. This was executed by assessing the proportion of texts that were classified as spoken in addition to analyzing the distribution of the probabilities of a given text being classified as spoken. Due to the computational limits of running `spacy` on large texts, we took 10 random samples from each Gutenberg text, each sample being 10,000 characters long.

Results

After training the model on the COCA spoken and fictional text articles, the classifier was able to classify the 20% test set with an accuracy of 91.7%. Table 4 includes the list of most strongly-weighted features that the classifier uses:

Class 1: fic	Class 2: spok
9.795 hedge_could	16.321 hedge_can
8.952 POS_VERB	14.874 POS_PROPN
8.810 hedge_around	9.716 POS_PUNCT
8.572 hedge_like	8.990 hedge_think
8.154 hedge_thought	8.164 hedge_about
7.390 POS_X	7.628 hedge_really
6.876 hedge_says	6.101 hedge_many
5.991 sentence_length_397	5.924 POS_AUX

5.852 hedge_would	5.905 evidential_think
5.670 sentence_length_378	5.738 sentence_length_74
5.044 sentence_length_320	5.483 hedge_will
4.937 POS_NOUN	5.303 evidential_say
4.880 POS_DET	5.199 sentence_length_64
4.803 hedge_might	4.999 evidential_obviously
4.781 sentence_length_808	4.866 hedge_kind of
4.637 hedge_read	4.785 sentence_length_196
4.436 sentence_length_62	4.682 hedge_much
4.416 sentence_length_239	4.563 sentence_length_315
4.399 sentence_length_49	4.095 sentence_length_241
4.398 evidential_I hear	4.002 sentence_length_457

Table 4. First 20 features organized by weight on classifier.

The classifier uses a variety of the four features. Interestingly, we found that some of our hedges corresponded most heavily with our written texts. We can also observe that unique parts of speech are highly associated with both of our categories, verbs as a top identifier for fiction and proper nouns for spoken pieces. Proper nouns were not identified as a particularly important part of speech distinguishing spoken language in Redeker's work, nor were punctuation and auxiliary verbs. Long sentence lengths dominate the identifying features of fictional texts, in accordance with O'Donnell's conclusions. Lastly, the presence of evidentials proved to be an effective marker of spoken language, shown by the absence of any evidentials in over 30 of the strongest weighted features to describe written texts and their presence in describing spoken language.

We then used this model to predict the classification of the Gutenberg dataset. We initially predicted that most of the texts would be classified as fiction, given the fictional and written nature of the Gutenberg texts. However, for all the collected Gutenberg texts, the model classified 136 samples as spoken and the remaining 942 samples as fictional, leading to an overall accuracy of 87.4%, a number quite close to 91.7%, the accuracy of our classifier when tested on the 20% test set from COCA. However, we are still interested in determining if there is a temporal pattern in the samples that were closer classified to spoken.

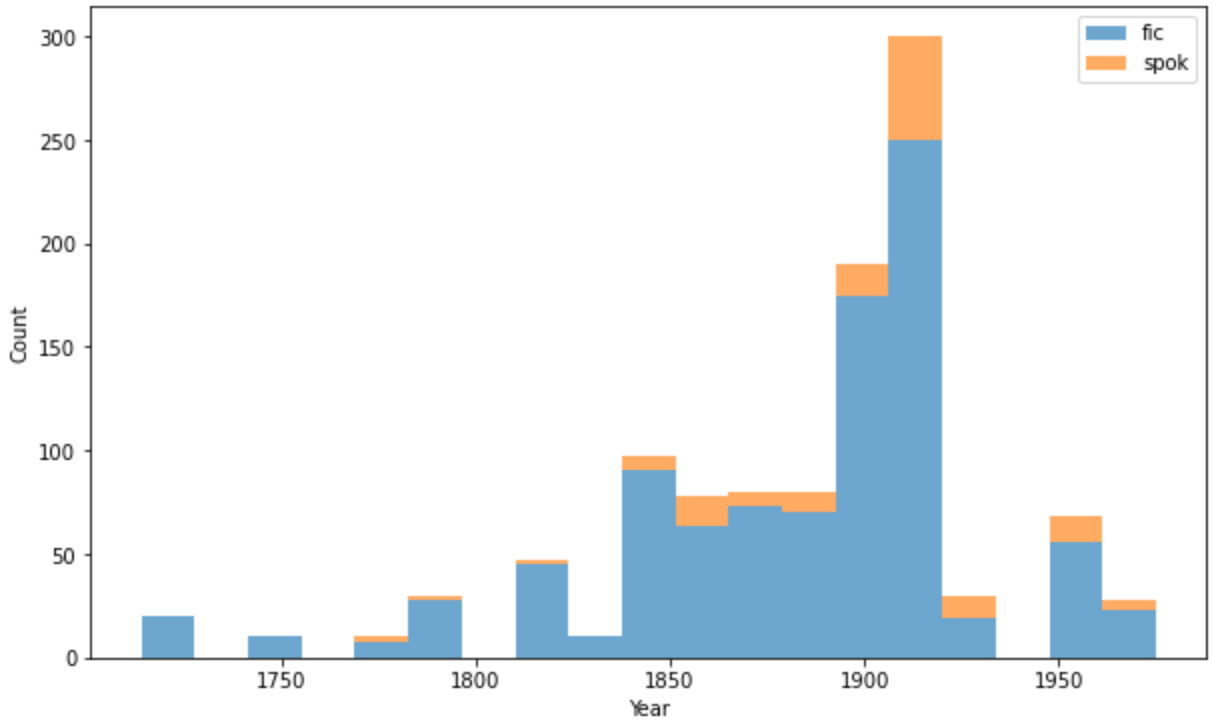

Diagram 5.

Diagram 5 is a histogram plotting the distribution of the number of Gutenberg samples classified as fictional and spoken against the year that the corresponding Gutenberg text was published. As can be seen from the diagram, a majority of the samples were classified as fictional over spoken. Almost none of the samples from texts published during the 1700s were classified as spoken, but as the years progressed, more samples were classified as spoken. However, since the data in this diagram is not normalized, it is hard to distinguish if the *proportion* of samples classified as spoken increased or not. This diagram highlights one of the limitations of our Gutenberg dataset: a majority of our texts are from a time period between 1850 and 1923, and we should have selected a few more texts from before and after this time period.

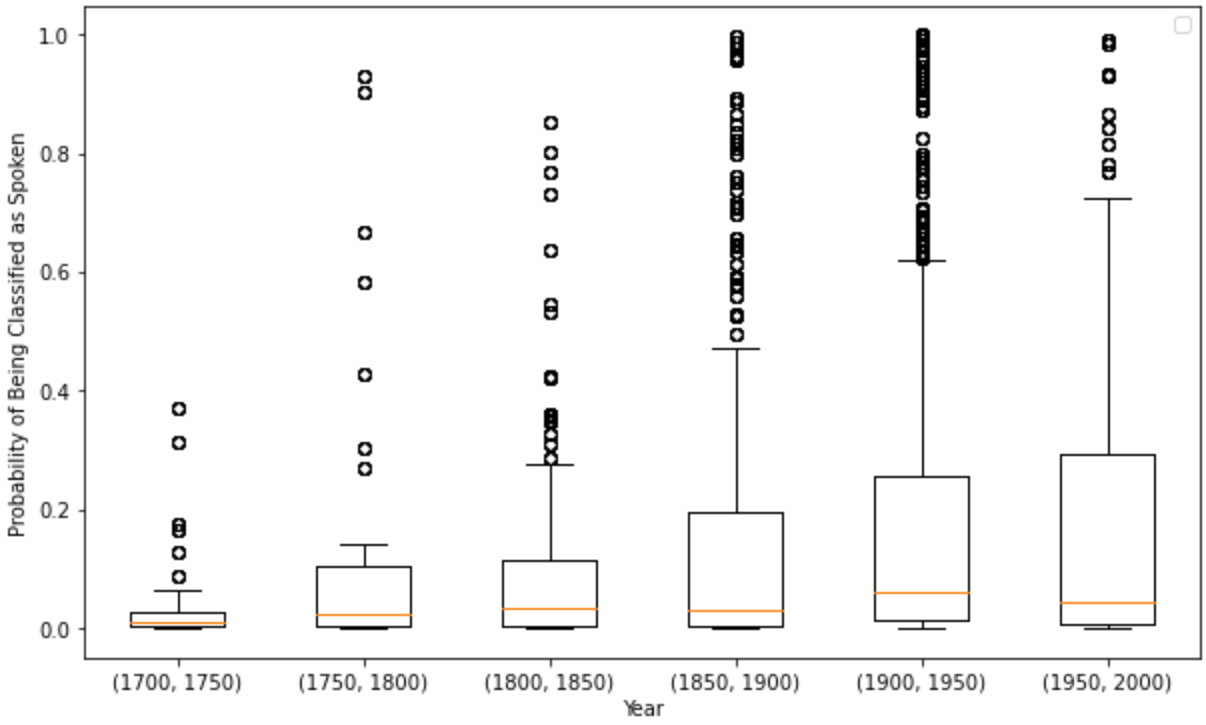

Diagram 6.

The above diagram is a series of boxplots representing the distribution of probabilities that the sample would be classified as spoken, each boxplot corresponding to an increasing range of time periods. These probabilities were calculated using the `predict_proba` method from the `sklearn.linear_model.LogisticRegression` model. A clear trend can be seen that the quartiles of each distribution increases over time periods. The medians of each time period are 0.0095, 0.0236, 0.0320, 0.0315, 0.0612, and 0.0440 from left to right, with a peak median probability of a sample being classified as spoken of 0.0612 during the time period between 1900 and 1950. In addition, we can visually assert that the other quartiles, particularly note the 75th quartile and top whisker, for each box plot consistently increase as the time periods increase as well. These quartiles inherently describe more "normalized" information regarding the classification of spoken and fiction because a median represents the value at which roughly 50%

of values within the distribution is greater than or equal to the other values in that time period's time distribution.

Discussion

Concerning our methods, we would like to discuss particular areas of weakness in our data collection and computation. First of all, it is worth mentioning the vast landscape of genres that constitute literature to this date. We narrowed our scope to observations between fictional literature and spoken language because of the lack of publicly available data from varied genres, especially from historical texts. This analysis should be applied to various genres to explore the extent of differences between spoken and written languages. Furthermore, our evaluation strategy lacked computations to argue significance, such as A/B testing and/or other methods to determine if a temporal sample is unrepresentative of the population of text samples. Additionally, our research is not a cohesive analysis of all features that have historically been noted as identifiable markers of spoken and written language. We encountered many obstacles in deciding which features we could include under our time constraints and scope of knowledge. For example, Ulinski, Benjamin, and Hirschberg's research included guidelines on how to additionally implement hedge detection with a rule-based approach. However, due to time constraints, we could not include their rules in our analysis, which may have affected the importance of some hedges in identifying fiction. Many other linguistic markers outlined in Redeker's work could have proven useful to the development of our classifier, such as self-references, colloquial expressions, and monitoring. Ultimately, there is a plethora of features that could be incorporated into a classifier such as ours, but could not be included under our time frame.

In assessing the validity of historical conclusions regarding the differences between spoken and written language, we found that O'Donnell's conclusions were more heavily supported by our research [4]. In the context of Redeker's research, our research supports the importance of evidentials [3], but our implementation of hedges were not as fruitful. Additionally, the parts of speech that were found to distinguish spoken language in Redeker's work were not emphasized in ours. Though it may be attributed to our methods, we may still argue that these differences are more evidence to the evolution of spoken language over time, and, further, the relationship between spoken and written language over time.

Regarding the meaning of the decreased accuracy of our classifier applied to our historical texts, we investigated the outliers of our conclusions to determine possible patterns impacting accuracy. We manually found that these outliers generally consisted of a large amount of dialogue between characters. Within the Gutenberg samples that were classified as spoken, the samples contained 71.2 quotation marks per sample on average. While quotation marks alone are not indicative of the author writing dialogue between characters, dialogue is often dictated by quotation marks. We also manually verified that most texts with few quotation marks contained dialogue dictated in different formats (i.e. using single quotations instead of double quotations, starting the speech of one character with --, dialogue written in transcription format, etc). We additionally noticed that a few of the samples that were classified as spoken were written in first-person, with references to others in a conversational fashion.

Moreover, there are many conversations regarding the relationship between oral and written cultures we could not analyze through a computational lens. One such conversation is that we as speakers of American English live in a primarily written culture, and linguistic analysis centers around our ability to produce our ideas as text. As we can't analyze the full

extent of differences between spoken and written language, we must look to writers like Walter Ong. Ong, in *Orality and Literacy*, details the psychodynamics of oral cultures [1]. Ideas such as referencing, the ability to look back on text, are not conceived of in an oral culture. Furthermore, words in an oral culture carry a different kind of meaning to the speaker. Words and ideas are stored as sounds and images, which for a speaker of an oral culture carry a heightened sense of visual acuity. If knowledge is not passed onto someone else, it is simply lost there. Because of this, there is an emphasis on giving information to a select individual in the group to hold, a completely different system of information dissemination. Written cultures conversely store their knowledge and findings in text, with little to no importance placed on an individual to carry all of the information single-handedly; this difference has resulted in unparallel styles of communication, ill-suited for computational analysis at this time.

Conclusions and Future Work

Although our methods require much development, the conclusion of this research provides some support to argue the evolution of American English in spoken and written contexts, as well as the evolution of the relationship between spoken and written language over time. Generally speaking, we observed a trend toward an increased probability of being classified as a spoken piece over time. This observation can be used as evidence of the evolution of American English in literature toward our contemporary speech styles and patterns. As for the implications and/or reasons for this trend, we may look toward research in cultural analytics, such as Damon R. Young's "Ironies of Web 2.0" [11] or Abigail De Kosnik's *Rogue archives: Digital cultural memory and media fandom* [12]. Within these scholars' research we find an unprecedented cultural system of written art and communication styles has materialized as a

result of online communication. It is safe to argue that our style of communication, oral and literate, has evolved at least in part as a result of emergent online culture.

To further this research in the context of our scope, it may be useful to explore the validity of grammar and vocabulary as classifying features in more work, as explored previously by Lester Harrell [13] and Joseph A. Devito [14], respectively. However, it is important to make such observations with particular sensitivity to historical viewpoints on the implications of grammar and vocabulary differences, and to bring in the conversations of Jennifer Lynn Stoever and Ana María Ochoa Gautier [15] in their explorations of the implications of the development of our linguistic system centered around literacy.

Reaching toward research in a broader scope, we may see a difference in the relationship between written and spoken language should this research be implemented on corpora outside of fiction, or among particular genres of fiction. Again, textual culture has evolved with the popularization of online communication. Social media platforms have allowed many more writers to enter the public sphere, and with this accessibility, new types and standards of writing. Future work in this field may include the analysis of the relationship between spoken language and Twitter or Facebook posts, or between spoken language and Archive of Our Own works, should one desire to focus on contemporary relationships. Should this work emphasize a temporal analysis, it may be worth analyzing speech data from various time periods and their relationships with literature from those time periods, and/or to compare speech data across those various time periods.

Our Jupyter Notebook and related work can be found at https://github.com/21dmohan/info190_final

References

[1] Ong, Walter J. *Orality and literacy*. Routledge, 2013.

[2] Stoever, Jennifer Lynn. The sonic color line: Race and the cultural politics of listening. Vol. 17. NYU Press, 2016.

[3] Redeker, Gisela. "On differences between spoken and written language." *Discourse processes* 7.1 (1984): 43-55.

[4] O'Donnell, Roy C. "Syntactic differences between speech and writing." American Speech 49.1/2 (1974): 102-110.

[5] Davies, Mark. "The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights." *International journal of corpus linguistics* 14.2 (2009): 159-190.

[6] Project Gutenberg, www.gutenberg.org/.

[7] https://github.com/dbamman/litbank

[8] Ulinski, Morgan, Seth Benjamin, and Julia Hirschberg. "Using hedge detection to improve committed belief tagging." *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*. 2018.

[9] https://github.com/words/hedges

[10] Berglind Söderqvist, Erika. *Evidential marking in spoken English: Linguistic functions and gender variation*. Diss. Department of English, 2020.

[11] Young, Damon R. "Ironies of Web 2.0." (2019).

[12] De Kosnik, Abigail. Rogue archives: Digital cultural memory and media fandom. mit Press, 2016.

[13] Harrell, Lester E. “A Comparison of the Development of Oral and Written Language in School-Age Children.” *Monographs of the Society for Research in Child Development*, vol. 22, no. 3, 1957, p. 1., doi:10.2307/1165494.

[14] Devito, Joseph A. “Psychogrammatical Factors in Oral and Written Discourse by Skilled Communicators.” Speech Monographs, vol. 33, no. 1, 1966, pp. 73–76., doi:10.1080/03637756609375483.

[15] Ochoa Gautier, A. (2014). Aurality: Listening and Knowledge in Nineteenth-Century Colombia. Durham (N.C.): Duke University Press.