Spring 5-7-2021

Analyses and Creation of Author Stylized Text

Keith Carlson
Keith.E.Carlson.GR@Dartmouth.edu

ANALYSES AND CREATION OF AUTHOR STYLIZED TEXT

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

in

Computer Science

by

Keith Carlson

GUARINI SCHOOL OF GRADUATE AND ADVANCED STUDIES

Hanover, New Hampshire

May 2021

Abstract

Written text is one of the major ways that humans communicate their thoughts. A single thought can be expressed through many different combinations of words, and the writer must choose which they will use. We call the idea which is communicated the *content* of the message, and the particular words chosen to express the content, the *style*. The same content expressed in a different style may tell something useful about the author of the text (e.g., the author's identity), may be easier to understand for different audiences, or may evoke different emotions in the reader.

In this work we explore ways that the style of writing can be used to make inferences about the author and demonstrate applications where these techniques uncover interesting results. We supplement the analytic approach with a synthetic approach and consider the problem of generating text which matches the style of a target author. To this end we find and curate suitable parallel datasets of the same content written in different styles. These are – to the extent possible – made publicly available. Next, we demonstrate the performance of machine translation systems on this data. Finally, we show settings in which modifications to existing machine translation architectures can improve results and even perform style transfer in an unsupervised setting.

Acknowledgements

Throughout my PhD program I have been fortunate to be able to rely on the support and guidance of many people.

I'd like to thank my advisor Daniel Rockmore for his extraordinary guidance, enthusiasm, and patience. Without his mentorship this document would not exist.

I'd like to thank my committee members Soroush Vosoughi, V.S. Subrahmanian, and Allen Riddell for agreeing to take the time from their busy schedules to serve on my committee and for offering guidance on ways to improve this work.

I have had the honor of collaborating with many great researchers during my time at Dartmouth and truly appreciate my coauthors. In particular I would like to thank Dan Rockmore, Allen Riddell, Michael Livermore, and Ramon Lecuona Torras for their expertise and insights which have broadened my research into interdisciplinary projects I would not have had the chance to experience working on my own.

My peers and professors at Dartmouth also deserve recognition. There are countless, brilliant people in this community who I have learned from or who asked meaningful questions to guide me.

Outside of the academic world I have relied on wonderful friends and family through this process for whom I am very grateful. In particular I would like to thank my parents, Todd and Alyson, and sisters, Erin and Lauren, for their support and love.

Finally I would like to thank my wife Delaina for her unwavering belief in me. It

has been a long journey but she has been there for me the entire way. I know I have made her, myself, and hopefully everyone else who invested so much time into me and my work, very proud.

Contents

Abstract ii

Preface iii

1 Introduction **1**

2 Stylometry of the U.S. Supreme Court **7**

2.1 Introduction 7

2.2 Data 11

2.3 Preliminary Analyses 12

2.3.1 Productivity 12

2.3.2 “Friendliness” 14

2.3.3 Defensiveness 17

2.4 Stylistic Fingerprint 19

2.5 Time Trends in Judicial Style 22

2.6 Potential Mechanisms 30

2.6.1 Prior decisions 31

2.6.2 Partisan affiliation 34

2.6.3 Substantive factors 36

2.7 Clerk Influence 39

2.7.1 Comparing inter-year variability 40

2.8 Conclusion 46

3 Generating Stylized Text 49

3.1 Background 49

3.1.1 Online Reviews 52

3.1.2 Wine Reviews 54

3.2 Data 56

3.3 Model and Experiments 59

3.4 Results and Analysis 64

3.5 Human Evaluation Survey 66

3.6 Conclusions 70

4 Style Transfer in Text 73

4.1 Task and Background 73

4.2 Metrics 77

4.2.1 BLEU and PINC 77

4.2.2 Frequent Idiosyncratic Words 78

4.3 Datasets 79

4.3.1 Previously Used Datasets 79

4.3.2 The Bible Dataset 81

4.4 Supervised Style Transfer in Text 86

4.4.1 Data Splitting for Bible Style Transfer 87

4.4.2 Moses 88

4.4.3 Seq2Seq 89

4.4.4 Experiments 92

4.4.5 Conclusions 95

4.5 Unsupervised Style Transfer in Text 97

4.5.1 Background 97

4.5.2 Data Splitting 98

4.5.3 Baseline System 100

4.5.4 Model with Content Embeddings 100

4.5.5 Results 101

4.5.6 Conclusions 103

4.6 Conclusion 104

5 Discussion 106

References 110

List of Tables

2.1 Top Ten Highest and Lowest Friendliness Scores 16

2.2 Content-Free Words 21

2.3 Spectral clustering analysis 27

2.4 Historical Influence. '*' indicates p<0.05, '**'p<0.01 33

2.5 Similarity scores between SCDB topic areas. '*' indicates p<0.01, '**'p<0.001 38

2.6 Chow test on centroid distance 42

2.7 Clerk Influence on Consistency 44

2.8 Chow test on consistency score 45

3.1 Examples of Formatted Parallel Data Used 57

3.2 Reviews Generated with Different Decoder Settings 64

3.3 Model Generated Reviews with Increased Ratings Along with the Human Review 65

3.4 Model Generated Reviews with Decreased Ratings Along with the Human Review 66

3.5 Reviews Generated with Different Authors Targeted by the Model (Along with the Human Review) 67

3.6 Panel regression with review-level random effects and individual-level fixed effects 69

4.1 Names of publicly available Bible Versions and other versions we used followed by their standard abbreviations in parenthesis. Text was collected from Biblegateway.com (and BBE from www.o-bible.com). . . 83

4.2 BLEU scores between full text of Bible versions. The verses of the version of each row are treated as the candidates and the column version's verses are treated as the reference. 85

4.3 Pairings of Bible versions created. For each pairing, parallel training, test, and development files are created and number of lines in each are reported. 88

4.4 The BLEU and PINC scores of the best Moses and best Seq2Seq ("Sq2Sq") models for each test set. The best model is defined here as the one which achieves the highest BLEU score on the test set. ("Pub" is Public.) 93

4.5 Examples of Moses and Seq2Seq output which showcase Moses making fewer modifications to the source sentence. 94

4.6 Examples which show Moses and Seq2Seq outputs which have correctly made changes to better imitate the target style or which have missed some obvious part of the transfer. 96

4.7 Our partition of Bible books into divisions. 99

4.8 The BLEU (PINC) scores of the unmodified source and the output of each model for each test book. All scores are when translating from Bible version YLT to Bible version BBE. 102

4.9 Words which are frequent in YLT but do not appear in BBE 103

4.10 Examples Outputs of each of the systems with YLT source and BBE target (validation). The verses are Hebrews 2:6 and Hebrews 5:9. . . 103

List of Figures

2.1 Productivity over time (excluding sitting Justices). 13

2.2 Time trends in opinion length 14

2.3 Friendliness score by median year 16

2.4 Grade level by median year 18

2.5 LANS graph of stylistic similarity between Justices 25

2.6 Similarity scores between Justices, as a function of time 29

2.7 Average similarity and temporal distance 29

2.8 Consistency Scores by Median Year 44

3.1 Category of Wines in the Data by Year 58

3.2 Distribution of Rating of Wines in the Data 58

3.3 Representation of the Transformer network with actual example input and output. Source: http://jalammar.github.io/illustrated-transformer/. Used with permission. 60

3.4 Transformer Loss During Training 63

4.1 MDS representation of the (top) BLEU comparisons for the public bibles and (bottom) Burrows Delta distances between the same. . . . 86

4.2 A Diagram of the Experimental Workflow 92

4.3 "XLM + Content" model training on the MLM objective. Based on Figure 1 of [89]. 101

Chapter 1

Introduction

The ability to communicate complex ideas is a unique and essential part of being human. It is hard to imagine modern civilization having arisen without the ability to share our knowledge and experience with one another. While much of this communication takes place face-to-face through words, gestures and expressions, written text has allowed these connections to occur even across great physical and temporal distance. For centuries, writing has been a vehicle for the transmission and sharing of knowledge across time and space. In more recent human history, the Internet has extended its reach. Although expanding the range of our communication networks, textual exchanges lack some of the richness brought by the intonation, facial expressions and gestures of in-person discussion. The loss of these auxiliary channels during our interactions means that even if two people are saying the same things through text and speech it is possible that some of differences in emotions, context, etc. may be lost. Such loss can lead to misunderstandings of the meaning or the intention of the speaker, increasing the likelihood "toxic online disinhibition" [93]. Readers have likely noticed that conversations across the internet can tend towards hostility more than conventional talks [154].

In text, we have only words available to communicate our meaning. While speaking we have other supplemental channels available through which we often signal subtler things about our history, intention, relationship to our conversation partner, etc. This loss, however, does not have to be complete. Given a “message”, there are many ways to write a sentence capable of conveying the embedded information, even when they are all written in the same language. Sentences can communicate essentially the same information but do so using different combinations of words. While sharing the same semantic content, these choices are not necessarily interchangeable. When constructing a sentence we frequently consider not only the semantic content we wish to communicate, but also the manner in which we express it. Different wording may convey different levels of politeness or familiarity with the reader, display different cultural information about the writer, be easier to understand for certain populations, etc. Consideration of the way we communicate our thoughts is important during verbal discussions, but even more so in writing where many auxiliary channels are not available.

This choice between semantically equivalent but linguistically distinct phrasings underlies the concept of *style* in text. This is in distinction to the *content* of prose which is the original thought which the author wished to communicate, whereas the style is the particular way in which they communicate it. The style of text still contains information from which a careful reader can learn something. Through stylistic choices, the author can intentionally or unintentionally reveal their background, intentions, emotions or even their view of the reader. Style can also affect the amount of information understood by the reader, for example by using simple words for non-native speakers or avoiding technical language for non-experts, but doesn’t change the underlying “message” the writer wishes to communicate.

This definition of “style” has some necessary vagueness to it. There are aspects of text such as formality and simplicity which contribute to style, but none of which captures the entire idea. An attempt to list all such relevant aspects would surely be missing some elements. This broad definition allows us to consider any of these aspects individually, but also affords a more holistic interpretation when required.

With this definition of style in mind, and the belief that style of text is becoming increasingly important, this work focuses on how computational methods can be applied to the analysis of style in text and then to the creation of stylized text.

The statistical analysis of style in text has often focused on the problem of author attribution in which the author of an anonymous or disputed piece of writing needs to be identified. This practice predates modern computers [113, 177] but computers (and text digitization) have greatly expanded the range of available techniques while also reducing the amount of work required for application of older methods. Famous examples of computer-assisted stylometry include the analysis of the authorship of The Federalist Papers [120] and Shakespearean plays [17]. But stylometric analysis is only one side of the coin of stylistics. On the flipside is the generative problem: how does one generate text of a particular style? As increasingly machines are used to write text, the challenge of writing in a particular voice is both interesting and somewhat disturbing. On the one hand, it could be used as a work assistant to the writer able to train a machine to write in their voice, but it could also be used for more nefarious schemes.

In this thesis we contribute new work to both of these dimensions – the analytic and the generative – of machine stylistics. In our first contribution, in chapter 2 we present a summary of a case study analyzing opinions written by the United States Supreme Court using various stylometric measures [25]. Among our findings is the discovery of evidence supporting the hypothesis that modern decisions are more likely to be written by clerks (rather than the justices themselves) than in previous eras. Our analysis here was a first step toward a much broader use of modern text analysis tools in the context of legal analysis to which we refer the reader to the original papers for details [29, 26, 23, 100].

We then focus on the creation of text targeting specific styles. The techniques used here build on methods employed for other natural language processing(NLP) tasks[156, 10, 163, 89]. These methods generally require large human-created datasets for training and evaluation of the models[186, 78]. To this end we identify suitable, underutilized data sources which represent different writing styles. The first such dataset is a collection of wine reviews and corresponding information about the wine being discussed. These 201,431 reviews were gathered from winemag.com and were published from 1999-2016. The second is made up of 34 translations of the Bible into English created from 1599 to modern day. In versions of the Bible, pre-existing alignment of the text due to the standardized book, chapter, and verse labels further increases suitability for NLP tasks.

With the wine review dataset, we address the task of generating text in chapter 3, in this case, in the “style of wine reviewers". We adapt modern neural network text-to-text systems to this problem by providing metadata of the wine and reviewer as input and train with the reviews themselves as the targeted output [24]. Style can be controlled by changing some parts of the input such as review author and rating given to the wine. We find that this approach produces reviews which survey participants were unable to distinguish from the reviews written by humans.

While these generated texts are conditioned on the input, the content of the review is created by the system. In many applications text needs to be created which has a specific meaning but written in a targeted style. Examples include periodicals or advertisements which seek to have a single voice, or ghost writers whose writing needs to be made to match a specific author’s style. We next focus on this problem of *style transfer*, the task of rewriting a sentence such that we preserve the meaning but alter the style, in chapter 4. We provide the Bible data to two modern machine translation systems, one statistical and one neural, and train them in a supervised manner to reproduce provided verses in the style of another targeted version. We find for both systems that the created verses are more stylistically similar to the target than the unmodified original is. We verify this with automatic evaluation metrics and also note several qualitative indicators[27].

The Bible is a useful dataset because it is parallel and aligned, but most text is not so clean. Many examples of text written in a style that we would like our models to target don't have parallel data, for example the works of most authors. To approach this task, we use the Bible data, but treat it as non-parallel and perform unsupervised training. First, we take an existing neural network architecture used for unsupervised machine translation and train it to perform style transfer. Then we show that a modification to this architecture, in the form of separate embeddings for human-assigned content categories, improves the output [28].

The major contributions of this dissertation are:

- Showing that the style of text matters, as information independent of the content can be communicated, via a case study of Supreme Court opinions [25].
- Identification of large datasets of stylistically distinct texts (in the form of Bible versions [27] and wine reviews [24] which have advantages over more commonly used corpora.
- Demonstrating that a machine translation neural architecture can be made to generate novel text of high quality if it is provided with appropriate information about the target [24].
- Showing that machine translation methods can be used for style transfer in text, where the meaning of the input is preserved, but the style is changed to match

a specified target style [27].

- Demonstration of unsupervised neural machine translation architecture applied to style transfer [28].
- Introduction of new embeddings in neural architecture which take advantage of the differences between machine translation and style transfer to improve performance compared to unmodified system[28].

The work in this dissertation has also provided the tools and foundation for other work including:

- An examination of U.S. appellate court decisions where we find evidence that decisions on publication are affected by political affiliation of the judges [26].
- An analysis of the formal structure of the United States Code (USC) compared to the results of a topic model trained on it [23].

Chapter 2

Stylometry of the U.S. Supreme Court

Section 2.1

Introduction

As mentioned in chapter 1, analysis of the style of human-produced writing has a history of successfully producing something like the writerly “fingerprint" of an author [120, 17]. In this chapter, we will show ways in which stylometric methods – developed and deployed in the context of legal writings – are used to uncover other kinds of information that have new and interesting implications in the context of legal studies. Specifically, we present our work that produced the first general quantitative investigation of writing style on the U.S. Supreme Court [25, 29].

The written word is the medium through which the law travels: courts, agencies, and legislatures create law by producing text. In recent years, these legal texts have increasingly become available to the public in digital form. Together with advances in processing power, data storage, machine learning, and computational text analysis, the digitization of the law has opened a new frontier in empirical legal scholarship, and a number of researchers have eagerly crossed over into this unexplored territory in search of new insights, methods, and questions.

While judicial writing style often serves as fodder for commentary, it has rarely been subject to systematic study. Systematic qualitative analysis is made difficult by the sheer bulk of the corpus, which prevents a human reader from digesting any more than a tiny sample. Perhaps for this reason, historically, qualitative analysis of style tends to focus on the "gems" in judicial writing, examining the prominent writings of prominent justices and neglecting the mine-run of workaday opinions[169]. Scholars have only relatively recently combined accessible digital versions of the corpus of judicial writing with the tools offered by computational text analysis to undertake quantitative analysis of style.

Prior to our work attempts to analyze quantitatively judicial writing style have typically been based on relatively small datasets, and are limited to a small number of specific stylistic features [15]. Our work makes use of a corpus containing all opinions written by the U.S. Supreme Court in the period 1792 to 2008, compiled from publicly available raw textual data that has been augmented with identifying information concerning the year, author, and opinion type [1]. In addition to examining specific stylistic features culled from the prior literature on judicial writing style, we deploy a general stylistic measure that serves as a proxy for what Judge Posner refers to as "style as signature [135]." This general style proxy is first used to examine the gradual change in writing style over time, and then to investigate potential hypotheses about sources of stylistic variation over time. Our most striking finding suggests that the institution of the modern clerkship appears to have had an important effect on judicial writing style on the Court, both in the consistency of writing style in individual

[1] Jonathan Ashley, research library at the University of Virginia, was primarily responsible for identifying resources, collecting cases and providing the markup needed for analysis. We are extremely grateful for his efforts.

chambers and in the consistency of writing style of the Court as an institution.

Our primary analysis relies on a commonly used measure of writing style based on the frequency of use of content-free words (also called "function words"). This measure provides a "useful stylistic fingerprint". This approach has been used to great effect in other studies, notably for the large-scale study of literary style executed by Hughes, Foti, Krakauer, and Rockmore [72] (see that paper for other related references). This stylistic approach has its roots in statistical approaches to the problem of author attribution. As will be discussed more thoroughly, our stylistic fingerprint measure allows for analysis of the similarity between texts or a group of texts, including texts that are grouped by time and by author. In our analysis, we use the stylistic fingerprint as a means of developing descriptive statistics and as the basis for testing a number of hypotheses concerning the evolution of judicial writing style in the Supreme Court.

We first address the general relationship between style and time. The starting place for this analysis is the intuitive hypothesis that there is a "style of a time" in the Court. Stated somewhat more formally, the hypothesis is that as the distance in time between judicial writings increases, there is a lower likelihood that they will be stylistically similar. Our analysis finds that stylistic similarity decreases with distance in time, as expected.

We also examine potential mechanisms that could drive this robust temporal trend. We examine the possibility that the writing style of particularly influential Justices propagates over time, so that the most read and cited Justices tend to project style forward. Perhaps surprisingly, we do not find that being widely cited increases the stylistic similarity between a past Justice and members of the current Court[2]. We also examine the potential for party affiliation to have played a role in stylistic evo-

[2] For our measure of influence, we use the tables produced in [86]

lution. While initial analysis reveals some differences between Democratic-appointed and Republican-appointed Justices, these differences appear to themselves be the result of the temporal trend (alongside the changing partisan balance on the Court over time) rather than the cause.

We then examine whether substantive features of opinions affects their style. We first find that there are robust stylistic difference between majority opinions and dissents, even when comparing the writings of the same Justice. The growth of dissents, and dissent-like writing styles, may account for some of the drift in writing style on the Court over time. Second, we examine whether there are differences in judicial writing style across subject matter, such that the changing composition of the Court's docket could account for changes in writing style over time. This analysis finds that, while the similarity between broad topic areas is less than would be expected by chance, we cannot exclude the possibility that temporal changes in writing style are the cause, rather than the consequence, of this effect.

We finally conduct an in-depth investigation of the influence that the modern institution of the judicial clerkship has had on writing style on the Court. For each Justice, we define a measure of consistency as the similarity between a Justice's writings in one year and in all other years (see subsection 2.7.1 for details). We then test the hypothesis that the number of law clerks that a Justice employs is negatively associated with intra-Justice stylistic consistency. Overall, we find evocative evidence that the substantive role that clerks now play on the Court has led to decreasing inter-year intra-Justice stylistic consistency, while leading to increasing intra-year institutional stylistic consistency on the Court.

For more background on the history of computational analysis of legal text and broader legal context of this work, see our paper [25].

Section 2.2

Data

Computational analysis of legal texts is hampered by difficulties accessing the relevant data [4]. While judicial opinions are not protected by copyright, the commercial databases that provide ready digital access to these opinions are protected by terms of use agreements. Limits on machine reading may be necessary to protect the proprietary content that has been produced by these publishers, but they can also inhibit academic research and access to the non-copyrighted government documentary information included within these resources.

Public.Resource.Org, a private not-for-profit corporation has created a digital version of the Supreme Court and federal appellate court corpus, based on the information within the Westlaw database not protected by copyright, and published that information online at "bulk.resource.org." The bulk resource data has been used in prior n-gram studies of text usage in the federal courts and Supreme Courts [80], and provides the public with access to a digital version of the nation's judicial opinions. However, the bulk resource data has some important limitations, including a lack of readily identifiable author and date information.

Because our analysis was limited to the Supreme Court, which has a relatively manageable universe of approximately 25,000 decisions, we were able to augment the Public.Resource.Org data to generate a new dataset. Human researchers conducted a series of "by year" searches on a commercial database to download digitized versions of all Supreme Court cases. All proprietary information was stripped out. Next, a series of iterative human and Python-based analyses were carried out to separate majority, dissenting, and concurring opinions, and assign an authoring Justice and year to each opinion. Per curium decisions were removed from the dataset, as were opinions with a file size smaller than one kilobyte. Data concerning the number of clerks employed in chambers was provided by the Supreme Court Library.

The resulting data covers all opinions for the years 1792 to 2008. Our data includes 25,407 decisions. There are roughly 8,000 dissents and 4,600 concurrences. We have data for 110 Justices: Justices Sotomayor and Kagan were appointed after the end of our study period. We have partial data for Justices who began their terms prior to 2008 but either retired after our study period or remain on the Court [3].

Section 2.3

Preliminary Analyses

Before introducing the primary stylistic metric that is used for the bulk of our analysis, we report the results from three preliminary analyses.

2.3.1. Productivity

Our first analysis examines the "productivity" of each Justice, as measured by the total number of words authored by that Justice in all of their opinions. Figure 2.1 presents the number of words produced by each Justice, with each Justice located on the horizontal axis according to his or her median year on the Court. The results of an ordinary least squares (OLS) analysis comparing Justices' production and their median year of service, found highly significant results[4]. More recent Justices tend to produce more total characters that Justices that served in earlier periods. The analysis of productivity excludes Justices that are currently sitting and the two non-sitting Justices who left after the end of the study period (Justices Souter and Stevens), leaving a total of 101 observations.

[3] These Justices are Samuel Alito, Stephen Breyer, Ruth Bader Ginsburg, Anthony Kennedy, John Roberts, Antonin Scalia, David Souter, John Paul Stevens, and Clarence Thomas.

[4] The coefficient is 6,831, the R-squared value is 0.32, and the p-value is less than 0.01%.

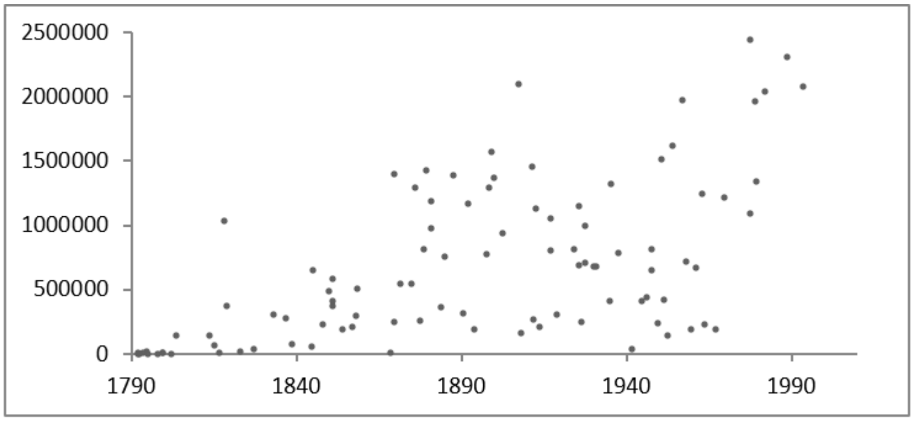


Figure 2.1: Productivity over time (excluding sitting Justices).

There are many factors that could account for the growth in productivity over time, including longer average opinions, more opinions produced per year, and longer length of service. Black and Spriggs provide a detailed treatment of time trends associated with opinion length [15]. They find that while the number of decisions has declined since peaking around the turn of the century, concurrences and dissents have become much more prevalent. They also find that average opinion length tends to go through cyclic patterns. The cyclical pattern identified by Black and Spriggs was growth from 1790 with trend reversals in 1830, 1870, 1900, and 1940, and a final period of growth thereafter.

Figure 2.2 presents an analysis of average opinion length by Justice, ordered by their median year on the Court[5]. There is a general positive time trend in average length[6]. The cyclical pattern identified by Black and Spriggs is roughly present (presented in Figure 2.2 as a four-year moving average) around a general trend of growth. It should be noted that while the time trend noted by Black and Spriggs

[5] This analysis examines majority opinions only, excluding three Justices who authored only dissents of concurrence (Blair, Iredell and Thomas Johnson). Justices with partial data are included, for a total of 107 observations.

[6] The p-value for a simple linear time trend is less than 0.01%; the R-square is 0.49. We examined variability in opinion length as well, finding no statistically significant time trend in the standard deviation of a Justice's opinion length.

was for opinion length by year (in words), the analysis presented here is the average opinion length by Justice (in characters), with the median year of service of the Justice as the explanatory variable.

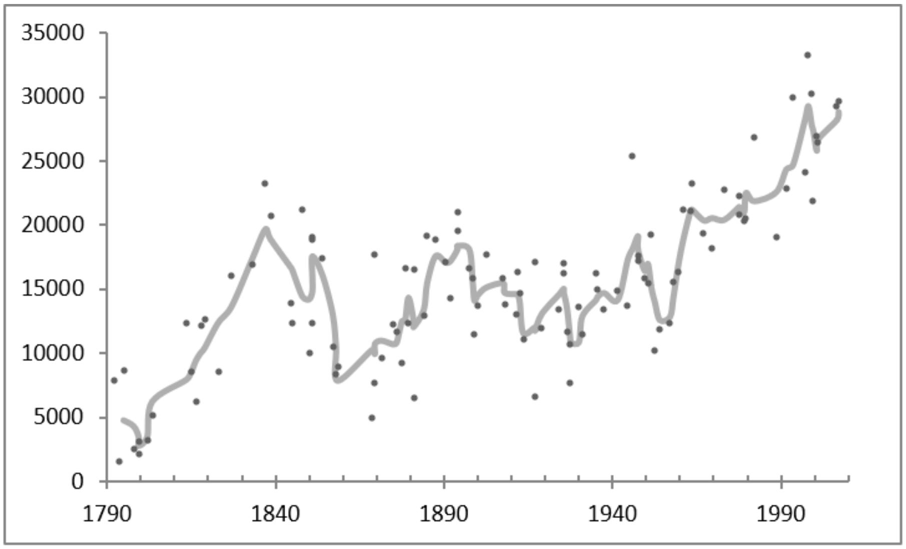


Figure 2.2: Time trends in opinion length

2.3.2. "Friendliness"

Our next analysis examines the "friendliness" of each Justice, as measured by his or her use of positive and negative words. This analysis is based on a list of words constructed to examine the "sentiment" of written texts [69, 97]. Positive and negative words have been used to evaluate online reviews, among other texts, and analyzing their use has generally been found to be a useful means of engaging in "opinion mining"— computational analysis of large text corpora to "determine[. . .] whether a document or sentence is opinionated, and if so whether it carries a positive or negative opinion [96]." Some examples of negative words are "2-faced," "admonish" and "problematic." Positive words include "adventurous" and "preeminent." Together, there are around 7,000 words characterized as either positive or negative.

For each Justice, a Python script was used to determine the total number of negative words and the total number of positive words, in opinions authored by each Justice. The numbers of negative and positive words were then each expressed as percentages of the total number words authored by a Justice. The percentage of negative words was subtracted from the percentage of positive words to generate what we call a "friendliness score."

This analysis – while based on measures of sentiment that have been used in a variety of other contexts – should be approached with a healthy dose of skepticism. Other research areas in sentiment analysis that we are aware of compare contemporaneous texts, rather than examining trends over time. Comparing texts over a long time horizon may be problematic for a variety of reasons, including that a text that reads as relatively friendly in one time period may read as downright nasty in another (or vice versa).

That said, our analysis finds a clear time trend toward lower friendliness scores. Table 2.1 includes the twenty Justices with the highest and lowest friendliness scores, ordered alphabetically, with their median year of service in parenthesis. For this analysis, we exclude Justices with low total production[7]. The analysis on the same data is shown graphically in Figure 2.3, with the median year of the Justice's term on the horizontal axis and the friendliness score on the vertical axis.

The results are evocative: there is a highly significant negative correlation between time and friendliness scores [8]. There are a variety of potential avenues that future

[7] We exclude the Justices who produced less than 100,000 words based all of their writings (majority, concurring, and dissenting opinions). This leaves ninety-two Justices in our sample. We exclude the low production Justices because small total production makes it difficult to draw useful inferences; some of the Justices authored as little as a few hundred total words, leaving less than a dozen positive or negative words in their entire corpus.

[8] An OLS regression on this data showed an R-squared of 0.61 and a p-value of less than 0.01%. Using data based on all 110 Justices adds some noise to the analysis but the results do not substantially change: the R-squared falls to 0.34 and the p-value remains below 0.01%. We also conduct the analysis on all only majority opinions, dropping the three Justices who only authored dissenting or concurring opinions, and arrive at similar results (R-squared failing to 0.27 and similarly low p-values). Dropping the lower production Justices from the majority only analysis reduces the noise considerably. For majority only opinions, dropping Justices with less than 100,000 words of

Highest	**Lowest**
Henry Baldwin (1837)	Samuel Alito (2007)
Samuel Blatchford (1888)	Stephen Breyer (2001)
David Josiah Brewer (1900)	Robert Jackson (1948)
Samuel Chase (1803)	Anthony Kennedy (1998)
David Davis (1870)	Joseph Rucker Lamar (1914)
Stanley Matthews (1885)	Sandra Day O'Connor (1994)
Smith Thompson (1833)	Antonin Scalia (1997)
Willis Van Devanter (1924)	David Souter (1999)
Morrison Waite (1881)	Clarence Thomas (2000)
James Moore Wayne (1851)	Byron White (1998)

Table 2.1: Top Ten Highest and Lowest Friendliness Scores

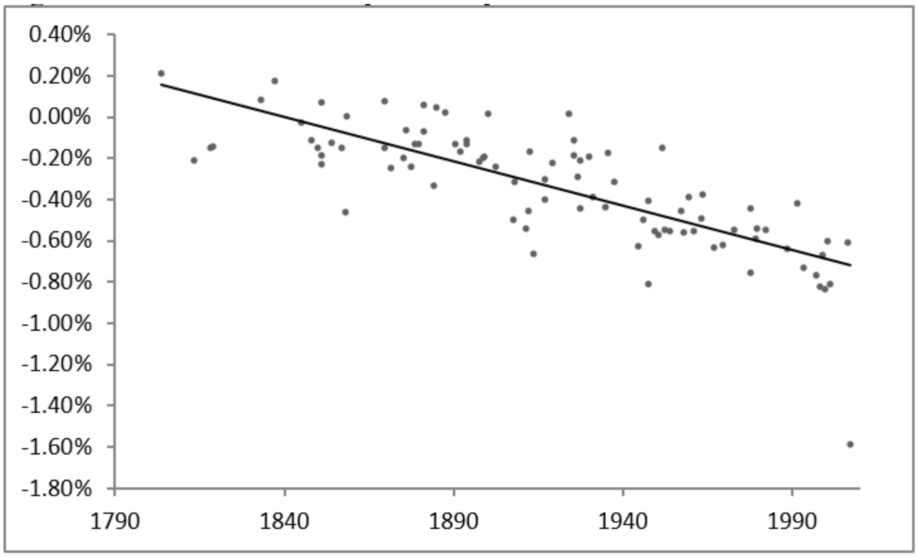


Figure 2.3: Friendliness score by median year

researchers could explore to untangle the causes of this interesting correlation. Some of this effect may be due to an increasing number of dissenting opinions, or the use of less formal language on the part of the Justices, and may be skewed by the use, or non-use, of particular negative or positive words. The changing sentiment on the Court may also reflect broader changes in language usage in political institutions (such as Congress) or within the broader culture. The time trend in friendliness scores, and

production, the R-squared is 0.6.

its causes and potential consequences, may be worth future analysis.

2.3.3. Defensiveness

Our final preliminary analysis reexamines prior research on "defensiveness" that was conducted by Long and Christensen in 2013 [101]. The basic theory underlying Long and Christensen (2013) is that people broadcast weakness through their use of language, and in particular through specific "defensive" forms of speech, including the use of intensifiers, such as the word "clearly," and more complex semantic structure.

To test whether this theory describes behavior on the Supreme Court, Long and Christensen hypothesize that dissents will demonstrate these stylistic characteristics more than majority opinions. For their analysis, Long and Christensen examined 526 Supreme Court opinions in the years 2006 and 2007. They counted the intensifiers in majority and dissenting opinions, as a percentage of total words. For their measure of complexity of semantic structure, they relied on the familiar Flesch-Kincaid reading "grade level" score [82]. Flesch-Kincaid reading grade levels are based on the average number of words per sentence and average number of syllables per word—increasing either increases the grade level.

Long and Christenson found a significant increase in the use of intensifiers in dissenting opinions, but found that the grade level scores were actually higher in majority opinions, although that finding was not statistically significant.

We re-ran the analysis from Long and Christensen on our larger dataset to see how well their findings held up. For every Justice who filed at least one majority opinion and one dissent, an average grade level and intensifier percentage was developed for that Justice's majority opinions and dissenting opinions[9]. A paired t-test was then

[9] There were a total of ninety-nine observations. As before, Justices Thomas Johnson and Iredell were excluded because they authored no majority opinions. There were seven Justices who did not author any dissenting opinions (Chief Justices John Jay, John Rutledge, Oliver Ellsworth, and Salmon Chase and Justices William Cushing, Thomas Todd, and James Byrnes).

run to determine whether there was a statistically significant difference in means for either grade level or intensifier use [10]. Tracking Long and Christensen, we found that majority opinions had somewhat higher grade levels, but that difference was not statistically significant. More interestingly, there was a marked time trend in the sophistication of writing (as measured by grade level), with more recent Justices writing at lower grade level[11]. The time trend analysis is presented in Figure 2.4.

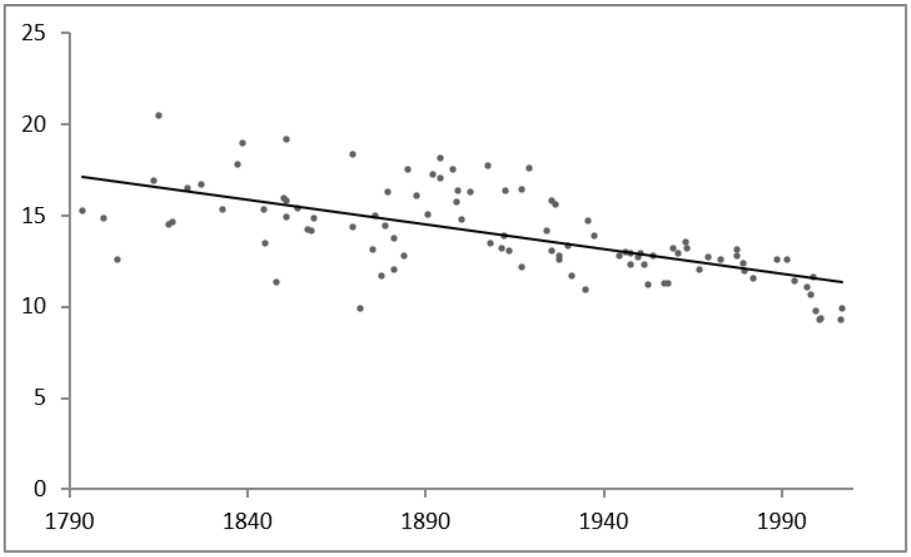


Figure 2.4: Grade level by median year

From this analysis it appears that the Court has generally reduced the complexity of its language (as measured by Flesch-Kincaid Grade Level) over time[12]. This finding runs contrary to the findings of Johnson's examination of grade level trends using a smaller sample: cases written during the 1931–1933 and 2009–2011 terms [76]. Johnson found that writing complexity, and Flesch-Kincaid Grade Level specifically, had increased over time. A glance at Figure 2.4 reveals that there is variability

[10] A paired t-test is a statistical method for examining differences between the means in two samples. See [18]

[11] To develop a single grade level for each Justice's writings, we averaged the grade level for their dissents and majority opinions.

[12] An OLS regression returned an R-squared value of 0.4 and a p-value of less than 0.01%. The coefficient was -0.03.

around the general time trend toward lower grade level scores, making any inference of a general time trend from limited data difficult. Of course, the actual grade level comparison between the two temporal sets made by Johnson remains valid, even if it do not appear to be representative of a longer and more general time trend.

It should be noted that Flesch-Kincaid scores have been criticized as a measure of sophistication and complexity [76]. A general time trend toward lower grade level does not necessarily mean that the Court's reasoning is less sophisticated, or that its writing is of lower quality. Good writing does not necessarily involve long words or long sentences. An interesting question for future research would be whether the trend toward lower complexity in the Court's writings is mirrored in broader social trends, or marks a trend toward more vernacular writing that is more closely in line with non-judicial writing styles.

The paired t-test on intensifier use also confirmed Long and Christensen's findings. There was a markedly higher use of intensifiers in dissents, with means of 0.12% of words for majority opinions and 0.18% for dissents. The t-test revealed a high degree of statistical significance between the means[13]. Unlike friendliness and grade level, there was no obvious time trend in intensifier use—it appears that intensifiers have been used at roughly similar rates across the data.

Section 2.4

Stylistic Fingerprint

Our stylistic analysis moves beyond attempts to measure specific stylistic features of writing, and instead relies on a measure that is meant to serve as a broad proxy for a range of stylistic characteristics: the use of function words.

"Function words" (also called "content-free words" or "CFWs") play a special role

[13] The p-value for a two tailed paired t-test was less than .01% .

in language. They tend to form a "closed" class— i.e., languages do not easily add function words[3]. Content words, on the other hand, are constantly added. For example, the 2014 update to Merriam-Webster's Collegiate Dictionary includes "hashtag," "selfie," and "crowdfunding"—all very much content words[14]. Function words can often be very short, such as "I", "the", "a", "of", while content words are rarely as short. There also appear to be neurologically-related differences between function and content words. Function words are acquired by children later than content words, and specific types of neurological injuries can lead to a loss of use of function words, while content words remain accessible [114]. Various neurological studies have found that function and content words are stored and processed in different brain regions [83]. For purposes of the following analysis, the most important characteristic of function words is that they have been found to provide a source for the development of a useful stylistic "fingerprint" for purposes of author attribution, and we therefore use it as a proxy for writing style more generally [145, 72].

Our study relies on 307 CFWs used in [72] listed in Table 2.2. The individual occurrences of each CFW for each author are aggregated and normalized so that the components sum to one [72]. These normalized vectors are the feature vectors for each author. The feature vectors depend on the level of text aggregation, such as all writings associated with a Justice or the writings of a Justice in a given year, or all of the writings of all of the Justices in a given year. A Python script was used to count each of the CFWs and output the feature vectors into a simple text format.

The degree of difference between two feature vectors can be measured as the *Kullback-Leibler (KL) divergence.* KL divergence is a standard measure for comparing vectors, and has been used in prior studies of the evolution of writing style [72] [15].

[14] Merriam-Webster, A Sample of New Dictionary Words for 2014, http://www.merriam-webster.com/new-words/2014-update.htm.

[15] To avoid undefined division by zero, we smooth by adding .0001 to all of the components in all of the feature vectors, regardless of whether a word was used.

a about above across after afterwards again against all almost alone along already also although always am among amongst amoungst amount an and another any anyhow anyone anything anyway anywhere are around as at back be became because become becomes becoming been before beforehand behind being below beside besides between beyond both bottom but by call can cannot cant con could couldnt cry describe detail do done down due during each eight either eleven else elsewhere empty enough etc even ever every everyone everything everywhere except few fifteen fifty fill find fire first five for former formerly forty found four from front full further get give go had has hasnt have he hence her here hereafter hereby herein hereupon hers herself him himself his how however hundred ie if in inc indeed into is it its itself keep last latter latterly least less ltd made many may me meanwhile might mine more moreover most mostly move much must my myself name namely neither never nevertheless next nine no nobody none noone nor not nothing now nowhere of off often on once one only onto or other others otherwise our ours ourselves out over own part per perhaps please put rather re same see seem seemed seeming seems serious several she should show side since six sixty so some somehow someone something sometime sometimes somewhere still such take ten than that the their them themselves then thence there thereafter thereby therefore therein thereupon these they thin third this those though three through throughout thru thus to together too top toward towards twelve twenty two under until up upon us very via was we well were what whatever when whence whenever where whereafter whereas whereby wherein whereupon wherever whether which while whither who whoever whole whom whose why will with within without would yet you your yours yourself yourselves

Table 2.2: Content-Free Words

Following convention, we use a symmetrized version of KL divergence that is the average of the KL divergence of A with respect to B and the KL divergence of B with respect to A.

$$D_{KL}(A,B) = \frac{1}{2}\sum_{w \in W} A(w)log\Big(\frac{A(w)}{B(w)}\Big) + B(w)log\Big(\frac{B(w)}{A(w)}\Big) \tag{2.1}$$

The KL divergence is then scaled to generate a similarity score. For Justice J_A and J_B with style vectors A and B respectively, the similarity is,

$$S_{J_A,J_B} = e^{\left(\frac{-D_{KL}(A,B)}{\sigma}\right)} \tag{2.2}$$

where σ was chosen to spread the values for a given piece of analysis between 0 and 1.

It is important to reiterate that our stylistic fingerprint is not meant to capture the totality of judicial writing style. It would be strange indeed to claim that the frequency with which Justice Holmes used the word "it" accounts for the claim by Judge Posner—nearly a century later—that the dissenting opinion in Lochner is a "rhetorical masterpiece [134]." Instead, the feature vector is meant to serve as a proxy for the larger set of stylistic characteristics that distinguish one writer from another.

There are other potential measures of style. For example, the LitStats software reports statistics on eight specific stylistic factors: average footnote length, average sentence length, average word length, word length diversity, sentence length diversity, footnote frequency, type-token ratio [166], and the once-word rate. These factors have been used in analysis of juridical writings. Alternatively, scholars have looked to compression software to generate a measure for similarity between writings [37]. All of these methods are plausible, and there is no consensus on a dominant quantitative methodology for quantitative measurement of style. The stylistic measure used in this study has the advantage of simplicity, and is commonly used in both forensic and literary attribution work [72].

Section 2.5 Time Trends in Judicial Style

This section applies the methodology just described to examine how writing style on the Court changes over time. Specifically, we ask whether there is a "style of the time," in the sense that contemporaneous Justices tend to write more similarly than Justices who are temporally remote from one another. As will be clear from the analysis below, the answer to that question is "yes."

To undertake our analysis of the relationship between temporal distance and writing style similarity, we first calculated feature vectors for all justices and created similarity scores for every justice-pair within the study period. Each Justice was also assigned a place in time, based on the mid-point of their term on the Court [16].

Our first analysis is a representation of similarity scores as a style "network" with Justices "linked" to each other based on stylistic similarity. In the terminology of network analysis, the Justices are "nodes" and a local thresholding technique ("LANS") on the stylistic similarity is used to determine when "edges" (or pathways) are placed between the nodes [52]. Each Justice is a node in the network, and an edge was created between that Justice and the 5% of other Justices with the highest similarity scores in their set. If the edge already exists (because it was added when a previous Justice was considered) it will not be added again, but a new edge was not moved into the top 5% to replace it[17].

We then undertake a quantitative estimate of groups within the style network, using the methodology of spectral clustering analysis [165]. Boyd, Hoffman, Obradovic, and Ristovski describe a use of the spectral clustering methodology, which is a technique used to "classify and group" items within a dataset [16]. In essence, spectral clustering "cuts" a network into some defined number of groups (i.e., accomplishes a "clustering"), relative to the condition that similarity between members of the groups should be relatively high and the similarity between members of different groups relatively low. A related (and often thorny) problem in spectral clustering is the determination of the number of clusters as based on the data. We did not address that second problem—which is not necessary to our analysis—and instead set the number of clusters to be identified at seventeen, which is the number of Chief Justices that have served on the Court. That number is admittedly somewhat arbitrary, but it is

[16] For Justices serving at the end of the study period, we used the last year as the end of their term.

[17] Three Justices are excludes from the LANS graph based on a lack of similarity to any other Justice: Chief Justice John Rutledge, Justice Moore, and Justice Thomas Johnson.

sufficient for our purposes, which is generally to examine whether Justices' writing styles appear to cluster together on a temporal basis. The groups generated by the spectral clustering analysis are ordered by the median year of the Justices in the cluster, and the range of median years is presented alongside the group as well.

Figure 2.5 and Table 2.3 present the results from these analyses.

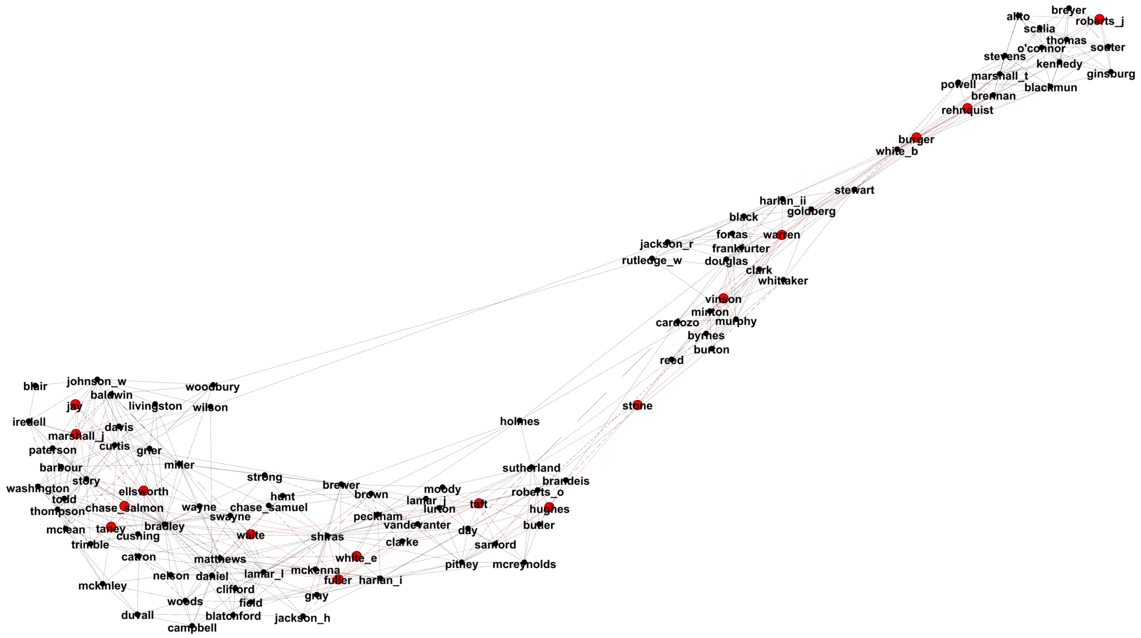


Figure 2.5: LANS graph of stylistic similarity between Justices

ID	Year (Range)					
1	1812	Blair	Jay	Johnson_T	Wilson	Iredell
	(78)	1792	1792	1793	1794.5	1795
		Rutledge_J	Ellsworth	Cushing	Paterson	Moore
		1795	1798	1800	1800	1802
		Todd	Duvall	Woodbury	Chase_Salmon	Clifford
		1816.5	1823	1848	1869	1870
2	1837	Washington	Johnson_W	Trimble	Thompson	Barbour
	(65)	1813.5	1819	1827	1833	1838.5
		Daniel	Story			
		1851	1879			
3	1842	Livingston	Marshall_J	Baldwin	Curtis	Grier
	(55)	1815	1818	1837	1854	1858
		Davis				
		1870				
4	1848	Mclean	Catron			
	(6)	1845	1851			
5	1853	Mckinley	Taney	Campbell	Nelson	
	(14)	1844.5	1850	1857	1858.5	
6	1875	Wayne	Miller	Bradley	Waite	Woods
	(33)	1851	1876	1881	1881	1884
7	1890	Field	Matthews	Lamar_L	Gray	Harlan_I
	(18)	1880	1885	1891	1892	1894
		Jackson_H	Shiras			
		1894	1898			
8	1891	Swayne	Strong	Hunt	Brown	Brewer
	(40)	1872	1875	1878	1899	1900
		Peckham	Mckenna			
		1903	1912			
9	1892	Chase_Samuel	Lamar_J	Holmes	Cardozo	

	(132)	1804	1914	1917	1935	
10	1893	Blatchford	Fuller			
	(12)	1888	1899			
11	1919	White_E	Moody	Lurton	Day	
	(30)	1908	1908	1912	1913	
		Pitney	Mcreynolds	Sutherland	Roberts_O	
		1917	1928	1930	1938	
12	1926	Clarke	Vandevanter	Hughes	Taft	
	(12)	1919	1924	1925.5	1926	
		Sanford	Brandeis	Butler		
		1927	1928	1931		
13	1948	Stone	Byrnes	Murphy	Reed	
	(24)	1936	1942	1945	1947.5	
		Vinson	Burton	Minton	Whittaker	
		1950	1952	1953	1960	
14	1954	Rutledge_W	Jackson_R	Frankfurter	Black	
	(21)	1946	1948	1951	1954	
		Douglas	Fortas			
		1957	1967			
15	1961	Clark	Warren	Harlan_Ii	Goldberg	
	(6)	1958	1961	1963	1964	
16	1980	Stewart	Brennan	Burger	White_B	Marshall_T
	(22)	1970	1973	1978	1978	1979
		Powell	Blackmun	Rehnquist	Stevens	
		1980	1982	1989	1992	
17	2000	O'Connor	Scalia	Kennedy	Souter	Thomas
	(14)	1994	1997	1998	1999	2000
		Ginsburg	Breyer	Roberts_J	Alito	
		2001	2001	2007	2007	

Table 2.3: Spectral clustering analysis

The most striking observation from Figure 2.5 is the degree to which Justices are more stylistically similar to their contemporaries than to temporally distant Justices. This is especially the case in the modern era, with the Justices on the current Court quite isolated stylistically from Justices in earlier years. The thicker portion of the graph, in the lower left-hand quadrant, includes many Justices from the Court's early years, who exhibit somewhat more temporally promiscuous stylistic connections. In general, the spectral clustering analysis displayed in Table 2.3 created groups that were time-based, with temporal ranges of a few decades, and some closer to a single decade [18].

To analyze more closely the relationship between time and stylistic similarity, we characterized every Justice by the median year of his or her term of service on the Court. For sitting Justices, 2008 was used as the end of their tenure. We then calculated the distance in time for every pair of Justices, and related those distances to the similarity score for those Justices. The results are presented in Figure 2.6.

An OLS regression generated an R-squared of 0.18 and a p-value of less than 0.01%, and a coefficient for temporal distance of 0.047. These findings can be interpreted as indicating that, while there are sources of variation in the data other than time, there is also a strong trend of declining similarity in time, with a rate of decay in similarity score of roughly 4-5%. As Justices move farther apart in time, they become increasingly distinct in their writing style.

To examine the influence of time from a somewhat different angle, we next calculated feature vectors for all years and created similarity scores for every year pair within our study period. We then calculated average similarity scores based on temporal distance: one-year distant pairs were averaged into a single similarity score;

[18] The outlier groups are 9 and 1 which have somewhat larger ranges.

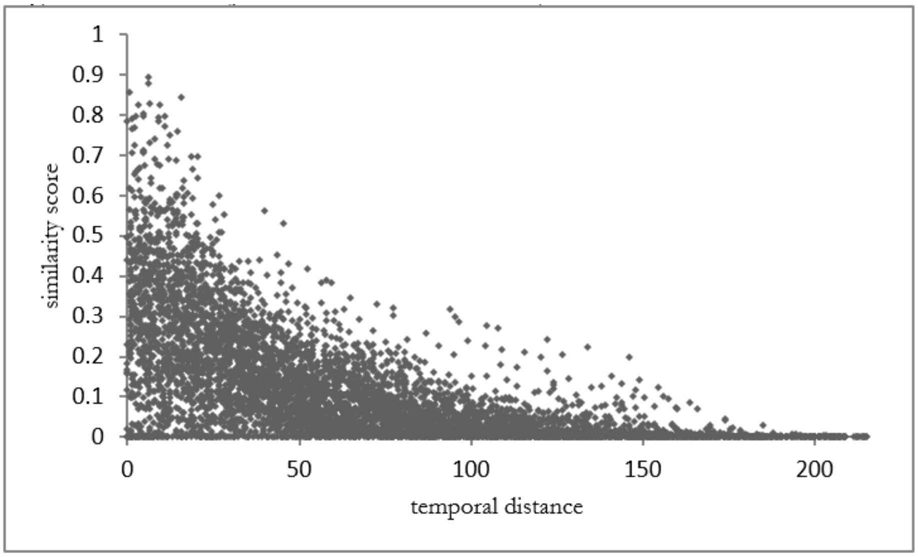


Figure 2.6: Similarity scores between Justices, as a function of time

two-year distant pairs were averaged into a second; and so on. The average similarity score for temporally matched pairs is represented in Figure 2.7.

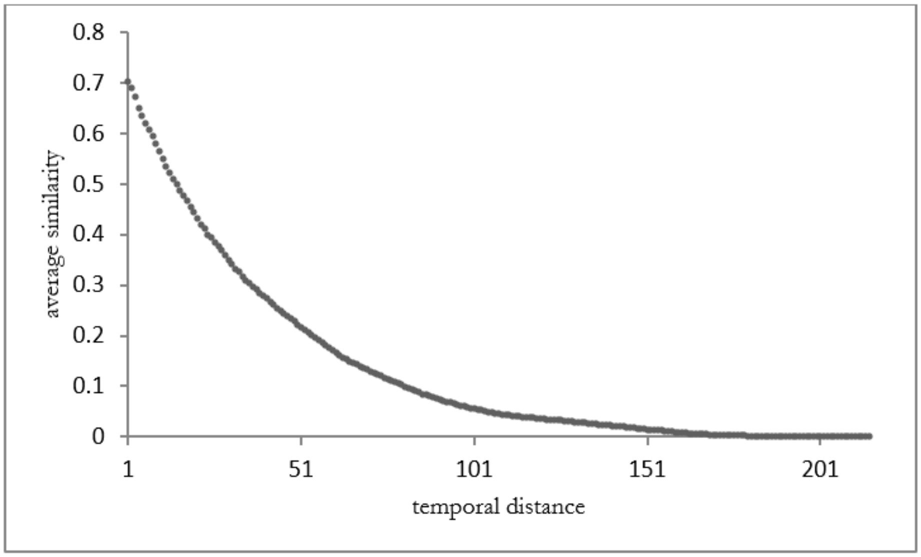


Figure 2.7: Average similarity and temporal distance

Overall, these results indicate a decline in the similarity of year feature vectors as they move farther apart in time, with a rate of decay in similarity of around

4–5% [19]. This analysis again provides strong evidence that style on the Court is not time independent, but instead changes over time. The writings of Justices working together in a given decade are far more stylistically similar to each other than they are to writings of Justices on a temporally remote Court.

Section 2.6

Potential Mechanisms

The foregoing analysis raises an interesting question as to why stylistic similarity within judicial writing declines with temporal distance between Justices. There are a variety of potential mechanisms that could cause writing style to change with time. When examining the Court, it is perhaps most natural to look to external factors, including broader society-wide trends in writing style. Questions surrounding change of writing style on the Court, then, would necessarily implicate a larger set of questions concerning change of writing style outside the Court in various media, including literature, popular culture, and personal communication [72]. Those questions, while no doubt of interest, are outside the scope of this project.

We restrict our analysis to internal factors that could help explain a change in style. In this Part, we examine three potential causal mechanisms. The first is influence by highly respected prior decisions. We do not find convincing evidence that the more frequently cited prior decisions exert any particularly great influence on latter style. We then examine changes in the Court's composition, and do not find evidence that the partisan affiliation of Justices has an effect on style. Finally, we examine the potential influence of substance, including by comparing dissenting and majority opinions, and comparing opinions concerning different subject matters. We

[19] The coefficient for temporal distance is -0.04 in an OLS regression of the natural log; the p-value is less than .01%. The R-squared for this analysis is 0.92, reflecting the reduction in noise due to the averaging procedure.

find some evidence that writing style bears some relationship to subject matter and opinion type.

2.6.1. Prior decisions

The first mechanism that we examine is the possibility of a causal role played by influential past decisions. Just as past decisions generate legal standards and norms of judicial reasoning, they serve as the backdrop against which a Justice's writing style is perceived. While some innovation in writing style may be rewarded, Justices are likely to express some degree of conformity to prevailing conventions. Justices may also consciously model their writing style on prior Justices who they find to be particularly worthy of emulation, or may be subconsciously influenced by the decisions that they read.

To test for the influence of prior Justices, we rely on the current Court as our baseline. To create the baseline, we used the writings of each of the currently sitting Justices in our dataset at the time of our experiment (Scalia, Thomas, Roberts, Thomas, Alito, Ginsburg, Breyer) to generate a single stylistic feature vector. We then exclude from our analysis all of the Justices who served at the same time with any sitting Justice as a way to separate out any cross-influence between Justices and ensure that the causal relationship runs in the anticipated direction. For each remaining Justice, we constructed a feature vector for their writing, and calculated the KL divergence between that feature vector and the current Court baseline.

For each Justice in the analysis, we then constructed a "ghost" vector made up of the texts produced by the Court in each of their years on the bench, minus that Justice's writings [20]. We calculated the KL divergence between the ghost vectors and

[20] For this analysis, we use a smaller list of seventy-five non-content words: first between also where who those part than him will could without whether must after before within should these only them when against same so one would their there has they other all made may if we us he under but been had his were no have are any its upon such at an with from on which this not or as for be it was by is a that in and to of the

the current Court baseline vector. Finally, we took the simple difference between the KL divergences: a difference less than zero indicated that a Justice's writing was more similar to the current Court's style than to the other writings of the Court in the years when that Justice was on bench. We call these "prediction scores" based on the idea that Justices who perform well tend to "predict" the current style of the Court better than Justices who perform poorly (with lower numbers associated with better prediction). There were few Justices with prediction scores of less than zero, because each Justice typically authors only a small fraction of cases in a given year, meaning that random sources of variation are much less likely to substantially influence the ghost vectors than an individual Justice's vector.

We then compared the resulting prediction scores to a measure of "historical value" for each Justice, which was generated by Kosma in 1998 based on citation counts [86] [21]. This variable is meant to capture the possibility that Justices who are widely cited exert greater stylistic pressure on subsequent Justices. We control for the relationship between a Justice's total production, in words, and our prediction scores. There are two potential mechanisms for this variable to affect the prediction scores. First, Justices that produce a great deal of text contribute more to the total body of the law that later Justices read. For that reason, perhaps they exert greater stylistic influence. Higher levels of production also imply less opportunity for random sources of variability in the use of function words to affect a Justice's feature vector. Finally, we control for time, to account for temporal effects that are not captured in the ghost vector–based normalization, and examine the interaction between production and Kosma's historical value [22].

[21] In Kosma's analysis, Chief Justice Hughes is given two different scores, corresponding to the two different stints that he spent on the Court. Because of the lack of correspondence to our single entry for Hughes, we dropped him from both sides of the analysis.

[22] For this analysis, we use the data described above to construct three variables. *Hist* is based on Kosma's historical value scores, normalized through a cube root function; *Prod* is based on total word production, again normalized through a cube root function. *Predict* is the natural log of the prediction scores. We created a fourth variable *Predict1* which is a cube root transformation of the

	Predict	Predict	Predict	Predict1
Hist	-0.073 (6.94)**	-0.015 (0.53)	-0.107 (3.04)**	-0.018 (3.20)**
Prod		-0.013 (2.27)*	-0.027 (3.30)**	-0.003 (2.67)**
Year			0.000 (0.10)	0.000 (0.78)
Hist*Prod			0.001 (3.55)**	0.000 (3.44)**
_cons	-2.965 (17.17)**	-2.871 (16.53)**	-2.284 (0.75)	0.146 (0.31)
Adj. R^2	0.35	0.38	0.47	0.31
N	87	87	87	91

Table 2.4: Historical Influence. '*' indicates p<0.05, '**'p<0.01

Historical value is significant in the first specification as a standalone variable. In the full model, historical value and production are significant, as is the interaction term. Recall that lower prediction scores imply greater similarity. The interaction variable implies that historical value has a less strong influence on prediction scores as production increases [23].

As an additional test, we examine higher and lower production Justices separately. For Justices with production greater than 100,000 words, random sources of variation in function word use will be much less important. Within this group of seventy Justices, neither *Hist* nor *Prod* nor their interaction is significant[24]. Within the

prediction scores: the log transformation better normalizes the data but creates difficulties around the negative prediction scores. For *Predict* we drop the negative observations, which are retained in *Predict1*.

[23] To give a flavor for the effect of the interaction, we re-centered Prod around "high" production Justices (defined as one standard deviation above the mean), "medium" production Justices (defined at the mean) and "low" production Justices (defined as one standard deviation below the mean). In three specifications of the full model, we replace Prod with Prod_low, Prod_med and Prod_high. Those regressions return coefficients of 0.058, -0.022, and 0.014 for Hist, respectively. The improvement in prediction, then, is roughly twice as strong for the low productivity Justices as the average Justice. High productivity Justices do not appear to benefit from greater historical value.

[24] These variables are not close to traditional significant thresholds in this specification, with p-values ranging from 0.4 to 0.9. Year is also not significant in this specification. The adjusted R-squared is negligible at less than 0.03. The same findings hold with the cube root transformation of the prediction scores, which adds four more observations into this group.

seventeen lower production justices *Prod* is significant and *Hist* and the interaction variable approach traditional significance thresholds, with signs that match those in Table 2.4 [25].

What to make of this analysis? Justices at the lowest level of productivity have poor prediction scores, and their prediction scores improve as they become more productive. Our analysis cannot determine whether that effect is from a reduction in statistical noise or a greater likelihood that a future Justice read and internalized their writing style. Among the lower productivity Justices, authoring more highly cited opinions may improve their prediction scores. For Justices at higher levels of productivity, additional citation does not appear to contribute to greater stylistic similarity with future Justices [26].

2.6.2. Partisan affiliation

We next examine the possibility that writing style is associated with some other set of cognitive, ideological, value-based, or perceptive characteristics, and that the change in writing style over time on the Court reflects a broader shift in Weltanschauung [60, 158]. This type of relationship is hard to test, for obvious reasons, but we conduct a very general analysis by examining whether there is any systematic stylistic difference between Justices appointed by Presidents of different parties.

We only test differences between the contemporary Democratic and Republican parties, and so restrict our analysis to the latter half of the twentieth century [27]. Looking over the entire study period, there is somewhat less similarity between

[25] Hist has a p-value of 0.07 and Hist*Prod has a p-value of 0.1. Prod has a $p<0.01$. Year is not significant. The adjusted R-squared in this specification is 0.67.

[26] [72] find that Nobel Laureates similarly did not have out-sized long range similarity with the style of latter writers

[27] There is some controversy over the meaning of "realignment" elections and their relationship to party systems[108], but the dawn of the FDR coalition provides a reasonable starting place for when the contemporary meaning of "Democrat" and "Republican" take shape. There is enough data to generate inter- and intra-party similarities for post-1932 appointees starting in the mid-1950s.

Democratic-appointed Justices and the Republican-appointed Justices than within the party groupings. However, there is an obvious temporal problem, because the relative representation of the two parties on the Court has shifted markedly over time.

To account for this feature in the data, we compare the inter- and intra-party difference for each year, starting in 1955. We generate a feature vector for the texts authored by the Justices appointed by Republican and Democratic Presidents and calculate a similarity score between them. This analysis is done for each year starting in 1955. For each party, we then subdivide the opinions, randomly, into two test groups and generate feature vectors for the test groups. Finally, we calculate similarity scores for the feature vectors for the same-party test groups, and then average the two scores (Democratic and Republican) to generate a measure of intra-party distance. The hypothesis is that the inter-party similarity scores will be lower than the intra-party scores.

On average, the similarity scores are a shade higher for the inter-party group (contrary to the hypothesis), and that difference is not statistically significant. We also conducted a very simple significance test by computing similarity scores for "parties" generated by random assignment of opinions for each year [88]. There were statistically significant differences between the similarity scores in the actual data and the simulated similarity scores [28]. The actual similarity scores were somewhat lower—both inter- and intra-party—compared to randomly generated groupings. There is no clear interpretation for this feature of the data, but it does not provide evidence that partisan affiliation is associated with stylistic difference.

To account for the possibility that writing style has become more polarized over the course of our dataset (as the parties have polarized) we examined whether there

[28] The p-value for a paired t-test on the mean of the similarity scores for the two groups was 1%.

was any time trend toward increasing dissimilarity between inter- and intra-party similarity scores. We find a very mild time trend toward greater dissimilarity, but time accounted for very little of the variation and the trend was not statistically significant [29]. Overall, for this relatively small portion of the dataset (53 years), we do not find evidence that there are greater differences between parties than within the parties.

2.6.3. Substantive factors

We conduct two investigations into the role of substantive factors in influencing writing style. First, we examine the degree of difference between opinion types (dissenting and majority opinions), compared to the degree of divergence within opinion type. For this analysis, we eliminated pre-1950 texts, when dissenting opinions were relatively rare. We then randomly separated majority opinions into two groups and dissenting opinions into two groups, and calculated the KL divergence between the feature vectors constructed between those two groups. As expected, given the large number of texts in each group, the KL divergence was quite small[30]. We then examine the differences between majority and dissenting opinions and found that the KL divergence for these groupings was two orders of magnitude higher[31]. This is a statistically significant result[32].

To account for the possibility that the growing number of dissents combined with general stylistic trends caused these differences, we constructed corpora of dissents and majority opinions for each Justice, and conducted the same within-group and between-group analyses. Because their groups were smaller, there was greater op-

[29] The R-squared value was 0.02, and the p-value was 38%.

[30] The KL divergence was 0.00026 for dissenting opinions, and 0.00011 for majority opinions.

[31] Majority1-Dissent1, 0.011; Majority1-Dissent2, 0.010; Majority2-Dissent1, 0.011; Majority2-Dissent2, 0.010.

[32] An analysis on simulated groupings showed that the likelihood of randomly generating such a difference between similarly sized groups was well below 0.01%.

portunity for random variation to affect the feature vectors, and the KL divergences were greater in general[33]. We also examined the KL divergence between majority and dissenting opinion, finding that there was not the same order of magnitude difference, but that there were statistically significant differences between the mean KL divergences[34].

The bottom line of this analysis is that there does appear to be a difference in writing style between dissents and majority opinions, even for the same Justice. One potential source of temporal variation in writing style, then, may be the growing prevalence of dissents on the Court[35]. Given that this particular form of judicial writing appears to be stylistically distinct, its growth in popularity may account for some of the temporal drift in writing style on the Court.

Our final analysis along these lines examines the potential for stylistic differences based on the subject matter of the opinion. We use the categorizations in the Spaeth Database[36] to classify opinions by topic area[37]. Our analysis is therefore limited to the period of time covered in the Spaeth Database (cases after 1946). Dividing the cases up into the thirteen broad "topic areas" assigned in the Spaeth Database, we create feature vectors for each of the areas, and construct similarity scores between them. These results are reported in Table 2.5.

[33] The average KL divergence between majority opinions was 0.012; the average KL divergence between dissent opinions was 0.05.

[34] The average KL divergence between opinion types, for all groups (majority1-dissent1; majoirty2-dissent1; majority1-dissent2; majority2-dissent2) for all Justices was .04. We conducted a t-test on the difference in means between the KL divergence within majority opinions and between opinion types, and the difference in means between the KL divergence within dissenting opinions and between opinions types. Both were significant ($p< .0.01\%$).

[35] Our tests of concurrences found that they had half the KL divergence from majority decisions as dissents; also, it is possible the majority writings have begun to take on the tone of earlier dissents, such as being more argumentative.

[36] http://Supremecourtdatabase.org

[37] The SCBD topic areas are: 1. Criminal Procedure; 2. Civil Rights; 3. First Amendment; 4. Due Process; 5. Privacy; 6. Attorneys; 7. Unions; 8. Economic Activity; 9. Judicial Power; 10. Federalism; 11. Interstate Relations; 12. Federal Taxation; 13. Miscellaneous.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	1	0.52**	0.40**	0.59	0.26*	0.47	0.26**	0.29**	0.39**	0.18**	0.14*	0.22**	0.17
2	0.52**	1	0.70**	0.71	0.67	0.69	0.63*	0.73**	0.80**	0.61**	0.31*	0.52**	0.53
3	0.40**	0.70**	1	0.53**	0.60**	0.52**	0.55**	0.63*	0.63**	0.53**	0.21**	0.35**	0.42
4	0.59	0.71	0.53**	1	0.44**	0.60**	0.42**	0.51	0.62**	0.38**	0.21**	0.41**	0.32
5	0.26*	0.67	0.60**	0.44**	1	0.48**	0.49**	0.56	0.58	0.58**	0.21**	0.31**	0.46
6	0.47	0.69	0.52**	0.60**	0.48**	1	0.43**	0.52	0.62	0.39**	0.19**	0.41**	0.33
7	0.26**	0.63*	0.55**	0.42**	0.49**	0.43**	1	0.8	0.74	0.68**	0.29**	0.56**	0.49
8	0.29**	0.73**	0.63*	0.51	0.56	0.52	0.8	1	0.86*	0.81	0.38	0.66	0.55
9	0.39**	0.80**	0.63**	0.62**	0.58	0.62	0.74	0.86*	1	0.76*	0.38	0.61*	0.58
10	0.18**	0.61**	0.53**	0.38**	0.58**	0.39**	0.68**	0.81	0.76*	1	0.39**	0.49**	0.59
11	0.14*	0.31*	0.21**	0.21**	0.21**	0.19**	0.29**	0.38	0.38	0.39**	1	0.29**	0.31
12	0.22**	0.52**	0.35**	0.41**	0.31**	0.41**	0.56**	0.66	0.61*	0.49**	0.29**	1	0.36
13	0.17	0.53	0.42	0.32	0.46	0.33	0.49	0.55	0.58	0.59	0.31	0.36	1

Table 2.5: Similarity scores between SCDB topic areas. ‘*’ indicates $p<0.01$, ‘**’$p<0.001$

We again generated simulated similarity scores for similarly sized but randomly assigned groups. Table 2.5 notes the similarity scores that are statistically significantly lower than would be expected for groups of texts of equivalent sizes.

These results are consistent with the possibility that some of the temporal drift in writing style on the Court is due to the changing composition of cases that the Court hears over time [128]. At the same time, our analysis does not exclude the possibility that differences between the subject matter is the result, rather than the cause, of temporal shifts in writing style on the Court. In addition, subject matter differences may be a consequence of particular Justices consistently writing opinions in particular subject areas. This source of difference in writing styles between subject areas would not necessarily contribute to any temporal effect. Separating out these possibilities will have to await additional analysis.

Section 2.7

Clerk Influence

As judicial clerks have become an enduring feature of the operation of the federal courts, the role of these recent law graduates has been the subject of both scholarly and public debate [168, 132, 126, 87]. An important empirical predicate to this debate is the belief that clerks play a substantial role in authoring opinions. At least for the Supreme Court, there is a long history of anecdotal evidence supporting the claim that law clerks exert some influence over judicial decision-making [141, 144, 142]. There is also a nascent literature that uses quantitative techniques to address the question of clerk influence over both substance and style. This section investigates whether the stylistic measures discussed above can provide insights into whether clerks have had a measurable effect on writing style on the Court.

2.7.1. Comparing inter-year variability

The question of clerk influence over opinion drafting has been the subject of several attempts at computational content analysis [166, 37, 15, 145]. Our analysis expands this prior research.

Wahlbeck, Spriggs, and Sigelman [166] as well as Rosenthal and Yoon [145] test hypotheses about individual Justice pairs based on anecdotal evidence concerning reliance on clerks. While both studies reject the null hypotheses about a single Justice-Justice pair with a high degree of confidence, it is difficult to extrapolate their findings to a more general conclusion about clerk influence. Choi and Gulati's [37] computationally intensive test finds no greater stylistic consistency for reputed likely author judges; their more straightforward measures are loosely commensurate with their anecdotal hypothesis, but the evidence is not overwhelming. Black and Spriggs [15] find no relationship between clerks and opinion length (the only variable that they studied).

We focus on variation in writing style and focus specifically on inter-year stylistic variability. Our model of clerkship influence is different than that used by Rosenthal and Yoon and Wahlbeck, Spriggs, and Sigelman. For those authors, variation is hypothesized to be a consequence of different clerks drafting different opinions in a given year. While this may well be a source of variation, it is extremely difficult to identify the roles of individual clerks, and there are reasons to believe that multiple clerks may be involved at some point in the drafting and editing process [166].

Our model differs in its focus on clerk turnover as the source of variability, rather than simply the existence of clerks within chambers. One of the peculiar features of the contemporary clerkship is that it is so short, typically lasting a mere year. We exploit this fact in our inter-year measure of variability. In addition, we construct a new measure of the total consistency of the Court. This measure examines the writing style consistency of the Court as an institution, rather than individual Justices. We then compare both of our new measures of consistency to the time periods used by Black and Spriggs (the Peppers groups [132]), to determine if the changing nature of the clerkship institution has affected either intra-year consistency of the Court or inter-year consistency of individual Justices.

The first measure of variability that we introduce is centroid distance. This measure is based on the writings of the entire Court in each year. The distance between the feature vector for each text in a given year and the remainder of the Court's writings in that year are computed, and those distances are summed for each year [38]. This provides a measure of how tightly clustered the Court's style is in a given year: the greater the centroid distance, the bigger the stylistic "spread."

An OLS regression on this data examining the relationship between year and centroid distance found that there is a statistically significant relationship over the entire period[39]. Over time, the intra-year consistency on the Court has measurably increased.

To examine whether the overall trend toward greater consistency differed as the institution of the modern clerk developed, we conducted a structural break test on the data. A structural break is a concept from econometrics that is primarily used in time series analysis of macroeconomic data [62]. The point of a structural break analysis is to determine whether there has been an underlying shift in the data generating mechanisms, such that the distribution of data from the period after the "break" is systematically different that the distribution prior to the break. For example, the U.S. economy generates data on productivity, employment, and other economic variables. In our analysis, the data generating mechanisms is the U.S. Supreme Court. The

[38] For the centroid distance estimate, we use cosine similarity, which, like KL divergence, is a representation of distance in multi-dimensional vector space. To avoid confusion, we take [1 - cosine similarity] as the measure of "distance" so that larger distances are associated with greater difference.

[39] The p-value is less than 0.01% and the R-squared value is 0.5.

First Period	Second Period	F-value	P-value
(1791-1885)	(1886-2008)	5.7	<0.01
(1791-1885)	(1886-1919)	6.7	<0.01
(1886-1919)	(1920-1952)	5.5	<0.01
(1920-1952)	(1953-2008)	3.13	<0.05
(1886-1919)	(1920-2008)	16.1	<0.01

Table 2.6: Chow test on centroid distance

structural break analysis is meant to examine whether the variable of interest—intra-year consistency—exhibited a different relationship to time during the periods when clerks played very different roles in chambers.

We first ran a Chow structural break test, which is a standard tool to determine whether there are changes in the relationships between time and another variable over different time periods [38]. For the potential break points, we used the Peppers groups [132]. The results of the Chow tests are presented in Table 2.6.

The Chow test rejects the null hypothesis that there are no structural breaks in the centroid distance data at the Peppers groups dates. We conduct two additional tests. The first examines whether there is some structural break in the data and estimates the break date[40]. For this test, we do not specify a hypothesized date. The analysis rejected the null hypothesis of no structural break[41]. The estimated break date that was returned was 1926, very close to the dates that clerks took on a greater substantive role, as indicated by the Peppers group transition from "stenography" to "assistants." We also conducted Wald and likelihood-ratio based structural break tests [133] for the three hypothesized break dates of 1885, 1919, and 1952. Both tests confirmed breaks at those dates with a p-value of less than 0.01.

To attempt to better estimate the effects of clerks specifically, we develop a new measure of writing consistency, focused exclusively on inter-year variability in writing

[40] For this analysis we used the Supremum Wald test. See [133]

[41] The p-value was less than 0.01

style. For purposes of our analysis, a chamber in a given year can be thought of as a "team" made up of a Justice and several clerks. A team co-produces the opinions in a given year. When clerks turn over, it changes the composition of the team. In chambers with a larger number of clerks that turn over more frequently, there will be a higher percentage of team turnover from year to year. Although some inter-year stylistic variability can be expected even with a single author, clerk turnover is hypothesized to decrease inter-year consistency.

The dependent variable in our analysis is an inter-year consistency score. To construct the consistency score, we rely on the feature vectors based on the texts a Justice authored in each year of his or her tenure. So, for a Justice with a term from 1950 to 1959 (inclusive) there would be ten feature vectors for that Justice. For each year, we then calculate the KL divergence between that year's vector and a feature vector based on the remainder of the Justice's writings. The consistency score for a Justice active between years i and j is

$$\sum_{y=i}^{j} e^{\left(\frac{-D_{KL}(A_y,B_y)}{0.2}\right)} \tag{2.3}$$

where A_y is the single year vector for year y and By is the vector based on the Justice's opinions except for those written in year y. This is the sum of similarity scores, as defined in equation 2.2, with σ set to 0.2. Figure 2.8 presents each Justice's consistency scores ordered by median year of services.

We examine the relationship between consistency score and the number of clerks that served in a Justice's chambers over the course of his or her tenure. The time trend will be controlled for through a polynomial. We will also control for each Justice's total production, under the theory that Justices who produce more may be more consistent, and there will be less statistical noise between years. We will also

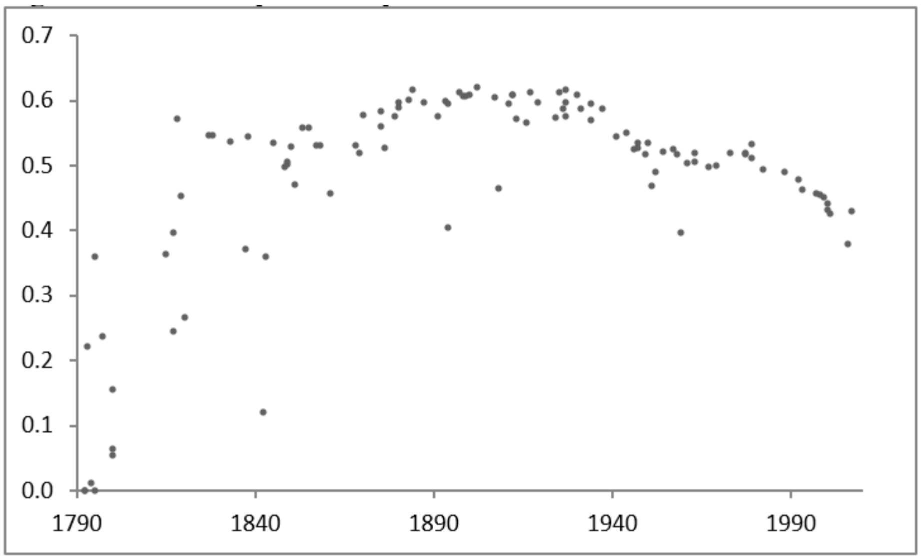


Figure 2.8: Consistency Scores by Median Year

examine the interaction between clerks and time[42].

	Consist	Consist	Consist
Clerks	-0.061 (10.32)**	-0.066 (12.22)**	-2.908 (3.39)**
Prod		0.001 (4.19)**	0.001 (3.90)**
Year			0.113 (3.33)**
Year^2			-0.000 (3.35)**
Clerks*Year			0.001 (3.34)**
_cons	0.223 (30.26)**	0.171 (12.31)**	-106.962 (3.31)**
Adj. R^2	0.63	0.70	0.75
N	65	65	65

Table 2.7: Clerk Influence on Consistency

[42] For this analysis, we examine the period after 1885, with the introduction of clerks as "stenographers" under the Peppers grouping. We normalized the consistency scores using a cube function to construct Consist. The variable Clerks is the total number of clerks that served in a Justice's chamber, divided by that Justice's tenure on the Court, and normalized through a square root function. Prod is as described above.

First Period	Second Period	F-value	P-value
(1791-1885)	(1886-1919)	8.5946	<.01
(1886-1919)	(1920-1952)	5.5973	<.01
(1920-1952)	(1953-2008)	4.2538	<.01
(1886-1919)	(1920-2008)	5.8302	<.01

Table 2.8: Chow test on consistency score

Table 2.7 reports the results of an OLS regression with median year and clerks per year as explanatory variables of consistency scores. In the first model, additional clerks are associated with a reduction in inter-year consistency. This relationship holds when a Justice's production is taken into account; that variable is significant and associated with increased consistency. Finally, median year of service is significant, with a linear trend toward greater consistency and a squared trend toward lower consistency. The interaction between clerks and year indicates that clerks have had a less strong influence over time, perhaps indicating the declining marginal influence of an additional clerk as the Court has institutionalized a practice of each Justice having between four and five clerks.

Because the institution of the modern clerkship was introduced gradually over time, it is difficult to fully disaggregate the effects of clerks from other time dependent variables. We conduct one further structural break analysis based on the four Peppers groups, reported in Table 2.8[43].

The first three tests identify the likelihood that the coefficients in the Peppers groups are the same, rejecting the null hypothesis in all cases. We also compare the period of clerks as stenographers to the latter two periods, and reject the hypothesis that the time trends have the same coefficient.

[43] Unlike in the case of centroid distance, we do not have the type of data necessary to carry out the additional structural break analyses discussed above. For centroid distance, we had a single measure for every year except (constructing an average for two missing years). In the case of consistency scores, which is ordered according to the Justices' median year of service, there are many years missing, and a number of years with multiple entries.

It is worth remembering the difficulty of fully distinguishing the effects of unobserved time-related variables from the effects of clerks. Nevertheless, there can be little doubt that over the course of the twentieth century, the intra-year stylistic consistency of the Court as an institution has increased, while the inter-year consistency of writing style for individual Supreme Court Justices has declined. Nor can there be any doubt that over the same period of time, law clerks have become ever more integrated into the substantive work of the Court. Because the institution of the modern law clerk in the U.S. Supreme Court evolved gradually over time, it is hard to know the degree to which clerks have contributed to the decline of writing style, independent from some other set of time-related variables. But the information presented in this study is highly evocative, indicating that clerks have likely played a role in both increasing the consistency of the institutional voice of the Court and reducing the consistence of each Justice's individual voice.

Section 2.8
Conclusion

Over the past several decades, there has been an explosion in quantitative analysis in legal scholarship. The lion's share of that research has focused on the statistical analysis of hand-coded cases, typically oriented toward legal content. This research has spurred a number of interesting debates about law, politics, and various influences (and non-influences) on judicial behavior. As computational text analysis has become more sophisticated and more accessible, scholars have begun to apply these tools to legal questions (see e.g., [99]). The reduction of the costs of engaging in new forms of content analysis has allowed for new types of questions to be asked.

This chapter aligns with an important thread of this larger overarching effort and examines the stylistic features of judicial writings in quantitative terms. We offer several important innovations: (1) We construct a unique dataset of all Supreme Court cases in which dissents and concurrences are separated from main opinions; (2) these texts are coded with identifying information for year of publication and authoring Justice; (3) this substantial dataset, along with advances in computational power, allows us to conduct re-analysis of prior research to examine its validity in light of the new data. Finally, (4) we newly apply the "stylistic fingerprint" of frequency of function words (a known general proxy for writing style) to investigate trends as represented in the full decision corpus.

With this proxy variable, we test several hypotheses. The first hypothesis is that there is a style of the time in the Court, such that contemporaneous Justices write more similarly to their peers than to temporally remote Justices. Our analysis finds extremely strong support for this hypothesis. We also examine some potential causes, finding little support for the claim that highly cited Justices exert greater-than-average stylistic influence, but finding some non-conclusive support for the possibility that changes in legal content, and even judicial values, ideology, or perspective, may account for some of the change. Finally, we examine the influence of judicial clerks on writing style. Specifically, we test two hypotheses concerning the modern institution of the rotating judicial clerk. First is the claim that this phenomenon has led to greater intra-year institutional writing consistency on the Court. Second is the claim that clerks have led to less inter-year individual writing consistency for the Justices. We find reasonably strong support for both propositions, although it is impossible to exclude the effect of unobserved time-dependent variables on either.

Overall, we aniticipate that the analysis in this chapter opens the door to new avenues of research of legal texts. In particular, we believe that there are a number of important and interesting research questions to be asked concerning the interaction of writing style in the Court with broader social and political trends. Our dataset can be linked to other textual corpora, including appellate opinions and state opinions, as well as newspapers, published books, Twitter feeds, et cetera. This textual data can also be linked to more traditional social science sources, such as published economic (GDP, unemployment) or political (voting, electoral outcomes, campaign donations) data. Deploying the analytical tools described above, it may be possible to examine the interaction of writing style on the Court with courts more generally, with other forms of writing prevalent in the culture, and with broader social and political trends revealed in social science data. The possibilities of such analysis are exciting: human researchers can now find textual patterns that emerge at a macro-level, perceptible only recently with the digitization of vast textual corpora, the broad available of massive computing power, and the continually evolving application of advanced concepts in mathematics and computer science to these "big" datasets. As these textual patterns become ever more perceptible, they offer the hope of new understandings in the use and evolution of language, from the staid chambers of the U.S. Supreme Court to the unruly sprawl of the blogosphere.

Chapter 3

Generating Stylized Text

Machine generation of “human-level” text is a fascinating challenge. Success has implications ranging from story generation [48], the automated generation of real (e.g., weather reports [139]) or fake news articles [180], and even poetry [55]. In this chapter we show a new application to marketing – and specifically, the machine generation of marketing materials for wine that uses an approach that is clearly generalizable to other contexts.[1]

Section 3.1

Background

Online product reviews are one of the most ubiquitous and helpful sources of information available to consumers today for making purchase decisions. Consumers rely on both the quantitative aspects of reviews such as volume [35] as well as the text content of reviews [30] to learn about product quality and fit. Firms are likewise concerned about the effect of reviews on sales [34] and actively engage in review management

[1] This work formed the basis for “Complementing Human Effort in Online Reviews: A Machine Learning Approach to Automatic Content Generation", by Keith Carlson, Praveen K. Kopalle, Dan Rockmore, Allen Riddell, and Prasad Vana, for the (competitive) 2020 Conference on Artificial Intelligence. A full article is currently under journal review.

[33]. Much of the literature in this area has focused on reviews and opinions written by regular consumers [79, 117], and managers of businesses [137].

Our focus in this work is on the reviews and opinions written by experts (such as professional wine tasters) who professionally critique products. Research has shown that consumers rely particularly on the opinions of such experts for purchase decisions in the context of a variety of experience goods, such as books[14], movies [143, 13] and wines [119, 53] as well as for new technology products about which there is little information [103]. While these domain-specific experts possess the skills required to evaluate a product objectively, they may lack the ability to articulate and communicate their opinion through an engaging and convincing review given the complexities in describing these products [119]. Additionally, these experts increasingly face an unmanageably high volume of products needing written reviews[2].

To alleviate these challenges of expert reviewers, we turn our attention in this research to the writing of product reviews by machines and ask if machine-written product reviews can be as readable, engaging and informative to humans as human-generated (i.e., human-written) reviews. Machines capable of taking the product attributes as inputs and generating a human readable review as an output could act as "writing assistants" and offer the first draft of the review to help the expert reviewer (or tongue-tied evaluator) in writing their review.

Automated systems which can generate text have long been the subject of research [106, 41, 112]. In some cases the goal of the system is to create novel text which has certain characteristics, with minimal or no input. Some examples of this are the automatic generation of stories [111], or the automatic generation of novel sentences which are indistinguishable from sentences in academic papers [182]. In other text generation problems, the system requires input which informs the output in some

[2] For example, 17 wine review experts in our data wrote about 125,000 reviews, averaging over 7,300 reviews each.

way, but the input is not represented as natural language. Examples of this class of problems include automatic image captioning [64], the writing of weather reports based on meteorological data [139], or automatic generation of product reviews [178, 46], which will be the task we study here.

These text generation problems have some similarities to tasks that take some human-produced text as input, and attempt to alter it in some way to produce related, but novel text. Thus, this is related to the style transfer (see chapter 4 of this thesis and [175, 27]), but also to the problems of summarization [39, 123] and machine translation [36, 85]. In these translation-related tasks however, the desired “content” is provided to the system as natural language. While text generation problems often have some input which should be considered in creating the output, the relationship between input and output is much more tenuous. This means that generation systems have more room for “creativity” in their outputs, but also makes them harder to evaluate, especially with automated metrics.

Our approach to the problem of text generation is consider the input required for text generation systems as a kind of “non-natural” language. This allows relevant information to be easily encoded in a modular and understandable way. It also makes the input easier to interpret and modify for humans interacting with such a system, and enables the application of powerful machine-translation models to generation problems.

Machine translation is a heavily studied task. For translation, much of the earliest literature focused on rule based methods [81] which were later replaced by statistical methods [20, 85]. More recently, neural networks have been widely applied to these types of tasks, achieving state of the art results on a variety of language generation tasks including translation.

One of the first neural machine translation system appeared in 2014 [36]. This work employed 2 recurrent neural networks, one which took a source sentence and created a vector representing it, and one which took a vector representing a sentence and created a sentence in the target language. Since then, there have been many breakthroughs which achieve state of the art results by adding a variety of features to this basic model. Improvements include adding more layers [157, 170], giving the internal nodes more complex hidden states [157], and allowing the decoder network to look at the outputs of the encoder network at each step rather than simply looking at the final output vector [11]. The idea of looking at individual outputs of the encoder rather than only the final aggregated vector is called an *attention mechanism*. Recently, a neural translation model called a *transformer* network was introduced which relies heavily on attention [163]. In this architecture there is no recurrence in the networks. Instead, the encoder creates embeddings of each word in the source sentence and updates them in each layer while attending to the embeddings of each other word in the sentence. The decoder attends to both the embeddings of the encoder and its own embeddings from the previous layer to generate a translated sentence. The transformer network was found to create better translations and to take less time to train than previous methods.

By formatting the relevant traits of our input as text, we are able to build on these machine translation results and apply these state of the art architectures to the less bounded task of product review generation.

3.1.1. Online Reviews

Aside from the disciplinary advance that we offer, a system could have immediate uses so we ask if machine-written product reviews (i.e., reviews generated by algorithms, using various "features" of the product of interest) can be as readable, engaging and informative to humans as human-generated (i.e., human-written) reviews. Machines capable of taking the product attributes as inputs and generating a human readable review as an output could thus offer the first draft of the review and help the expert reviewer in writing their review. We ultimately envision a process where machines would complement and supplement human-written reviews where heretofore, the writing of expert reviews has been a singularly human endeavor.

The idea that machines could write reviews to help reduce the workload of expert reviewers is a natural one given the recent meteoric rise in the use of machine learning and artificial intelligence systems for marketing. Technological advances in this space have enabled easy processing of large-scale unstructured data and have demonstrated strong predictive performance in several marketing tasks given their flexible model structures [105]. Several traditionally human-handled marketing tasks such as the targeting of ads [61], chatbots [102], and services [71] have now been successfully transferred to machines.

Online reviews are impactful. Early research on online reviews found that aggregate measures of review data such as valence (mean rating), volume, and variance [155] affect sales in a variety of contexts such as books [34], movies [35], and video games [185]. At the same time, other work documented no effect of mean reviews on sales [47, 98]. Later studies shifted the focus to individual reviews [161] and looked at systematically studying them [125, 21].

Going past the effect of reviews on sales, recent work has examined the evolution of online reviews [57], customer-to-customer interaction [79], including the effect of past reviews on future ones [118]. Further, researchers have explored the effect of online reviews on non-sales related outcomes such as brand equity [12], consumer preferences for future product development [43], and for designing ranking systems for recommending products to consumers [56].

Closer to this work, we find that consumers rely on experts, critics, and opinion leaders for a variety of choices[22]. In particular, this reliance is strong for experience goods due to their “need for touch” [131] such as books, movies, and the product category of our empirical context, wines [119].

3.1.2. Wine Reviews

Reviews are of particular importance for the purposes of wine marketing and sales. Many consumers turn to wine reviews for guidance in their purchasing decisions. There are many sources for these reviews. *Wine Spectator* and *Wine Enthusiast*, offer expert opinions on tens of thousands of wines a year[3]. These reviews are usually both quantitative and qualitative, and consist of both a numerical (or sometimes categorical) rating and a textual description of the sensory experience of tasting the wine and have been the subject of a substantial and growing body of research. Friberg and Grönqvist [53] find that a positive review increases demand for a wine, that neutral reviews have a smaller positive effect, and that negative reviews have no effect. Chaney [31] examines reviews in newspapers and wine journals and note that the profile of companies mentioned in reviews raises after a review and that reviews may have the power to influence retail sales.

Moon and Kamakura [119] note that consumers rely on expert opinions specifically for wines as the sensory perception of the product is of a complex nature since the experience of a wine has several dimensions including color, aroma, taste, mouth feel, and appearance. Similarly, [8] and [130] argue that the enjoyment of a wine is both sensory and psychological and that reviews may impact the psychological processes of wine tasting. Accordingly, [150] and [42] demonstrate through experiments that consumers’ rating and enjoyment of wines is directly affected by the information and sentiment about the wine they read in the review descriptions.

In sum, while communicating the quality information of a wine effectively is a

[3] For example, *Wine Spectator* and *Wine Enthusiast* combine to create nearly 40,000 reviews per year in recent years (see Figure 3.1).

challenging task given the complexity of the product, at the same time, providing such reviews to consumers is critical as reading wine reviews generally impacts the overall experience of wine consumption. Expert wine reviewers thus play a crucial role in the wine industry.

These experts are expected to assess wines, break down the complexity of the product and communicate it effectively to consumers. Past research focusing on the competency of these experts has looked at convergence and consistency in the reviews within and across various experts. Hodgson [67] finds that only a small portion of experts rate the same wines consistently upon repeat sampling. Other studies have found a similar lack of consensus among experts [68, 9].

At the same time, other studies have found more support for the expertise of wine reviewers. In [153] consensus among four wine publications is examined (including *Wine Enthusiast*, the source of data used in this paper). While the correlation between different pairings of the publications vary, they report a moderate degree of consensus across the four publications. Similarly,[7] and [107] study at the ratings given by a small number of world-renowned experts and found higher levels of correlation than previous studies.

Overall, this stream of research suggests that while expert reviews of wines may diverge, the textual descriptions still do communicate useful information about a wine to the average consumer. Wine reviews and descriptions have been found to increase both sales and perceived enjoyment by consumers. The important factor in all this seems to be that a review exists and has been read by the consumer before consumption. This seems especially to be the case with novice consumers.

Section 3.2

Data

Product reviews have been previously used as a dataset for a variety of tasks. Sometimes the goal is to classify the review as in the case of sentiment analysis [49]. Other authors have attempted to rewrite existing reviews to change their characteristics [92, 183]. Still others have used the reviews of products to build systems to recommend new products to individual users [109, 184]. There is also prior research has also studied the task of automatically generating reviews for a specific product [178, 46].

The sources of review data have been similarly diverse and the specific information associated with each review has varied. Product reviews from Amazon.com are widely used, sometimes covering all products [95], but sometimes covering only specific categories such as books [46]. The user assigned rating for these products consists of only a single score, but in some cases the user is asked to rate the product on a variety of aspects. One such paper used Chinese reviews of cars [178] where the aspects rated included comfort, appearance and power. Another used online reviews of beer in which users rated each product on feel, look, smell, taste, and gave an overall score [110].

We collected reviews which were published in the years 1999-2016 from winemag.com. For each wine, this data consists of information about the wine itself, such as name and style, and information about the review including author, score, and the actual text. In some reviews a few of these fields are not provided and are marked with "N/A". We collect and use these reviews as they are except when there is no text representing the actual review in which case we discard the review from our data. In total this process left us 201,431 unique wines in our dataset, approximately 125,000 of which are identified as having been written by one of 17 authors.

In order to prepare the data for use by our model, we need to modify the format. We create two parallel files, one holding information about the wine and reviewer, and the other contains the actual written text of the review. Each wine is represented by a single line in each file.

In the meta-information file we include the following fields: name of the wine, points(rating given by reviewer), review author, price, designation, variety(type, merlot, etc), Appellation (geographic location where grapes were grown), winery, alcohol content, category(red, white, rose, dessert, fortified etc.), and date of review. Table 3.1 shows some examples of the formatted wine info and reviews.

Wine Information	**Wine Review**
<Ancient Peaks 2010 Cabernet Sauvignon (Paso Robles)> <84> <N/A> <$17> <N/A> <Cabernet Sauvignon> <Paso Robles, Central Coast, California, US> <Ancient Peaks> <14%> <Red> <12/31/2012>	This is a sound Cabernet. It's very dry and a little thin in blackberry fruit, which accentuates the acidity and tannins. Drink up.
<Feiler-Artinger 2012 Beerenauslese Traminer (Burgenland)> <92> <Roger Voss> <N/A> <Beerenauslese> <Traminer, Other White> <Burgenland, Austria> <Feiler-Artinger> <12%> <Dessert> <11/1/2013>	Very young wine, dominated by its textured spiciness and ripe tropical fruits. It is rich, full of pepper, cinnamon, and a dry core that makes it as much rich as sweet.

Table 3.1: Examples of Formatted Parallel Data Used

Figure 3.1 shows the distribution of wine category by year while Figure 3.2 presents the distribution of the rating in the corpus of reviews. As can be seen, a majority of wines in the data are red wines and the wine ratings range between 80 and 100. Due to the proprietary nature of this data (owned by the website), we are unable to make the full data publicly available, but others interested may collect it directly from winemag.com using standard scraping techniques.

We construct a single vocabulary of approximately 20k tokens built on both the

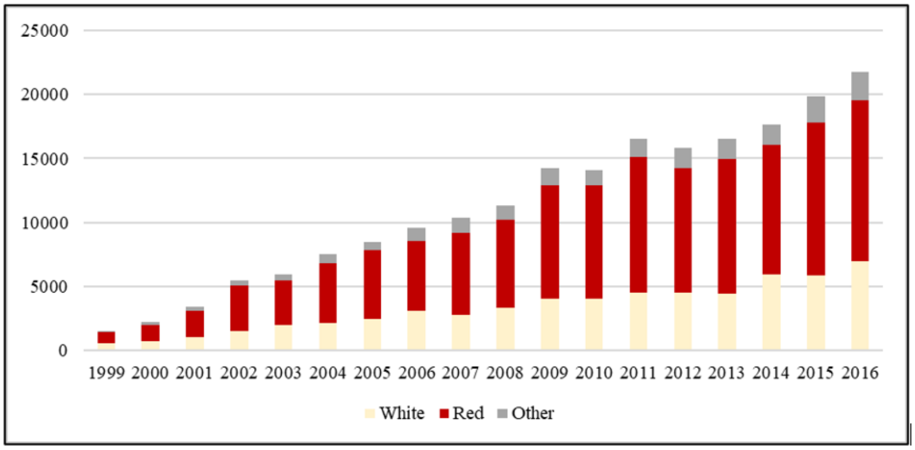


Figure 3.1: Category of Wines in the Data by Year

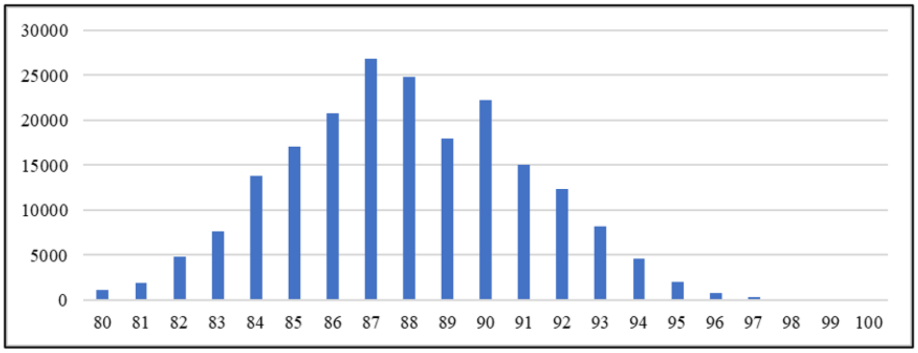


Figure 3.2: Distribution of Rating of Wines in the Data

wine attributes and the review text. Words which appear infrequently are not in the vocabulary directly, but subwords are included to represent them. For example, the word “government” is rare in our data and is represented by three tokens, “go”, “vern”, and “ment”, when it appears.

We split the reviews into three sets at random. The main set which we use to train the model consists of 90% of our data, while the test and development sets each contain 5%.

This data has some notable differences from previously used review datasets. For one, we believe the information associated with each wine is extremely rich. In [46] the task is to generate reviews for books, but the system is only provided with a product

ID, a user ID, and a rating. In our work we provide our model with information about many attributes of the product, such as variety, price, and category. This allows it to learn the contribution of these attributes to the review text, both on their own and their interaction with one another.

In addition to a wider range of attributes than most previous review datasets, we benefit from all reviews being professionally written by a fairly small number of experts. This means that the text of the reviews themselves is of consistently high quality and we don't have extremely short or informal text that would be found in reviews made by the general public. The small number of authors is potentially helpful in allowing us to learn specific preferences, both in wine and in the use of language describing wine, of individual reviewers.

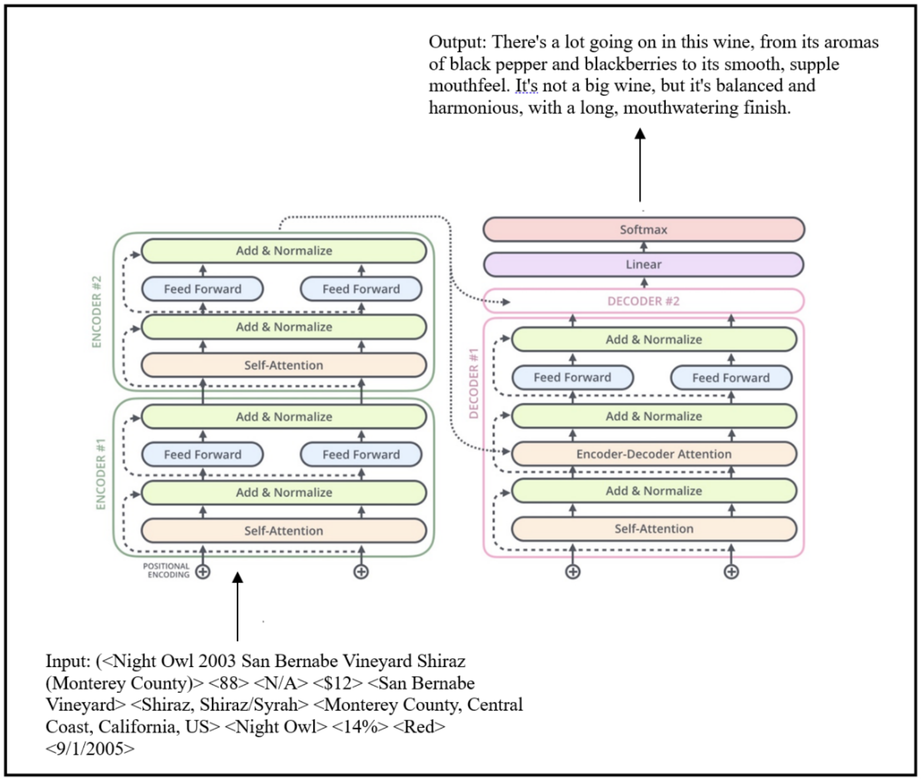
Section 3.3

Model and Experiments

We use a transformer network [163] created with the publicly available tensor2tensor [162] library [4]. Like many neural translation models, the Transformer model consists of two components: an encoder and a decoder. Figure 3.3 gives a standard sketch of the network, along with an actual example input and machine generated output from our work. As shown in the figure, the goal of our modeling approach is to have wine-related data (such as the name of the wine and style, review score, etc.) as inputs to the model and have a fully-written review describing the wine as an output. In this section, we will unpack some of the internal blocks of the model.

The model takes in as input a sequence of input tokens, $x_1, ..., x_n$. As shown in Figure 3.3, the inputs are the words describing the attributes of a wine (the metadata). The first component of the model is the encoder, which takes these tokens and creates

[4] A great explanation of transformer networks can be found at http://jalammar.github.io/illustrated-transformer/



Output: There's a lot going on in this wine, from its aromas of black pepper and blackberries to its smooth, supple mouthfeel. It's not a big wine, but it's balanced and harmonious, with a long, mouthwatering finish.

Input: (<Night Owl 2003 San Bernabe Vineyard Shiraz (Monterey County)> <88> <N/A> <$12> <San Bernabe Vineyard> <Shiraz, Shiraz/Syrah> <Monterey County, Central Coast, California, US> <Night Owl> <14%> <Red> <9/1/2005>

Figure 3.3: Representation of the Transformer network with actual example input and output. Source: http://jalammar.github.io/illustrated-transformer/. Used with permission.

vector representations, $z_1, ..., z_n$, of the tokens, representing each token as a real-valued vector in a common high-dimensional space. The z_i created here are akin to word embeddings [116]. The actual coordinates are in fact a reflection of parameter weights in an underlying neural network. The embedding is such that the relative positions in this high-dimensional space of each word (token) in the overall embedding of the vocabulary provide a measure of the likelihood – reflecting the interpoint distance of the embeddings – of their co-occurrence in the wine reviews.

Each encoder layer takes the embedding of the previous layer (the first layer takes a word embedding combined with an embedding of the position of the token) to update each token's embeddings. In this process, it uses a "self-attention mechanism" (generally, "attention" refers to the ability to base calculations on specific embeddings created earlier or in the case of self-attention, simultaneously). That is, the embedding is successively updated based on the embeddings of the other input tokens from the same level of the encoder. This allows for successive and simultaneous integration of parameter information that effectively incorporates “neighborhood” information from the training data.

Similar to the encoder part, the decoder part also consists of several decoder layers and operates similar to the encoder, although roughly backwards. That is, each decoder layer also adds an additional attention module which attends to the embeddings produced by the final layer of the encoder. Each decoder layer produces one output token at a time and the attention mechanism of each embedding is only allowed to see tokens that are from earlier in the output than the current token. For example, the 4th token of the decoder can only attend to the first 3 output tokens' embeddings in each layer. In this manner, the embedding of each decoder position can rely on the embeddings of all of the embeddings of the input as well as the embeddings of the output from earlier time steps.

Once embeddings from the final layer of the decoder are produced, softmax is used to turn these embeddings into conditional probability distributions over the words in the vocabulary. The model has the encoder's z_i vectors that are conditioned on the input X vector and since it has the previous tokens of the output, the Transformer architecture can condition on the output of previous timesteps, i.e., those $y < t$, thereby producing an overall conditional probability on output word strings. That is,

$$p(Y|X) = \prod_{t=1}^{T} p(y_t|X, y_{<t}) \tag{3.1}$$

During training, the model is given the input, the information about a wine, and generates probability distributions over the words in the review as output. By comparing the actual human written review with the model generated output at each iteration, the model adjusts internal parameters in all layers of the encoder and decoder so that its predictions get closer to the truth. These internal parameters are grouped into Θ in Equation 3.2 below and modified to minimize the loss function:

$$l(\Theta) = -\sum_{t=1}^{T} logp(y_t|X, y_{<t}; \Theta) \tag{3.2}$$

During inference, these probability distributions are used to identify the most likely tokens to create the new machine review.

We set most of the parameters of the Transformer network as described in [163]. These choices include using 6 layers each for the encoder (where each layer is itself a connection of sublayers of a so-called “multi-head self-attention layer” and a full connected feed-forward neural network) and decoder (which includes a second self-attention layer), embeddings with 512 dimensions, and dropout (node removal) probability of 0.1 between layers. Thus, ultimately the inputs generate 512-dimensional representations (embeddings) of the tokens.

At each iteration, loss is noted for the development data, which is similar to the test data as both represent a random 5% sample of the collected data. The difference is in their use, with development data being used to evaluate the model during training and test data being used for evaluation of the fully trained model. Training is stopped when the loss on development data has not improved over the best for 2 consecutive evaluations. During our training algorithm, this occurred after 65,000 iterations (steps). Figure 3.4 shows the loss throughout training for both the training and development data (labelled evaluation loss in the figure).

The model is fit using our training data with a batch size of 1024 (meaning each training step processes input with a total of 1024 tokens) and evaluated using the development set each time 2,500 iterations have been run. The trained model is then given wine metadata from the test set and the model automatically generates the

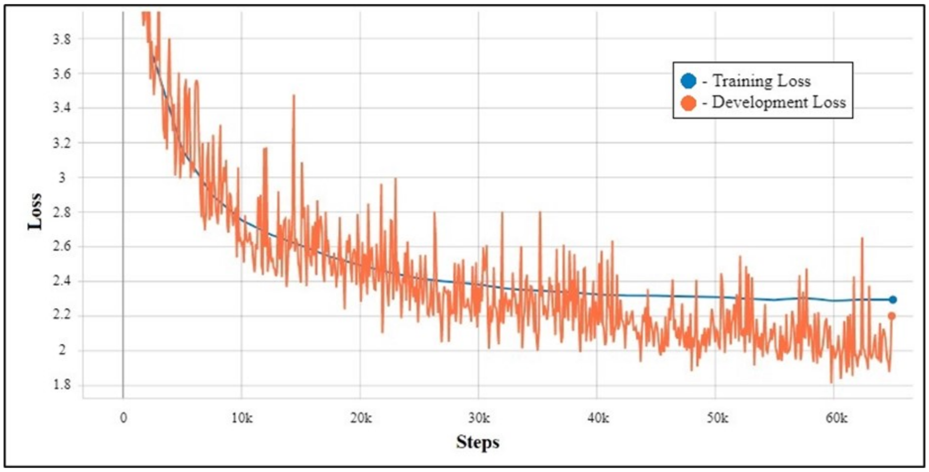


Figure 3.4: Transformer Loss During Training

review text for these wines in the test data. As mentioned above, we train the model using 90% of our data. This splitting of the data means that the model never sees the wines in the test set during the training phase, an important point for the system to have real applications.

After creating reviews for every wine in the test set, we then perform a class of experiments by altering some characteristics of the input. In particular, we experiment with both changing the author and the rating of the review. Since both fields are independent of the wine itself, this allows us to generate multiple reviews for the same wine – first as experienced by a different "expert" and second as experienced differently by the same "expert". A human using this system could then choose and possibly modify the review they believe to be most suited to their needs. To generate even more reviews, we also experiment with some changes to the parameters of the deep learning model. During review generation we alter the alpha parameter, which controls the length penalty applied to output, and the beam size, which affects how many candidate reviews the model considers at each step as it generates the output word-by-word. For a more thorough description of these model parameters see [163, 162] and the references therein.

Section 3.4

Results and Analysis

We apply our model on the test data and generate reviews for the wines in that data. We demonstrate here the suitability of this model architecture for our goal of generating machine-written content meant for human consumption by exploring the sensitivity of the model to model parameters as well as input data.

First, we focus on two model parameters: beam size and alpha. The former limits the number of alternatives considered at each step of the generation – the “expansiveness” – and the latter the length of the review. Table 3.2 shows reviews generated by the model for one example wine (a Merlot) in the test data for high and low settings of the two parameters. As can be seen, the model is responsive to these parameters and generates slightly different reviews for each set of parameters. For the reviews presented in the rest of the paper we set alpha to 0.7 and the beam size to 20.

Beam Size	**Alpha**	**Generated Review**
50	0.6	There are some good flavors of blackberries and cherries in this dry, full-bodied wine. It’s a little rough around the edges, though.
20	0.6	There’s lots to like in this dry, full-bodied wine. It has flavors of cherries and herbs, and is very dry.
20	0.7	There’s lots to like in this dry, full-bodied wine. It has pleasant flavors of cherries and blackberries, and is very dry.

Table 3.2: Reviews Generated with Different Decoder Settings

Second, we experiment with varying the metadata to test if the machine is capable of modulating the review to reflect the desired change. In Table 3.3, we choose a wine (Syrah rosé) which was rated poorly by the expert in our data (84) and generate a corresponding review from the model. We next generate a review after changing the given rating by increasing it from 84 to 97. The machine reviews are displayed alongside the human produced review. This experiment has face validity since increasing the rating leads to the model generating more positive words in the review, better reflecting the rating of 97.

Rating	**Source**	**Review**
84	Human	This blush wine has Viognier blended in with Syrah. It's a pretty salmon color, and has flavors of raspberries, peaches and vanilla. Fully dry, it will be nice with grilled salmon.
84	Transformer	Pretty dark for a rosé, and full-bodied, with cherry, raspberry, vanilla and spice flavors. It's dry, with good acidity.
97	Transformer	This is one of the best California rosés I've ever had. It's so dry, so rich in fruit, it's almost like a dessert wine, except it's totally dry. The flavors are enormously complex, ranging from cherries and raspberries to white chocolate, and the finish goes on and on.

Table 3.3: Model Generated Reviews with Increased Ratings Along with the Human Review

Similarly, the third analysis we conduct involves decreasing the rating of a wine. Table 3.4 shows the results of choosing a highly rated wine (Cabernet Sauvignon) where the expert in our data had rated it at 94 and its corresponding machine generated review. We then decrease the rating to 83 and regenerate the review by the machine. As can be seen, in the review with the reduced rating, the machine-generated review is quite muted and neutral, which is in concert with the rating of 83.

Fourth, we check if the style of review varies with the author. To do this, we regenerate the review by varying the author and leaving everything else fixed. Table 3.5 shows the results of this analysis from an example wine. Once again, the results show that the model is able to pick up the writing styles of particular authors and is able to modify the style accordingly. Overall, we note that our Transformer architecture model appears to be matching many of the human observations about the

Rating	**Source**	**Review**
94	Human	The estate vineyard is in the Vaca Mountains, east of the city of Napa. You can taste the cool climate in the brisk acidity that frames this 100% Cabernet Sauvignon. It's very ripe in blackberries, with rich layers of toast from new French oak. For all the sweet fruit, the finish is dry, even rasping. Such are the tannins that the wine feels tough, even at this age. Give it another five years, at least.
94	Transformer	This is one of the best Cabernets of the vintage. It's so delicious, you can hardly keep from drinking it now, but it should develop over the next 6–8 years.
83	Transformer	Soft and simple, with sweet-and-sour cherry and raspberry flavors. Drink now.

Table 3.4: Model Generated Reviews with Decreased Ratings Along with the Human Review

wines with good accuracy. For example, in the first 30 reviews of the test set, the human generated text recommends ageing 4 of the wines. The model created reviews for these 30 wines recommend ageing 7 of them, but all of the human recommended wines to age are amongst those 7.

Section 3.5

Human Evaluation Survey

The previous section provides some examples of machine-generated reviews and shows how varying the metadata is reflected in the review. This begs the main question addressed in this paper: “does a machine-written review seem human-written”. That is, given a corpus of human-written reviews plus product characteristics, can a machine learn to write a wine review of “human quality”? In the spirit of a Turing Test, we assess whether the machine is able to generate reviews similar to that of human beings with an experiment involving an independent sample of five hundred and one Amazon Prime Mechanical Turk participants.

Author	Source	Review
A	Human	Golden and extracted. Butter-sautéed bananas, intense tangerine and peach and pineapple custard and vanilla are among the flavors. Good acidity, but so sweet and delicious.
A	Transformer	If you're looking for a dry, crisp white wine, try this one. It's filled with citrus, peach, wildflower and vanilla flavors, with a rich, creamy texture and a long, spicy finish.
B	Transformer	From a small parcel of old vines, this is a rich, full-bodied wine, packed with flavors of tropical fruits, spices and a touch of vanilla. It's a great food wine.

Table 3.5: Reviews Generated with Different Authors Targeted by the Model (Along with the Human Review)

Our survey tests the degree to which machine generated reviews appear to the reader as reviews written by human beings who are expert wine reviewers. We also test how likely consumers are to purchase the wine based on the reviewer (human) generated review versus the machine generated review. Our survey puts respondents in the shoes of online shoppers and tests what decision they would make when a wine review is presented to them.

The survey was completed by five hundred and one respondents from ten states in the U.S. Each respondent earned a reward of $1.50. We did not delete any observations either due to non-compliance or for any other reasons. Overall, about 39.3% reported being female and 60.3% male, with 74.3% having Bachelor's degree or higher, and 50.7% with household income $50,000 or higher.

Each survey respondent saw fourteen randomly selected reviews from a set of six hundred reviews of which 300 were generated by human beings (expert wine reviewers) and the rest were generated by our Encoder/Decoder Transformer network. Note that the set of machine-generated reviews and the set of human-generated reviews are based on the same set of 300 wines. As mentioned before, all the 300 wines are part of the test set. That is, none of the wines in the test set were used to train our Transformer.

Respondents were told that some of the reviews were written by human beings who are wine reviewers and some were "written" by a machine. We asked them to read each review carefully and indicate to what extent they thought the review was written by a machine or a human being. We included two questions to uncover whether customers are able to figure out if the review was written by a human or a machine: (i) by indicating if the review was written by a human being or a machine or unable to tell, and (ii) an allocation of 100 points to the two types: review written by a human being and review generated by a machine. For each review, we also asked the participants to indicate the likelihood they would purchase the wine either as a gift or for themselves.

While 10.96 percent of respondents said they were unable to tell if a review was written by a human being or a machine, a chi-square test ($\chi^2 = 0.24$ with $df = 2$) revealed no significant difference ($p = .886$) for this question between the two true sources of reviews: human being versus a machine. Similarly, there was no significant difference in responses from the two types of review source in terms of their purchase likelihood ($\chi^2 = 6.32, p = .176$). This provides evidence that, from a consumer's perspective, while the reviews were either written by an actual reviewer or automatically generated by the machine using our model, there were no perceived differences regarding the true review source. There was also no significant difference in respondents' mean allocation of 100 points to each of the two types of sources: written by a human being versus generated by a machine ($p = .99$).

In order to fully understand whether consumers are really not able to tell the difference between reviews that were generated by a human being versus those that are automatically generated by the machine, we estimated a random effects regression model that controls for the following: (1) review specific effects, (2) respondent specific effects. The dependent variable in the regression is the number of points (out of 100) each participant allocated for each review in terms of how likely the review was generated by a machine. The key independent variable is the true review source, which is a dummy variable, 1 for human-generated and 0 for machine-generated. Respondent specific effects were controlled by 500 dummy variables to represent the 501 respondents. Review specific effects were controlled via a random-effects specification. In addition, the standard errors were clustered at the review level.

Table 3.6 presents the results. As seen in Table 3.6, after controlling for review specific effects, respondent specific effects, and clustering the standard errors at the review level, the coefficient for the review source is insignificant ($p = .57$). This provides further support to our analysis that the true source of the review had no significant impact on the respondents' perceptions of the source of the review being human-generated or machine-generated.

Dependent variable: How much (out of 100 points) Mturk respondent thought it was a machine generated review	Mean	SE
Review source is machine (dummy variable)	0.409	0.726
Individual specific effect	FE	
Review specific effect	RE	
Number of individuals	501	
Number of reviews	600	
N	7014	
R^2	0.167	
Standard errors clustered at the level of reviews		

Table 3.6: Panel regression with review-level random effects and individual-level fixed effects

Section 3.6

Conclusions

In this work, we sought to address the following question: "To what extent can a machine learn to write a review that is as engaging, informative, and appropriate as a human-written review?" While extant literature focuses on using natural language processing to generate text that resembles human-written text, there is scant research on its application to online reviews. Further, to our knowledge, there is no research on directly testing human versus machine-generated reviews.

We address these gaps in the literature and demonstrate that a deep machine-learning approach based on the Transformer network can produce (or "write") reviews of wines that are similar to those written by wine experts. We also test 600 randomly selected reviews (300 human-written reviews and 300 machine-generated reviews for the same set of 300 wines) with 501 Amazon Prime MTurk participants. We find no significant difference in respondents' identification of whether a review was written by a machine or human being. We thus show that machines can indeed learn to write "human-quality" reviews. This issue is relevant for consumers and retailers alike.

Possible applications of this work include not only aiding experts in their writing of wine reviews as previously discussed, but also direct use by wineries to provide descriptions for their labels. Wineries usually include some information on the labels, and research suggests that an "elaborate taste description" is one of the features most valued by consumers [122]. The presence of these descriptions likely increase consumers' likelihood of purchasing a wine and their enjoyment [42, 150]. This means the wineries have incentive to generate descriptions for their labels, and those who do also have costs associated with this task.

Researchers have noted that positive reviews in publications increase sales [53], and suggested that detailed information about wines is valued by consumers making a purchasing choice [31, 42]. Additionally, reading information about a wine before drinking it appears to actually increase the satisfaction of the imbiber [150, 42]. Reviews of wine then do provide value, both to the seller and consumer, even if the specifics of the description seem to be less important.

Furthermore, we have shown that by varying input to the model such as the rating of the wine and the author of the review, we can generate additional candidate descriptions of the same wine. We envision a process which does not eliminate the need for human reviewers, but where their task is made easier through computer assistance. A wine writer tasked with creating a review could taste a wine and automatically generate several suggested reviews. The reviewer could choose the one they feel best describes the wine and edit it as necessary. For a reviewer whose first language is not English, this process may be especially useful as it can provide candidate reviews which may include words the writer would not have thought of on their own. Thus, having a human expert still in the loop allows them to choose a description which is accurate. This process ensures that the review created is still useful to other experts and instills confidence in the validity of the review to readers who are aware of the computer assistance.

In sum, tens of thousands of wines are reviewed annually and creation of these reviews incurs cost as they are crafted by expert wine tasters. In this work, we demonstrate a potential way to decrease the burden of work on these experts by automating part of the process. We show that deep neural networks, given enough examples of existing reviews, can generate reviews for new unseen wines that are indistinguishable from human-generated reviews. The proposed system can generate multiple reviews for a single wine, and a taster could use these many reviews as a resource when creating their own. One of parameters that can be changed during review generation is the author, even allowing wineries or reviewers to create reviews which target a specific writing style. Thus, our model makes the process faster, easier, and less costly.

Chapter 4

Style Transfer in Text

Section 4.1

Task and Background

Written prose is one way in which we communicate our thoughts to each other. Given a "message", there are many ways to write a sentence capable of conveying the embedded information, even when they are all written in the same language. Sentences can communicate essentially the same information but do so using different "styles". That is, the various versions may have essentially the same meaning or semantic content, and insofar as they use different words are each "paraphrases" of each other. These paraphrases, while sharing the same semantic content, are not necessarily interchangeable. When writing a sentence we frequently consider not only the semantic content we wish to communicate, but also the manner, or style, in which we express it. Different wording may convey different levels of politeness or familiarity with the reader, display different cultural information about the writer, be easier to understand for certain populations, etc. *Style transfer*, or stylistic paraphrasing, is the task of rewriting a sentence such that we preserve the meaning but alter the style.

Style transfer has obvious connections to work in traditional language-to-language translation and paraphrasing. Framed in this way it makes sense to try out some of the advanced deep learning translation models on the style transfer problem. The particulars will be given below, but here we briefly summarize the relevant models and related background.

The *Seq2Seq model* was first created and used in conjunction with statistical methods to perform machine translation[36]. The model consists of a recurrent neural network acting as an encoder which produces an embedding of the full sequence of inputs. This sentence embedding is then used by another recurrent neural network which acts as a decoder and produces a sequence corresponding to the original input sequence.

Long Short-Term Memory (LSTM)[66] was introduced to allow a recurrent neural network to store information for an extended period of time. Using a formulation of LSTM which differs slightly from the original[59], the Seq2Seq model was adapted to use multiple LSTM layers on both the encoding and decoding sides[156]. This model demonstrated near state-of-the-art results on the WMT-14 English-to-French Translation task. In another modification an attention mechanism was introduced[10] which again achieved near state-of-the-art results on English-to-French translation.

Other papers proposed versions of the model which could translate into many languages[10, 51], including one which could translate from many source languages to many target languages, even if the source-target pair was never seen during training [75]. The authors of this work make no major changes to the Seq2Seq architecture, but introduce special tokens at the start of each input sentence indicating the target language. The model can learn to translate between two languages which never appeared as a pair in the training data, provided it has seen each of the languages paired with others. The idea of using these artificially added tags was applied to related tasks such as targeting level of formality or use of active or passive voice in produced translations [148, 176].

This work on machine translation is relevant for *paraphrase generation* framed as a form of *monolingual translation*. In this context statistical machine translation techniques were used to generate novel paraphrases[138]. More recently, phrase-based statistical machine translation software was used to create paraphrases[171].

Tasks such as text simplification [152, 172] can be viewed as a form of style transfer, but generating paraphrases targeting a more general interpretation of style was first attempted in 2012[175]. All of these results employed statistical machine translation methods.

The advances mentioned previously in neural machine translation have only started to be applied to general stylistic paraphrasing. One approach proposed the training of a neural model which would "disentangle" stylistic and semantic features, but did not publish any results[77]. Another attempt at text simplification as stylistic paraphrasing is [167]. They generate artificial data and show that the model performs well, but do no experiments with human-produced corpora. The Shakespeare dataset [175] recently was used with a Seq2Seq model [74]. Their results are impressive, showing improvement over statistical machine translation methods as measured by automatic metrics. They experiment with many settings, but in order to overcome the small amount of training data, their best models all require the integration of a human-produced dictionary which translates approximately 1500 Shakespearean words to their modern equivalent.

Many attempts at style transfer focus on only a particular aspect of style such as formality [176] or simplicity [152]. Some target several of these factors at once [50]. While these approaches do change the style of the text, they do not necessarily do so in a way that can be generalized to other related tasks. One can imagine for example, that a system for simplification of text may benefit from including a training objective of metrics which capture textual complexity. Such an approach requires a suitable equivalent metric. Furthermore, it may be difficult, or impossible, to create an exhaustive list of specific stylistic attributes.

For these reasons, we draw a distinction between tasks like simplification and the task of holistic style transfer. In the latter, the characteristics which contribute to style are neither considered separately nor explicitly. The goal of this task is to create a system which can implicitly capture all stylistic differences when changing text from one style to another.

This holistic view of style transfer introduces some difficulties. For example, while some metrics exist for analyzing the simplicity of text [82, 151], metrics to evaluate holistic style transfer are less developed. In section 4.2 we discuss the use of some existing metrics and introduce some ideas to aid in this evaluation. These derive – in part – from viewing style transfer as a machine translation problem where the source language and target language are simply different textual styles. The style transfer task is treated this way in much of the existing work on style transfer and related problems [32, 175, 172, 74]. Despite this obvious connection, many breakthroughs in machine translation systems have not been directly applied to style transfer. As many previous authors have noted, the major difficulty seems to lie in finding a suitably large parallel corpus of different styles [78, 140, 174, 54]. In section 4.3 we introduce a dataset of Bible versions for style transfer and related tasks. We believe this data is highly suited to the task for many reasons including pre-existing alignment because of verse numbers and the presence of clearly distinct writing styles.

Using these metrics and our new dataset, we then approach the problem of supervised holistic style transfer in section 4.4. Since aligned (parallel) data is rare in the style transfer setting we close with a section in which the Bible data as non-parallel and perform an unsupervised version of the task in section 4.5.

Section 4.2

Metrics

4.2.1. BLEU and PINC

For evaluation of our style transfer models we use several established measures. We first calculate BLEU (bilingual evaluation understudy) [129] scores for our results. BLEU is a metric for comparing parallel corpora which rewards a candidate sentence for having n-grams which also appear in the target.[1] Although it was created for evaluation of machine translation, it has been found that the scores correlated with human judgement when used to evaluate paraphrase quality [171]. The correlation was especially strong when the source sentence and candidate sentence differed by larger amounts as measured by Levenshtein distance over words.[2]

BLEU gets at some of what a good paraphrase should accomplish (similarity), but a good (i.e., interesting) paraphrase should use different words than the source sentence, as noted by Chen and Dolan [32]. They introduce the PINC (**p**araphrase **in** **n**-gram **c**hange) score which "computes the percentage of n-grams that appear in the candidate sentence but not in the source sentence" (see [32] for a clear and more thorough description). The PINC score makes no use of target sentence, but rewards a candidate for being dissimilar from the source. According to the metric's authors "In essence, it is the inverse of BLEU since we want to minimize the number of n-gram overlaps between the two sentences". To capture a candidate's similarity to the target and dissimilarity from the source they use both the BLEU and PINC scores together. They find that BLEU scores correlate with human judgement of semantic equivalence and that PINC scores correlate highly with human ratings of

[1] A good description can be found at https://towardsdatascience.com/bleu-bilingual-evaluation-understudy-2b4eab9bcfd1

[2] As per Wikipedia: "Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other."

lexical dissimilarity. Lexical dissimilarity on its own is important for paraphrasing, but as previously mentioned, a high lexical dissimilarity may also strengthen the correlation of BLEU score and human judgement of paraphrase quality[171]. We will use and report both PINC and BLEU for evaluation as can be found in previous work on stylistic paraphrasing [175, 74].

4.2.2. Frequent Idiosyncratic Words

This combination of BLEU and PINC scores for evaluating style transfer in text has been frequently used in other work [175, 74, 27], but not without criticism [159, 174]. Arguably, style transfer, especially for the situation in which there is no parallel (aligned) text, cries out for new kinds of measures. Classical stylometric measures could provide such a source. Some approaches to stylometry are structural, while others focus on word usage frequency. Function word approaches consider the usage of common words [121].

If we had a perfect system for holistic style transfer then we would expect stylometric measures to produce high correlation in such measures between the target sentences and the produced output. In defining the task of holistic style transfer we viewed particular aspects of style, such as simplicity, to be too narrow and fail to capture the entirety of what people think of as the style of text. When it comes to evaluation, however, all of these narrower indicators of style must be present in the candidate output if the style has been accurately changed. With this intuition, and the necessarily vague definition of holistic style transfer, one approach for evaluation of transfer quality would be to analyze our system with a suite of metrics measuring individual components of style. While the list of relevant metrics is not likely be exhaustive, just as a list of aspects of style would be incomplete, the measures are still likely to be useful.[3] Imagine a system trained for holistic style transfer, with-

[3] In one of the earliest stylometric efforts – Wincenty Lutoslowski's analysis of Plato's Dialogues

out any knowledge of the stylometrics which will ultimately be used for evaluation. Then imagine the outputs of the system are evaluated using existing style measures. Our belief that the system was performing high-quality holistic style transfer would increase if many existing metrics for individual aspects of style returned high scores. Since the model training was agnostic to these measures, it seems likely that untested aspects of style are as likely to be captured by the model as those we test. A perfect system would have to pass all of these measures, and so the more of them that our candidate model excels at, the more confidence we gain in its quality. In some sense, it faces the same challenges as the analysis of psuedo-random generators [2].

Thus inspired we augment the use of BLEU and PINC through the identification of *frequent idiosyncratic words*, words that appear often in one style but are absent or rare in another. The intuition is that a model which often produces words which never appear in the ground-truth text of the targeted style cannot be performing well. Similarly, a model which fails to produce words that are frequent in the target style is suspect. As the specifics of the words compared depends on the styles, we will provide more details on this metric below when evaluating our models.

Section 4.3 Datasets

4.3.1. Previously Used Datasets

As mentioned in many previous papers [78, 140, 174, 54], progress on this task has been slowed by a lack of ideal datasets. The existing datasets have their own strengths and weaknesses.

One of the most used style transfer corpora was built using articles from Wikipedia and Simple Wikipedia to collect examples of sentences and their simplified versions[186].

[104]– 500 characteristics (“peculiarities”) were articulated!

These sources also were used with improved sentence alignment techniques to produce another dataset which included classification of each parallel sentence pair's quality [73]. More recently, word embeddings have been used to inform alignment and yet another Wikipedia simplification dataset was released[78].

The use of Wikipedia for text simplification has been criticised generally, and some of the released corpora denounced for more specific and severe issues with their sentence alignments [173]. The same paper also proposed the use of the Newsela corpus for text simplification. This data consists of 1,130 news articles, each professionally rewritten 4 times to target different reading levels.

A new dataset targeting another aspect of style, namely formality, is available[140]. The Grammarly's Yahoo Answers Formality Corpus (GYAFC) was constructed by identifying 110,000 informal responses containing between 5 and 25 words on Yahoo Answers. Each of these was then rewritten to use more formal language by Amazon Mechanical Turk workers.

While these datasets can all be viewed as representing different styles, simplicity and formality are only two aspects of a broader definition of style. The first work to attempt a more generalstyle transfer problem introduced a corpus of Shakespeare plays and their modern translations for the task [175]. This corpus contains 17 plays and their modernizations from http://nfs.sparknotes.com and versions of 8 of these plays from http://enotes.com. While the alignments appear to mostly be of high quality, they were still produced using automatic sentence alignment which may not perform the task as proficiently as a human. The larger sparknotes dataset contains about 21,000 aligned sentences. This magnitude is sufficient for the statistical machine translation methods used in their paper, but is not comparable to the corpora usually employed by neural machine translation systems.

Most of these existing parallel corpora were not created for the general task of style transfer[186, 73, 78, 173, 140]. A system targeting only one aspect of style may use techniques specific to that task, such as the use of simplification-specific objective functions[174]. So while we can view simplification and formalization as types of style transfer, we cannot always directly apply the same methods to the more general problem.

The Shakespeare dataset [175] (which does not focus on only simplicity or formality) still contains only 2 distinct styles (or 3 if each modern source is considered individually). Standard machine translation corpora, such as WMT-14[4], have parallel data across many languages. A multi-lingual corpus not only provides the ability to test system generalizability, but can also be leveraged to improve results even when considering a single source and target language [75].

Some of these existing corpora require researchers to request access to the data [173, 140]. Access to high-quality data is certainly worth this extra step, but sometimes response times to these requests can be slow. We experienced a delay of several months between requesting some of this data and receiving it. With the current speed of innovation in machine translation, such delays in access to data may make these corpora less practical than those with free immediate access.

4.3.2. The Bible Dataset

Bible versions, with their well-demarcated sentence and verse structure, widespread availability, and many different styles, provide a corpus which is free of many of the problems discussed in subsection 4.3.1 such as only representing differences in one aspect of style or problems with alignment of text.

We identify a novel and highly parallel corpus useful for the style paraphrasing task: thirty-four stylistically distinct versions of the Bible (Old and New Testaments). Each version is understood as embodying a unique writing style. The versions in this

[4] http://www.statmt.org/wmt14/translation-task.html

corpus were created with a wide range of intentions. Versions such as the Bible in Basic English[5] were written to be simple enough to be understood by people with a very limited vocabulary. Other versions, like the King James Version,[6] were written centuries ago and use very distinctive archaic language. In addition to being viewed individually, the versions can also be partitioned according to different stylistic criteria, any one of which could be a goal of a paraphrasing. For example, metrics that enable the identification of versions deemed “simple” could identify a subcorpus that would allow training towards the task of text simplification. Versions identified as using “old" language could be used to train towards the task of “text archaification”. Such richly parallel datasets are difficult to find, but this corpus provides such a wide range of text that it could be used to focus on a variety of stylistic features already present within the data. While many parallel corpora require alignment before they can be used, here verse numbers immediately identify equivalent pieces of text. Thus, in this data the text has all been aligned by humans already. This eliminates the need to use text alignment algorithms which may not produce alignments that match human judgement. Our work splits books of the Bible into training, testing, and development sets. We then publish these sets using all 8 of the publicly available Bible versions in our more complete corpus and list the versions we use which are not public (but could be scraped by the energetic and interested reader). This easy to access and free to use, standardized, parallel corpus is a major contribution of our work.

We collected 33 English translations of the Bible from BibleGateway.com, and also the Bible in Basic English from www.o-bible.com. We found that 7 of these collected versions are in the public domain and thus can be freely distributed. Ad-

[5] http://www.o-bible.com/bbe.html
[6] https://www.kingjamesbibleonline.org/

ditionally, the Lexham English Bible[7] has a permissive license which allows it to be distributed for free. These 8 public versions are used to create the corpus that we release. Other versions can be acquired relatively easily and inexpensively, but may not be distributed due to prevailing copyright law. Table 4.1 displays the complete list of versions.

Public Domain Bible Versions	**Other Versions Used**
Bible in Basic English (**BBE**)	New Life Version (**NLV**)
World English Bible (**WEB**)	New International Reader's Version (**NIRV**)
	International Children's Bible (**ICB**)
Young's Literal Translation (**YLT**)	Easy-To-Read Version (**ERV**)
	New Century Version (**NCV**)
Lexham English Bible (**LEB**)	Contemporary English Version (**CEV**)
	Good News Translation (**GNT**)
Douay-Rheims 1899 American Edition (**DRA**)	God's Word Translation (**GWT**)
	Names of God Bible (**NOG**)
	Jubilee Bible 2000 (**JUB**)
American Standard Version (**ASV**)	New King James Version (**NKJV**)
	Modern English Version (**MEV**)
Darby Translation (**DARBY**)	English Standard Version (**ESV**)
	1599 Geneva Bible (**GNV**)
King James Version (**KJV**)	New International Version (**NIV**)
	Holman Christian Standard Bible (**HCSB**)
	21st Century King James Version (**KJ21**)
	New Living Translation (**NLT**)
	New Revised Standard Version (**NRSV**)
	Common English Bible (**CEB**)
	New English Translation (**NET**)
	International Standard Version (**ISV**)
	Revised Standard Version (**RSV**)
	New American Bible Revised Edition (**NABRE**)
	The Living Bible (**TLB**)
	The Message (**MSG**)

Table 4.1: Names of publicly available Bible Versions and other versions we used followed by their standard abbreviations in parenthesis. Text was collected from Biblegateway.com (and BBE from www.o-bible.com).

[7] http://www.lexhamenglishbible.com/

These Bible versions are highly parallel and high-quality, having been produced by human translators. Sentence level alignment of parallel text is needed for many NLP tasks. Work exists on methods to automatically align texts [186, 73, 40], but the alignments produced are imperfect and some have been criticized for issues which decrease their usefulness [173]. The Bible corpus is human-aligned by virtue of the consistent use of books, chapters, and verses across translations. While many verses are single sentences some are sentence fragments or several sentences. This is not problematic as we only require the parallel text to be aligned in small parts which have the same meaning, but there is no obvious reason that this must be at a strict sentence level.

Some Bible versions contain instances of several verses combined to one. For example, we may find a Bible version with “Genesis 1:1-4" instead of singular instances of each of the four verses. We remove these aggregated verses from our data to keep the alignment more fine-grained and consistent. There are over 31,000 verses in the Bible, so even with this regularization we still have over 1.7 million potential source and target verse pairings in the publicly available data and over 33 million pairs in the full dataset.

To help identify similar versions and those which are quite distinct, we can consider the BLEU scores between the full text of every pair of versions. The results of this analysis on the public Bibles can be seen in Table 4.2. Another representation is given in Figure 4.1, which provides a two-dimensional MDS representation of the table, as well as a two-dimensional MDS representation of the inter-Bible distances derived from the Burrows Delta [5].

Some versions are highly similar according to the BLEU metric but some are quite different. For example treating ASV as a candidate and KJV as the reference has a score of 69.09 but comparing BBE to YLT only gives a score of 11.72. Because of

	YLT	**DARBY**	**KJV**	**WEB**	**DRA**	**LEB**	**BBE**	**ASV**
YLT	100	26.43	23.61	19.33	13.57	15.15	9.42	25.87
Darby	26.46	100	52.79	37.86	23.94	22.44	16.27	55.49
KJV	23.89	53.38	100	41.04	30.18	19.6	17.76	68.72
WEB	19.78	39.49	41.24	100	20.37	30.07	19.15	53.11
DRA	16.3	29.39	35.56	23.69	100	17.76	15.29	31.64
LEB	17.89	26.71	22.72	33.67	17.46	100	18.49	25.98
BBE	11.72	20.35	21.8	22.59	15.49	19.4	100	22.75
ASV	26.48	56.84	69.09	53.01	26.95	22.51	18.72	100

Table 4.2: BLEU scores between full text of Bible versions. The verses of the version of each row are treated as the candidates and the column version's verses are treated as the reference.

this we would expect a system trained to transfer ASV text into the KJV style to outscore one trained for the BBE to YLT task. The range of scores illustrates that within the dataset, there are similar and quite distinct versions.

We hope that the publication of this corpus will lead to the application of some machine translation techniques that were previously not applicable to style transfer, and that over time these techniques can be fine-tuned to better handle the nuanced differences between machine translation and style transfer. The full verse-aligned texts of all public Bible versions are available on github[8].

Some systems do not require a parallel corpus for training at all, both in machine translation [6, 90] and stylistic paraphrasing [54, 149]. In such research, the aligned data we find in Bible versions is still helpful. While training of such models does not take advantage of the parallel nature of the data, the results of these models are still evaluated using parallel data. Even in textual style transfer research which is more focused on unsupervised learning methods, there is a need for parallel text representing different styles for testing purposes.

While we present the corpus for style transfer, it is rare to find data that is

[8] https://github.com/keithecarlson/StyleTransferBibleData

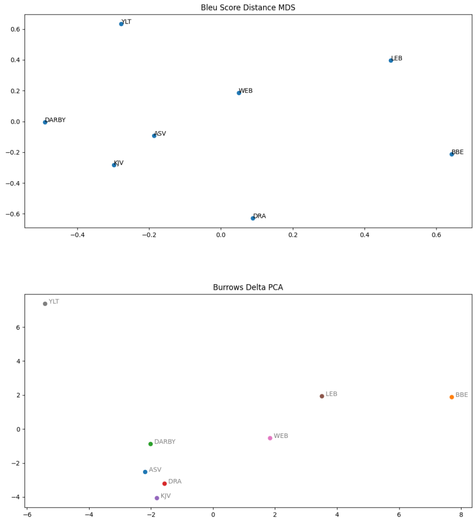


Figure 4.1: MDS representation of the (top) BLEU comparisons for the public bibles and (bottom) Burrows Delta distances between the same.

human aligned and so richly parallel. These qualities may make our corpus useful for a variety of other natural language tasks as well. For example, the large number of aligned translations in this data could prove useful for training towards the traditional paraphrasing task in which a specific style is not targeted. Alternatively, researchers could choose some aspect of style, such as simplicity or formality, and partition the corpus based on that criteria. The partitioned corpus could then be used to train models which produce text with the desired characteristic.

Section 4.4

Supervised Style Transfer in Text

In this section we present our work on supervised style transfer using the (a priori) aligned versions of the Bible.

4.4.1. Data Splitting for Bible Style Transfer

Per the standard learning framework, we to split our data into training, testing, and development sets. We do so by selecting entire Bible books to be included in each set to ensure that the text in the training data is not too similar to anything that appears in testing or development. Additionally, we expect that the language used in the New Testament may differ from that used in the Old Testament. We therefore want to ensure that each of our splits contains books from each of them. The test data is constructed by selecting two random Old Testament books and two random New Testament books, and the process is repeated for the development data. In the published data, the testing set books are Judges, 1 Samuel, Phillipians, and Hebrews. The development set books are 1 Kings, Zephaniah, Mark, and Colossians. The remaining 35 Old Testament books and 23 New Testament books are used as the training split. The verse numbers which are at the beginning of each line are removed as they are always identical for each pair of verses and so make no interesting contribution.

To decide on good target versions we looked at the BLEU scores between the full text of every pair of versions which can be seen in 4.2. We want to consider situations where there is relatively little modification required to the input as well as those which will need drastic stylistic revision. We pick ASV as a single-version target because, when treating all other other public versions as a candidate, it has the highest average BLEU score when treated as a reference. Similarly, we will use BBE as a target since it has the lowest average BLEU score. In addition to creating splits using all public versions as sources, and BBE and ASV as targets, we want to investigate models' performance specific version → version pairs of varying similarities. To this end we also create parallel files of only KJV to ASV (easy), BBE to ASV (hard), and YLT to BBE (very hard). These files can be used to train models on their own, or used to test those which were trained using all public versions for the source and the correct single version as a target. The version pairings for which we create and publish(excluding all → all) parallel training, testing, and development files can be seen in Table 4.3.

Source Versions	**Target Versions**	**# Train Lines**	**# Dev Lines**	**# Test Lines**
All	All	28,693,558	1,707,252	1,920,108
Public	Public	1,534,582	91,780	102,732
Public	ASV	192,414	11,456	12,843
Public	BBE	192,324	11,481	12,843
BBE	ASV	27,584	1,637	1,835
KJV	ASV	27,608	1,637	1,835
YLT	BBE	27,595	1,642	1,835

Table 4.3: Pairings of Bible versions created. For each pairing, parallel training, test, and development files are created and number of lines in each are reported.

The full verse-aligned texts of all public Bible versions are available on github[9]. Additionally we publish the public version parallel testing, training, and development files discussed so that future work with this corpus can make use of a standardized data.

4.4.2. Moses

To test the suitability our dataset and approach to holistic style transfer we will use two existing machine translation systems, carefully formatting and coercing our data to fit the restrictions of each system. First will be the statistical machine translation system Moses [85] which is an established baseline for testing new paraphrasing corpora and models. To use it this way, one recasts paraphrasing as a monolingual translation task on the paired data as mentioned above. Previous such uses include the work of Chen and Dolan [32], Xu et al. [175], and Wubben et al. [171] who found that it outperformed paraphrasing based on word substitution. It was also used as a

[9] https://github.com/keithecarlson/StyleTransferBibleData

baseline in [74] for stylistic paraphrasing of Shakespeare into present day English.

Moses is designed to translate from a single source language to a single target language and in the previous uses of Moses for style transfer only 2 distinct styles were used. Since our corpus contains examples of many styles the proper choice of training data for Moses is not as obvious. We could give Moses training examples using all versions as sources and all versions as targets and it should learn to produce good paraphrases, but we want the paraphrases produced by Moses to be in the style of a specific version. Even when targeting only a single style we have the option to use many versions as source sentences during training or only a single source version. To explore both of these options fully we use each of the single-target parallel files discussed above, those with single source versions and those with multiple, to train Moses models. We train Moses 5 times in total, once each using the (public versions → ASV), (public versions → BBE), (KJV → ASV), (BBE → ASV), and (YLT → BBE) parallel files and then use these models to decode all test sets with the appropriate target version. For example, the Moses model trained on (public versions → ASV) is evaluated on the (public versions → ASV), (KJV → ASV), and (BBE → ASV) test sets.

For all runs of Moses we use mgiza for word alignment [127], kenlm for the language model [65], and mert [127] to fine-tune the model parameters to the development data set. All of these tools are provided with Moses. The language model built is of order 5.

4.4.3. Seq2Seq

Encoder-decoder recurrent neural networks (Seq2Seq) have been widely used and adapted for Machine translation in recent years [10, 36, 45, 156]. One such paper introduced artificial tags at the beginning of each source example to indicate the language to target during decoding [75]. This minor change allowed the model to perform multi-lingual and zero-shot translation. These models generally require a large number of training examples to produce high quality results.

Application of Seq2Seq models to stylistic paraphrasing has not been fully explored. In one such existing work [74] which uses such a network in this context on real data the training corpus is much smaller than ours. To overcome the relatively small corpus the authors [74] use a human-expert produced dictionary giving the translations of Shakespearean words into modern equivalents.

We use our new, and relatively large Bible corpus to train Seq2Seq models. Our corpus contains many versions, and the number of training examples when using a single version as a source and a single version as a target is small. To fully take advantage of how richly parallel our data is, we use a similar idea to the tagging trick in [75]. In each source verse we prepend a tag indicating the version to be targeted. For example if the target style for a verse pair is that of the American Standard Version, we start off the source sentence with an "<ASV>" token. Using this method we are able to train a separate Seq2Seq model using each of the following parallel version pairs (all versions → all versions), (public versions → public versions), (public versions → ASV), and (public versions → BBE). We experimented with running this model on the single source and single target files, such as (YLT → BBE), but the results were poor because the amount of training data was too small for this model.

The Seq2Seq model requires a fixed vocabulary which contains the tokens which will be encountered by the model. Names and rare words are often difficult to handle in NLG tasks. Sometimes they are replaced with a generic "Unknown" token [10, 136, 164]. Byte pair encoding (BPE) can be used to create a vocabulary of subword units which removes the need for such a token [147]. In the resulting vocabulary rare words are not present but the smaller units which make up the word are. For example, in our data the rarely seen word "ascribed" is replaced by the more frequent subwords "as" and "scribed". We generated a vocabulary of the 30,000 most frequent subword units from the training portion of all of our Bible versions. This vocabulary was then applied to each of the samples by replacing any word which was not in the vocabulary with its constituent subword units.

We use a multi-layer recurrent neural network encoder and multi-layer recurrent network with attention for decoding. This set-up is similar to those described by Sutskever et al. [156] and Bahdanau et al. [10].

The encoder and decoder each have 3 layers of size 512 and use LSTM[59], and dropout between layers[179] with probability of dropping set to 0.3. Each uses a trainable embedding layer to project each token into a 512-dimensional vector before it is passed to the LSTM layers. The encoder is bi-directional [10, 146] and has residual connections between the layers [63]. The decoder uses an attention mechanism as described by Bahdanau et al. [10] to focus on specific parts of the output from each step of the encoder. The exact software and configuration of our model can be found on github[10].

During training mini-batches of 64 verse-pairs are randomly selected from the training corpus. Each of the target and source sentences are truncated to 100 tokens if necessary. The model's parameters are adjusted using the "Adam optimizer" [84]. A checkpoint of the model is saved periodically[11] during training. The checkpoints are all evaluated on the development data and the one with the lowest loss is selected.

During inference a single source sentence is fed into the model but the target sentence is not provided. Unlike during training, the decoder is fed its own prediction as input for the next timestep. The decoder performs a beam search [156] with a

[10] https://github.com/keithecarlson/StyleTransferBibleData

[11] We found that models trained on smaller amounts of data tend to overfit faster. To ensure we have a high quality checkpoint we need to save them more frequently when training on the smaller datasets. We saved a checkpoint every 5,000 steps on public → public and all → all training and every 1,000 steps when using only a single version as a target.

width of 10 to produce the most likely paraphrase.

4.4.4. Experiments

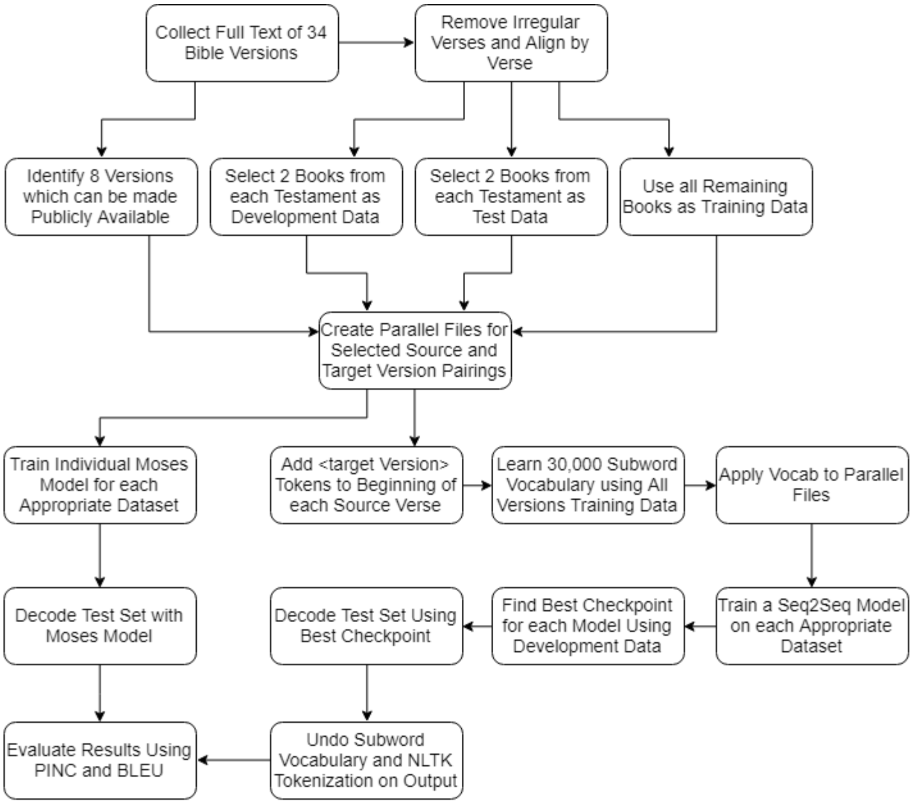

Collect Full Text of 34 Bible Versions
Remove Irregular Verses and Align by Verse
Identify 8 Versions which can be made Publicly Available
Select 2 Books from each Testament as Development Data
Select 2 Books from each Testament as Test Data
Use all Remaining Books as Training Data
Create Parallel Files for Selected Source and Target Version Pairings
Train Individual Moses Model for each Appropriate Dataset
Add <target Version> Tokens to Beginning of each Source Verse
Learn 30,000 Subword Vocabulary using All Versions Training Data
Apply Vocab to Parallel Files
Decode Test Set with Moses Model
Decode Test Set Using Best Checkpoint
Find Best Checkpoint for each Model Using Development Data
Train a Seq2Seq Model on each Appropriate Dataset
Evaluate Results Using PINC and BLEU
Undo Subword Vocabulary and NLTK Tokenization on Output

Figure 4.2: A Diagram of the Experimental Workflow

As indicated above, we Train Moses and Seq2Seq models on a variety of source-target pairings of our Bible Corpus. Our Seq2Seq model is implemented using a publicly available library [19] which itself makes use of the API provided by Tensorflow [1]. See Figure 4.2 for an overview of our entire work process.

The code and data to run the experiments which use only the publicly available portion of our data are available[12].

Results. For each of the single target test sets we identify the Moses model and the Seq2Seq model which achieves the highest BLEU score. The results of the evaluation

[12] https://github.com/keithecarlson/StyleTransferBibleData

metrics for these models' outputs can be seen in Table 4.4. The results of all models and test sets can be found in our github repository.

Test Set	**Best Moses Training**	**Best Sq2Sq Training**	**Moses BLEU**	**Sq2Sq BLEU**	**Moses Pinc**	**Sq2Sq PINC**
KJV → ASV	KJV → ASV	Pub → ASV	**71.16**	65.61	16.84	**24.68**
BBE → ASV	BBE → ASV	Pub → Pub	**31.28**	30.17	47.03	**56.8**
YLT → BBE	YLT → BBE	Pub → Pub	24.01	**24.94**	66.47	**72.51**
Pub → ASV	Pub → ASV	Pub → ASV	**46.87**	46.00	28.67	**42.22**
Pub → BBE	Pub → BBE	Pub → Pub	**30.03**	29.71	56.49	**64.54**

Table 4.4: The BLEU and PINC scores of the best Moses and best Seq2Seq ("Sq2Sq") models for each test set. The best model is defined here as the one which achieves the highest BLEU score on the test set. ("Pub" is Public.)

We note a few trends across all of these results. First, we find that the Seq2Seq model always has a higher PINC score than the Moses model. This matches our qualitative observation that Moses makes changes to the source sentence less aggressively than Seq2Seq and often leaves large portions of the input unchanged. A few examples showing this can be seen in Table 4.5.

Moses always performed best when it was trained on the exact source and target versions that it would eventually be tested on. Seq2Seq however seems to often benefit from additional versions being included. For example, when decoding the YLT → BBE test set the best Moses was trained using the YLT → BBE training set. The best Seq2Seq model, however, was trained using the training set where all public versions were used as both the source and target. As mentioned above, we experimented with training Seq2Seq on single source version and single target version data, such as YLT → BBE, but the results were poor. In this case though the Seq2Seq model trained on public → public also outperformed the one trained on public → BBE and on all → all. It seems the increase in number of training examples going from public → BBE to public → public was helpful for decoding into BBE even when none of the

Source	Moses Output	Seq2Seq Output
And that there may not be any evil liver, or any man without respect for God, like Esau, who let his birthright go for a plate of food.	And that there shall not be any evil liver, or any man without respect unto God, as Esau that let his birthright go for a plate of food.	And the liver shall not sin, nor the transgression of God, like Esau; who let his birthright go for the dish.
His legs were covered with plates of brass and hanging on his back was a javelin of brass.	His legs were covered with flakes of brass and hanged on his shoulder was a javelin of brass.	And he overlaid his legs with plates of brass, and put on his back a spear of brass.

Table 4.5: Examples of Moses and Seq2Seq output which showcase Moses making fewer modifications to the source sentence.

extra examples were targeting BBE. This is similar to results which found that adding additional languages to a multilingual Seq2Seq model could improve performance [75]. This improvement seems to be bounded in our case however, as the public → public model also outperforms the all → all model when decoding YLT → BBE. It is possible that when using all 34 Bible versions, there is simply to much for our formulation of the model to learn. Since it has to be able to produce 34 different styles it may not be able to focus on perfecting each of them. It could also be that the non-public Bible versions have some characteristics which make them ultimately detrimental to use as training for a model which will be evaluated on the YLT → BBE task.

In 4 out of the 5 pairings evaluated Moses outperforms Seq2Seq as measured by BLEU. It is only on the YLT → BBE test set that the best Seq2Seq model achieves a higher BLEU score than the best Moses model. This test set is the most demanding of the model, as the source and target sentences are the least similar as can be seen in Table 4.2. This performance seems related to our earlier observation that Moses is more conservative in making changes to the source. In situations where relatively little modification to the source is required, such as in the KJV → ASV task, the auto-encoding tendencies of Moses can be quite helpful. When more drastic revision is required however, the more aggressive tendencies of Seq2Seq begin to become more effective.

Qualitatively, both models seem to be changing text to better imitate the targeted style. In the first example in Table 4.6 Moses and Seq2Seq correctly use the archaic verb "seemeth" and replace "Lord" with "Jehovah". Both modifications are stylistically correct when when targeting ASV. We notice some cases where Moses seems to be unable to pick up on relatively simple stylistic markers. In the second example of Table 4.6 we see a translation from YLT to BBE. YLT uses quotation marks when someone is speaking, but BBE does not. While the text produced by both Moses and Seq2Seq resembles BBE in style, Moses has not removed the quotation marks which were present in the source. This is despite the fact that the Moses model was only trained with BBE targets and so has never actually seen a training example where it should have produced quotation marks.

4.4.5. Conclusions

Style transfer can naturally be viewed as monolingual machine translation, but lack of an appropriate training corpus has made the direct application of many machine translation models difficult. In this section we have shown that our Bible dataset can help bridge this gap and allow machine translation innovations to be more directly applied to the task of style transfer. A large parallel corpus, such as our Bible dataset, will allow researchers to explore both similarities as well as the nuanced differences between the tasks of style transfer and machine translation. We hope that this data inspires the creation of style-transfer-specific architectures. In our own work demonstrate this with both a statistical and a neural machine translation architecture. We find that these two systems perform similarly on most of the decoding data sets.

Source	**Target**	**Moses Output**	**Seq2Seq Output**
(BBE) Then Samuel gave him an account of everything, keeping nothing back. And he said, It is the Lord; let him do what seems good to him.	(ASV) And Samuel told him every whit, and hid nothing from him. And he said, It is Jehovah: let him do what seemeth him good.	Then Samuel told him of all things not. And he said, It is Jehovah; let him do that which seemeth him good.	And Samuel told all things, and did not hold back. And he said, It is Jehovah; let him do what seemeth good to him.
(YLT) And Jehovah saith, 'Judah doth go up; lo, I have given the land into his hand.'	(BBE) And the Lord said, Judah is to go up: see, I have given the land into his hands.	And the Lord said, 'Judah will go up, see, I have given the land into his hand.'	And the Lord said, Judah will go up; see, I will give the land into his hand.

Table 4.6: Examples which show Moses and Seq2Seq outputs which have correctly made changes to better imitate the target style or which have missed some obvious part of the transfer.

In general Moses performs slightly better, achieving a higher BLEU score on 4 of our 5 evaluations. This superiority is increased when the task requires less modification of the source sentence to match the target. Seq2Seq makes gains on Moses when the task requires more aggressive editing of the source, and is able to outperform Moses on the most demanding of our 5 tests.

It is likely that some previously published modifications to Seq2Seq would result in immediate performance improvements. Candidates from the machine translation literature include: coverage modelling [160] to help track which parts of the source sentence have already been paraphrased and the use of a pointer network [115] to allow copying of words directly from the source sentence. Pointer networks have already been used for style transfer [74], and seem likely to be useful for our multi-style corpus as well.

Section 4.5
Unsupervised Style Transfer in Text

In section 4.4 we demonstrated that our Bible dataset could be used to perform supervised style transfer. In many settings however, parallel data representing the styles we wish to target is not available. In this section we will explore this task of unsupervised style transfer. In addition, while application of state-of-the-art neural machine translation methods may yield near-term improvements in style transfer, we show here that translation models can be improved by considering the differences between the two tasks more explicitly.

4.5.1. Background

The potential applications of holistic machine style transfer lacking parallel data are numerous. For example, various periodicals often try to have a single “voice” and an unsupervised style transfer of the kind studied here would enable a staff writer to produce the content required of an article which was then "stylized" per the requirements of the venue. A style transfer platform could be a high-powered editorial assistant. Such a platform could also assist aspiring writers. All that said, one should not be blind to the more nefarious potential uses successful style transfer machinery which could be useful for spoofing an audience to productive, or unproductive writerly ends [124].

Previous related work using unsupervised training for generating text in a particular style includes the generation of stylized text [70] and modification of the sentiment or formality of prose [58, 94, 95, 149]. More generally, unsupervised machine translation recently has received significant attention. Most of these approaches rely on the idea of back-translation [6, 91] to automatically generate a synthetic parallel from unaligned data. In [89] this concept is also used, along with a novel cross-lingual language model objective for pre-training to achieve impressive performance on the unsupervised translation task.

4.5.2. Data Splitting

Once again our work makes use of the versions of the Bible discussed in section 4.3 (and made available on Github). Each Bible version is treated as representative of a different English writing style. The texts are divided hierarchically (and canonically), into version, book, chapter and verse, so that the verses from different versions are parallel. We do not take advantage of the alignment during the training of our systems, but the alignment does make possible an objective evaluation of our output.

Our major methodological advance is the introduction of another coarse level of hierarchy which we call *content*, which we then supply to our model to help it to learn content and style separately. In the case of the Bible, we use nine "divisions" of the

Bible which are classical groupings of thematically similar texts.[13] See Table 4.7 for the divisions used. We do not use the exact data splits detailed in subsection 4.4.1, but instead split the data as required by the formulation of our models. We use some books of the YLT and BBE versions for validation and testing as the YLT → BBE translation was identified as the "hardest" task in section 4.3. The validation set contains the BBE and YLT versions of 1 Kings, Zephaniah, Mark, and Colossians. The testing set contains the BBE and YLT versions of Judges, 1 Samuel, Philippians, and Hebrews. The remaining books from BBE and YLT and all books from the other six Bible (publicly available) versions makeup the training data.

The parallel texts allow for automatic and objective evaluation of translations. Nevertheless, the models we describe can be generalized to other non-parallel datasets, but in such cases objective evaluation would be more difficult.

Division	**Books**
Pentateuch	Genesis, Exodus, Leviticus, Numbers, Deuteronomy
History	Joshua, Judges, Ruth, 1 Samuel, 2 Samuel, 1 Kings, 2 Kings, 1 Chronicles, 2 Chronicles, Ezra, Nehemiah, Esther
Poetry	Job, Psalms, Proverbs, Ecclesiastes, Song of Solomon
Major Prophets	Isaiah, Jeremiah, Lamentations, Ezekiel, Daniel
Minor Prophets	Hosea, Joel, Amos, Obadiah, Jonah, Micah, Nahum, Habakkuk, Zephaniah, Haggai, Zechariah, Malachi
Gospels & Acts	Matthew, Mark, Luke, John, Acts
(Pauline) Epistles	Romans, 1 Corinthians, 2 Corinthians, Galatians, Ephesians, Philippians, Colossians, 1 Thessalonians, 2 Thessalonians, 1 Timothy, 2 Timothy, Titus, Philemon, Hebrews
General Epistles	James, 1 Peter, 2 Peter, 1 John, 2 John, 3 John, Jude
Revelation	Revelation

Table 4.7: Our partition of Bible books into divisions.

[13] There is no authoritative partition into divisions, but there are many similar varieties. Our choice among these options is somewhat arbitrary, but supported. An example of Old Testament divisions which match ours can found at `http://www.scriptureman.com/ot.gif` and our New Testament at `http://jpatton.bellevue.edu/inspired-table2.jpg`

4.5.3. Baseline System

In [89] a method is introduced for cross-lingual language model pretraining from non-parallel data[14]. Their model, XLM, feeds token, position, and language embeddings to a Transformer model [163] which tries to predict masked words. This task, Masked Language Modeling (MLM), was introduced by [44]. They then demonstrate unsupervised translation as an application of these pretrained language models. We use the XLM as our baseline.

In our experiment, we treat each version of the Bible in the data as a language. So the embeddings fed to the Transformer for MLM training are position and token embeddings as before, and version embeddings replacing the language embeddings of the original system.

We train the language model until the perplexity of the validation data for the BBE → YLT version has stopped improving. We then use this pretrained language model to initialize our machine translation and train on the task of unsupervised translation until the BLEU score of the validation data for the BBE → YLT task has stopped improving. We call these models "XLM".

4.5.4. Model with Content Embeddings

Using Bible divisions as a grouping of content similarity, we modify the XLM embedding structure accordingly and include a *content embedding* in addition to the token, position, and language (style) embeddings. In a different context other considerations or structural organization may suggest a different articulation of content. This additional embedding is treated similarly to the three embeddings in the baseline system. The input of each token passed to the Transformer is the combination of four embeddings instead of three. Just as in the XLM, these embeddings are updated during the training process. Our intuition is that for some datasets, the model may have

[14] Code found at: `https://github.com/facebookresearch/XLM`

difficulty distinguishing whether differences in language arise because of differences in the style of writing, or differences in the content. By providing training data where both style and content are designated, we anticipate that the model will be better able to reproduce the differences which are style-specific. Similar intuition has led to other approaches which allow a model to learn style and content separately [54, 181].

In this new formulation, we provide all four embeddings to the Transformer and then train towards the MLM objective as before. We call this model "XLM + Content" (see Figure 4.3). As in the "XLM" model, we stop training of the language model when the perplexity of BBE → YLT evaluation task has stopped improving and stop the translation training when the BLEU score of the evaluation data BBE → YLT has stopped improving. Note that the alignment (parallel nature of the texts) makes possible the BLEU scoring.

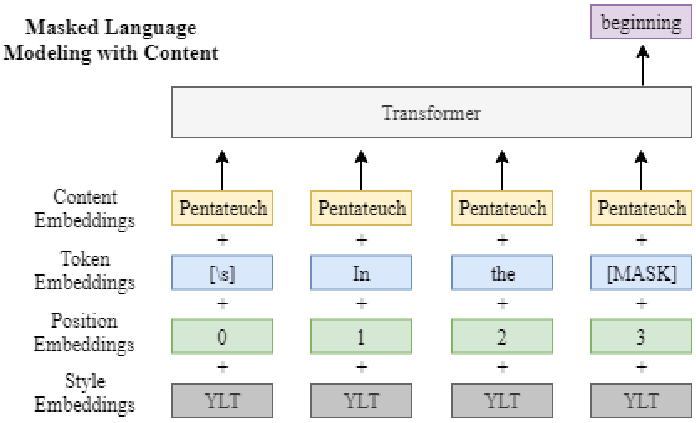


Figure 4.3: "XLM + Content" model training on the MLM objective. Based on Figure 1 of [89].

4.5.5. Results

After training both models as described, we evaluate the outputs of our test set. The existence of parallel texts allows us to evaluate our results using the standard translation quality measures BLEU [129] and PINC [32], which reward similarity to the target and dissimilarity to the source respectively. These scores for each model, as well as the unmodified source compared to the target, can be seen in 4.8.

Test Book	**Source**	**XLM**	**XLM+Content**
Judges	16.18(0)	**26.5(39.93)**	26.1(44.89)
1 Samuel	14.75(0)	24.21(39.72)	**24.36(44.40)**
Philippians	18.29(0)	20.56(25.50)	**22.82(29.83)**
Hebrews	12.27(0)	15.88(29.73)	**17.44(34.70)**

Table 4.8: The BLEU (PINC) scores of the unmodified source and the output of each model for each test book. All scores are when translating from Bible version YLT to Bible version BBE.

We find that our model with content embeddings has a higher PINC score for all four test books, indicating that it has more aggressively made changes than the baseline system. "XLM + Content" also attains a sizeably Higher BLEU score on Philippians and Hebrews. The BLEU score for the other two test books are similar between the two systems.

As discussed in section 4.2, BLEU and PINC have often been used for evaluating style transfer, but these metrics have received some criticism for the task. We introduced the idea of a *frequent idiosyncratic word* test as an additional form of validation. For this experiment this form of bespoke evaluation checks to see if 17 frequent words with known translations have been correctly translated in the YLT → BBE test task. These words can be seen in Table 4.9. All the words occur frequently and exclusively in YLT. Examples include *unto*, *hath*, *flee*, *doth.* These words occur 2,522 times in YLT source lines in the test set. In this test, a YLT → BBE translation is counted as correct if the BBE version does *not* include the idiosyncratic word from the YLT line.

Accuracy scores in this evaluation increase with the complexity of the model: 99.3% ("XLM") and 99.8% ("XLM + Content").

Table 4.10 shows two test data example inputs and their targets alongside the

YLT Exclusive Words
unto, flee, fleeth, hath, thine, hast, thus, midst, thy, inheritance, cometh, ye, also, shall, doth, thou, jehovah

Table 4.9: Words which are frequent in YLT but do not appear in BBE

corresponding outputs of our systems. In the first example, note that both outputs correctly remove the use of quotation marks as is consistent with the BBE target and modernize the archaic *Thou* and *dost*. The “XLM + Content" also correctly changes the word *testify* to *witness*. In the second example, the “XLM + Content” model correctly changes *age-during* to *eternal*.

Source(YLT)	**Target(BBE)**	**XLM**	**XLM + Content**
6 and one in a certain place did testify fully, saying, ‘What is man, that Thou art mindful of him, or a son of man, that Thou dost look after him?	6 But a certain writer has given his witness, saying, What is man, that you keep him in mind? what is the son of man, that you take him into account?	6 And one in a certain place did testify fully, saying, What is man, that you are mindful of him, or a son of man, that you do look after him?	6 And one in a certain place did give witness fully, saying, What is man, that you are mindful of him, or a son of man, that you will go after him?
9 and having been made perfect, he did become to all those obeying him a cause of salvation age-during,	9 And when he had been made complete, he became the giver of eternal salvation to all those who are under his orders;	9 And having been made perfect, he did become to all those who obey him a cause of salvation age-during,	9 And having been made perfect, he gave to all those who keep him a cause of salvation eternal,

Table 4.10: Examples Outputs of each of the systems with YLT source and BBE target (validation). The verses are Hebrews 2:6 and Hebrews 5:9.

4.5.6. Conclusions

The task of holistic textual style transfer requires a system take text in a native (source) style as input and then rewrite the text, retaining the meaning while changing the style consistent with a specified target. In many potential applications this task will need to be performed in contexts where there is no parallel data capturing the styles of interest available for training. Examples range from the journalistic (writing articles in a given editorial style) to the literary (writing the style or voice of a given author). Contexts such as these have large corpora of source and target examples, but – presumably – no source/target pairings.

In this work we demonstrated that a modern unsupervised machine translation technique could be applied to unsupervised holistic textual style transfer in the context of different styles (well known and publicly available versions) of the Bible. We show that by adding an additional “content embedding layer" to encode the type of content in text, holistic style transfer is improved. The parallel nature of Bible versions enables us to measure objectively the effect of our innovation of content embedding and improvement is witnessed in terms PINC and BLEU scores that are greater when using content embedding than without. Specifically, this improves upon the work of section 4.4 and makes use of our introduced, publicly accessible data. We further introduce a simple test of frequent idiosyncratic words as a measure of style transfer quality. This too supports the claim that content embedding improves style transfer.

Section 4.6 Conclusion

In this chapter we have presented our work related to the important problem of holistic style transfer. To that end, an important contribution to the community and the literature is the making available of a cleaned and (a priori) aligned collection of versions of the Bible. Future work will need to identify additional datasets that are suitable for research on this task. In particular, having some diversity of parallel corpora for testing style transfer would be of great interest. The structure of the Bible suggests a division of text into specific types of content (which we readily adopt), but other contexts may require a different approach to content labeling and embedding.

Our work highlights the utility of the Bible as a dataset for holistic style transfer, demonstrates that unsupervised machine translation methods for holistic style transfer are possible and can be objectively evaluated, and provides further evidence – and an actionable methodology – for the idea that learning content independent of style can be beneficial.

Chapter 5

Discussion

In this thesis we explored several dimensions of the computational study of writing style. As discussed in the introduction, it is not only the content of text which is important, but also the style in which it is written. While this has always been true, an increase in the amount of our communication which occurs through text due to the continually increasing use of the internet only amplifies the importance of style. Even more, as computation interacts increasingly deeply with the analysis and generation of text, quantitative approaches to textual style present an ongoing and important area of research.

In chapter 2, we examined how the quantitative analysis of the style of human produced text – stylometry – can lead to new insights. We did this in the context of written decisions of the United States Supreme Court. Through styolmetric analyses we found that by some measures the writing of the court is becoming both less complex and less friendly over time. Our main result is a different longitudinal stylometric study that relied on the use of a list of function words as a stylistic fingerprint and produced evidence that clerks (working with individual Justices) appear to be having an increasingly large role in the writing of official decisions.[1]

[1] This work is the foundation of the published paper [25]. Related publications not discussed in this thesis include [29], [26], and [23].

Next, in chapter 3, we tackle the problem of generating novel text which targets a specific style of writing. Here, our work focuses on the space of expertly written online wine reviews. We collect and clean reviews from winemag.com from 1999-2016, leaving us with a corpus of 201,431 reviews and associated metadata about the wine being reviewed, rating given to the wine, and review author. Using a novel approach of treating this metadata as language, albeit not very natural language, we train a modern neural machine translation architecture to produce reviews from metadata. While the details of a particular wine are kept constant, we are able to target specific styles in the produced reviews by altering the rating and author. We then show that these reviews are of high enough quality that 501 survey participants are unable to distinguish between human or machine produced reviews at a statistically significant level.[2]

In chapter 4 our task is style transfer, which requires a system to take text and a target style as input and output new text which retains the meaning of the original but which matches the target style. For this work we identify an unused but ideal corpus of Bible versions. Bible versions have many advantages as a corpus for this task, including their preexisting alignment from standardized books, chapters, and verses and their wide range of styles due to the sheer number and differing objectives of translations.

We first employ this corpus for a supervised version of the style transfer task. We use machine translation systems and compare a statistical machine translation approach to a neural based one. We find that both models produce text which is closer to the targeted style than the original was and verify this both qualitatively and through automatic evaluation metrics.[3]

Finally, in section 4.5 we undertake unsupervised style transfer. We employ our

[2] This work is currently under journal review.

[3] This work forms the basis of the publication [27].

newly identified Bible dataset as before, but treat the data here as unaligned. We first apply a recent unsupervised neural machine translation architecture to the task with minimal changes needed to match our data to the model. Then, we show that modifying this architecture to take advantage of the differences between style transfer and machine translation (in our case by adding separate embeddings for content and style) can improve the outputs.

In addition, we propose new evaluation of systems which produce style-targeted text. Here, we believe that a suite of traditional stylometric measures can be applied. The key idea is that systems which produce output which matches the targeted style by all of the ways that researchers have measured style in the past should be considered to be doing a good job at the task of style transfer. With this idea in mind, we show that our modified unsupervised style transfer model with new embeddings, better matches the idiosyncratic language of the target style than the unmodified machine translation model.[4]

In conclusion, we have argued that the writing style, and not only the content, of text is important. We have demonstrated that consideration of style can lead to new insights by examining the writings of the U.S. Supreme Court. Using wine reviews we have shown that with the correct input, formatting, and training neural networks can be used to produce text of a particular style which people are unable to distinguish from human written text. We identified a style in text dataset of Bible versions which alleviates many problems common in existing style datasets. We then show with this data that machine translation systems can be applied to the task of style transfer in both the supervised and unsupervised setting. Furthermore, we demonstrate in the unsupervised setting that the direct application of machine translation systems can be improved by considering the nuanced differences between the two tasks. Finally,

[4] This work is being prepared for journal submission.

we propose a new paradigm for the evaluation of tasks which target a particular style inspired by classic stylometry.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015, Software available from tensorflow.org.

[2] Lothar Afflerbach, *Criteria for the assessment of random number generators*, Journal of Computational and Applied Mathematics **31** (1990), no. 1, 3–10.

[3] Adrian Akmajian, Ann K Farmer, Lee Bickmore, Richard A Demers, and Robert M Harnish, *Linguistics: An Introduction to Language and Communication*, MIT Press, Cambridge, MA.

[4] Olufunmilayo B Arewa, *Open access in a closed universe: Lexis, Westlaw, law schools, and the legal information market*, Lewis & Clark L. Rev. **10** (2006), 797.

[5] Shlomo Argamon, *Interpreting Burrows's Delta: Geometric and probabilistic foundations*, Literary and Linguistic Computing **23** (2008), no. 2, 131–147.

[6] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho, *Unsupervised neural machine translation*, arXiv preprint arXiv:1710.11041 (2017).

[7] Robert H Ashton, *Is there consensus among wine quality ratings of prominent critics? An empirical analysis of red Bordeaux, 2004-2010*, Journal of Wine Economics **8** (2013), no. 2, 225.

[8] ———, *Dimensions of expertise in wine evaluation*, Journal of Wine Economics **12** (2017), no. 1, 59.

[9] Robert H Ashton et al., *Reliability and consensus of experienced wine judges: Expertise within and between*, Journal of Wine Economics **7** (2012), no. 1, 70–87.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, *Neural machine translation by jointly learning to align and translate*, CoRR **abs/1409.0473** (2014).

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, *Neural machine translation by jointly learning to align and translate*, arXiv preprint arXiv:1409.0473 (2014).

[12] Silke Bambauer-Sachse and Sabrina Mangold, *Brand equity dilution through negative online word-of-mouth communication*, Journal of Retailing and Consumer Services **18** (2011), no. 1, 38–45.

[13] Suman Basuroy, Subimal Chatterjee, and S Abraham Ravid, *How critical are critical reviews? The box office effects of film critics, star power, and budgets*, Journal of Marketing **67** (2003), no. 4, 103–117.

[14] Jonah Berger, Alan T Sorensen, and Scott J Rasmussen, *Positive effects of negative publicity: When negative reviews increase sales*, Marketing Science **29** (2010), no. 5, 815–827.

[15] Ryan C. Black and James F. Spriggs, *An empirical analysis of the length of US Supreme Court opinions*, Hous. L. Rev. **45** (2008), 621.

[16] Christina L Boyd, David A Hoffman, Zoran Obradovic, and Kosta Ristovski, *Building a taxonomy of litigation: Clusters of causes of action in Federal complaints*, Journal of Empirical Legal Studies **10** (2013), no. 2, 253–287.

[17] Ryan L. Boyd and James W. Pennebaker, *Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis*, Psychological Science **26** (2015), no. 5, 570–582.

[18] Charles Henry Brase and Corrinne Pellillo Brase, *Understandable Statistics*, Cengage Learning, 2014.

[19] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le, *Massive exploration of neural machine translation architectures*, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Copenhagen, Denmark), Association for Computational Linguistics, September 2017, pp. 1442–1451.

[20] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin, *A statistical approach to machine translation*, Computational Linguistics **16** (1990), no. 2.

[21] Joachim Büschken and Greg M Allenby, *Sentence-based text analysis for customer reviews*, Marketing Science **35** (2016), no. 6, 953–975.

[22] Nuno Camacho, Martijn De Jong, and Stefan Stremersch, *The effect of customer empowerment on adherence to expert advice*, International Journal of Research in Marketing **31** (2014), no. 3, 293–308.

[23] Keith Carlson, Faraz Dadgostari, Michael A. Livermore, and Daniel N. Rockmore, *A multinetwork and machine learning examination of structure and content in the United States Code*, Frontiers in Physics **8** (2021), 676.

[24] Keith Carlson, Praveen K. Kopalle, Allen Riddell, Daniel Rockmore, and Prasad Vana, *Complementing human effort in online reviews: A deep learning approach to automatic content generation*, Under review at time of writing (2021).

[25] Keith Carlson, Michael A Livermore, and Daniel Rockmore, *A quantitative analysis of writing style on the US Supreme Court*, Washington University Law Review **93** (2016), no. 6, 1461–1510.

[26] Keith Carlson, Michael A. Livermore, and Daniel N. Rockmore, *The problem of data bias in the pool of published U.S. Appellate Court opinions*, Journal of Empirical Legal Studies **17** (2020), no. 2, 224–261.

[27] Keith Carlson, Allen Riddell, and Daniel Rockmore, *Evaluating prose style transfer with the Bible*, Royal Society Open Science **5** (2018), no. 10, 171920.

[28] Keith Carlson, Allen Riddell, and Daniel N Rockmore, *Unsupervised text style transfer with content embeddings*, Under review at time of writing.

[29] Keith Carlson, Daniel N. Rockmore, Allen Riddell, Jon Ashley, and Michael A. Livermore, *Style and substance on the US Supreme Court*, Law as Data, SFI Press, 2019, p. 83–115.

[30] I. Chakraborty, M. Kim, and K. Sudhir, *Attribute sentiment scoring with online text reviews: Accounting for language structure and attribute self-selection*,

Cowles Foundation discussion paper, Cowles Foundation for Research in Economics, Yale University, 2019.

[31] Isabella M Chaney, *A comparative analysis of wine reviews*, British Food Journal **102** (2000), no. 7, 470–480.

[32] David Chen and William Dolan, *Collecting highly parallel data for paraphrase evaluation*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Portland, Oregon, USA), Association for Computational Linguistics, June 2011, pp. 190–200.

[33] Judith A Chevalier, Yaniv Dover, and Dina Mayzlin, *Channels of impact: User reviews when quality is dynamic and managers respond*, Marketing Science **37** (2018), no. 5, 688–709.

[34] Judith A Chevalier and Dina Mayzlin, *The effect of word of mouth on sales: Online book reviews*, Journal of Marketing Research **43** (2006), no. 3, 345–354.

[35] Pradeep K Chintagunta, Shyam Gopinath, and Sriram Venkataraman, *The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets*, Marketing Science **29** (2010), no. 5, 944–957.

[36] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, *Learning phrase representations using RNN encoder-decoder for statistical machine translation*, CoRR **abs/1406.1078** (2014).

[37] Stephen J Choi and G Mitu Gulati, *Which judges write their opinions (and should we care)*, Fla. St. UL Rev. **32** (2004), 1077.

[38] Gregory C Chow, *Tests of equality between sets of coefficients in two linear regressions*, Econometrica **28** (1960), no. 3, 591–605.

[39] John M Conroy and Dianne P O'leary, *Text summarization via hidden Markov models*, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2001, pp. 406–407.

[40] William Coster and David Kauchak, *Simple English Wikipedia: A new text simplification task*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, Association for Computational Linguistics, 2011, pp. 665–669.

[41] Laurence Danlos, *The Linguistic Basis of Text Generation*, Cambridge University Press, Cambridge, UK, 2009.

[42] Lukas Danner, Trent E Johnson, Renata Ristic, Herbert L Meiselman, and Susan EP Bastian, *"I like the sound of that!" Wine descriptions influence consumers' expectations, liking, emotions and willingness to pay for Australian white wines*, Food Research International **99** (2017), 263–274.

[43] Reinhold Decker and Michael Trusov, *Estimating aggregate consumer preferences from online product reviews*, International Journal of Research in Marketing **27** (2010), no. 4, 293–307.

[44] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Minneapolis, Minnesota), Association for Computational Linguistics, June 2019, pp. 4171–4186.

[45] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang, *Multi-task learning for multiple language translation*, ACL, 2015.

[46] Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu, *Learning to generate product reviews from attributes*, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, vol. 1, 2017, pp. 623–632.

[47] Wenjing Duan, Bin Gu, and Andrew B Whinston, *Do online reviews matter?—An empirical investigation of panel data*, Decision Support Systems **45** (2008), no. 4, 1007–1016.

[48] Angela Fan, Mike Lewis, and Yann Dauphin, *Hierarchical neural story generation*, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Melbourne, Australia), Association for Computational Linguistics, July 2018, pp. 889–898.

[49] Xing Fang and Justin Zhan, *Sentiment analysis using product review data*, Journal of Big Data **2** (2015), no. 1, 5.

[50] Jessica Ficler and Yoav Goldberg, *Controlling linguistic style aspects in neural language generation*, Proceedings of the Workshop on Stylistic Variation (Copenhagen, Denmark), Association for Computational Linguistics, September 2017, pp. 94–104.

[51] Orhan Firat, KyungHyun Cho, and Yoshua Bengio, *Multi-way, multilingual neural machine translation with a shared attention mechanism*, CoRR **abs/1601.01073** (2016).

[52] Nicholas J Foti, James M Hughes, and Daniel N Rockmore, *Nonparametric sparsification of complex multiscale networks*, PloS One **6** (2011), no. 2, e16431.

[53] Richard Friberg and Erik Grönqvist, *Do expert reviews affect the demand for wine?*, American Economic Journal: Applied Economics **4** (2012), no. 1, 193–211.

[54] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan, *Style transfer in text: Exploration and evaluation*, arXiv preprint arXiv:1711.06861 (2017).

[55] Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight Knight, *Generating topical poetry*, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (Copenhagen, Denmark), Association for Computing Machinery, September 2016, pp. 1183–1191.

[56] Anindya Ghose, Panagiotis G Ipeirotis, and Beibei Li, *Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content*, Marketing Science **31** (2012), no. 3, 493–520.

[57] David Godes and José C Silva, *Sequential and temporal dynamics of online opinion*, Marketing Science **31** (2012), no. 3, 448–473.

[58] Hongyu Gong, Suma Bhat, Lingfei Wu, Jinjun Xiong, and Wen Mei Hwu, *Reinforcement learning based text style transfer without parallel training corpus*, 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019, Association for Computational Linguistics (ACL), 2019, pp. 3168–3180.

[59] Alex Graves, *Generating sequences with recurrent neural networks*, CoRR **abs/1308.0850** (2013).

[60] Deborah H Gruenfeld, *Status, ideology, and integrative complexity on the U.S. Supreme Court: Rethinking the politics of political decision making*, Journal of Personality and Social Psychology **68** (1995), no. 1, 5.

[61] Cathy Hackl, *3 new ways artificial intelligence is powering the future of marketing*, (2020), Available at https://www.forbes.com/sites/cathyhackl/2020/06/28/3-new-ways-artificial-intelligence-is-powering-the-future-of-marketing/?sh=f77577b1a96e.

[62] Bruce E Hansen, *The new econometrics of structural change: Dating breaks in US labour productivity*, Journal of Economic Perspectives **15** (2001), no. 4, 117–128.

[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Deep residual learning for image recognition*, CoRR **abs/1512.03385** (2015).

[64] Xiaodong He and Li Deng, *Deep learning for image-to-text generation: A technical overview*, IEEE Signal Processing Magazine **34** (2017), no. 6, 109–116.

[65] Kenneth Heafield, *Kenlm: Faster and smaller language model queries*, Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2011, pp. 187–197.

[66] Sepp Hochreiter and Jürgen Schmidhuber, *Long short-term memory*, Neural Computation **9** (1997), no. 8, 1735–1780.

[67] Robert T Hodgson, *How expert are "expert" wine judges?*, Journal of Wine Economics **4** (2009), no. 2, 233–241.

[68] Robert T Hodgson et al., *An examination of judge reliability at a major US wine competition*, Journal of Wine Economics **3** (2008), no. 2, 105–113.

[69] Minqing Hu and Bing Liu, *Mining and summarizing customer reviews*, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.

[70] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing, *Toward controlled generation of text*, International Conference on Machine Learning, 2017, pp. 1587–1596.

[71] Ming-Hui Huang and Roland T Rust, *Artificial Intelligence in service*, Journal of Service Research **21** (2018), no. 2, 155–172.

[72] James M Hughes, Nicholas J Foti, David C Krakauer, and Daniel N Rockmore, *Quantitative patterns of stylistic influence in the evolution of literature*, Proceedings of the National Academy of Sciences **109** (2012), no. 20, 7682–7686.

[73] William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu, *Aligning sentences from standard Wikipedia to simple Wikipedia.*, HLT-NAACL, 2015, pp. 211–217.

[74] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg, *Shakespearizing modern language using copy-enriched sequence-to-sequence models*, arXiv preprint arXiv:1707.01161 (2017).

[75] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean, *Google's multilingual neural machine translation system: Enabling zero-shot translation*, CoRR **abs/1611.04558** (2016).

[76] Stephen M Johnson, *The changing discourse of the Supreme Court*, UNHL Rev. **12** (2014), 29.

[77] Jad Kabbara and Jackie Chi Kit Cheung, *Stylistic transfer in natural language generation systems using recurrent neural networks*, EMNLP 2016 (2016), 43.

[78] Tomoyuki Kajiwara and Mamoru Komachi, *Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings*, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 1147–1158.

[79] PK Kannan et al., *Digital marketing: A framework, review and research agenda*, International Journal of Research in Marketing **34** (2017), no. 1, 22–45.

[80] Daniel Martin Katz, Michael James Bommarito, Julie Seaman, Adam Candeub, and Eugene Agichtein, *Legal n-grams? A simple approach to track the 'evolution' of legal language*, Proceedings of JURIX 2011: The 24th International Conference on Legal Knowledge and Information Systems, Vienna, 2011, 2011.

[81] Martin Kay, *Functional unification grammar: A formalism for machine translation*, Proceedings of the 10th International Conference on Computational Linguistics, Association for Computational Linguistics, 1984, pp. 75–78.

[82] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom, *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*, Tech. report, Naval Technical Training Command Millington TN Research Branch, 1975.

[83] Jonathan W King and Marta Kutas, *A brain potential whose latency indexes the length and frequency of words*, CRL Newsletter **10** (1995), no. 2, 1–9.

[84] Diederik P. Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, CoRR **abs/1412.6980** (2014).

[85] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al., *Moses: Open source toolkit for statistical machine translation*, Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, Association for Computational Linguistics, 2007, pp. 177–180.

[86] Montgomery N Kosma, *Measuring the influence of Supreme Court justices*, The Journal of Legal Studies **27** (1998), no. 2, 333–372.

[87] Anthony T Kronman, *The Lost Lawyer: Failing Ideals of the Legal Profession*, Harvard University Press, Cambridge, MA, 1993.

[88] Max Kuhn and Kjell Johnson, *Applied Predictive Modeling*, vol. 26, Springer, 2013.

[89] Guillaume Lample and Alexis Conneau, *Cross-lingual language model pretraining*, Advances in Neural Information Processing Systems (NeurIPS) (2019).

[90] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato, *Unsupervised machine translation using monolingual corpora only*, ICLR Proceedings (2018).

[91] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato, *Phrase-based & neural unsupervised machine translation*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Brussels, Belgium), Association for Computational Linguistics, October-November 2018, pp. 5039–5049.

[92] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Y-Lan Boureau, et al., *Multiple-attribute text rewriting*, ICLR (2018).

[93] Noam Lapidot-Lefler and Azy Barak, *Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition*, Computers in Human Behavior **28** (2012), no. 2, 434–443.

[94] Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun, *Domain adaptive text style transfer*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3295–3304.

[95] Juncen Li, Robin Jia, He He, and Percy Liang, *Delete, retrieve, generate: a simple approach to sentiment and style transfer*, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (New Orleans, Louisiana), Association for Computational Linguistics, June 2018, pp. 1865–1874.

[96] Bing Liu et al., *Sentiment analysis and subjectivity.*, Handbook of Natural Language Processing **2** (2010), no. 2010, 627–666.

[97] Bing Liu and Minqing Hu, *Opinion mining, sentiment analysis, and opinion spam detection*, Dosegljivo: https://www. cs. uic. edu/ liub/FBS/sentiment-analysis. html# lexicon.[Dostopano 15. 2. 2016] (2004).

[98] Yong Liu, *Word of mouth for movies: Its dynamics and impact on box office revenue*, Journal of Marketing **70** (2006), no. 3, 74–89.

[99] Michael Livermore and Daniel Rockmore (eds.), *Law as Data: Computation, Text, and the Future of Legal Analysis*, Santa Fe, NM, Santa Fe Institute Press, 2019.

[100] Michael A Livermore, Beling Peter, Keith Carlson, Guim Mauricio, and Daniel Rockmore, *Law search in the age of the algorithm*, Michigan State Law Review (To appear).

[101] Lance N Long and William F Christensen, *When justices (subconsciously) attack: The theory of argumentative threat and the supreme court*, Or. L. Rev. **91** (2012), 933.

[102] Xueming Luo, Siliang Tong, Zheng Fang, and Zhe Qu, *Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases*, Marketing Science **38** (2019), no. 6, 937–947.

[103] Xueming Luo, Jie Jennifer Zhang, Bin Gu, and Chee Phang, *Expert blogs and consumer perceptions of competing brands*, MIS quarterly **41** (2013), no. 2, 371–395.

[104] W Lutoslowski, *The Origin and Growth of Plato's Logic*, Longmans, Green, 1897.

[105] Liye Ma and Baohong Sun, *Machine learning and AI in marketing – Connecting computing power to human insights*, International Journal of Research in Marketing **37** (2020), no. 3, 481–504.

[106] William C Mann and Christian MIM Matthiessen, *Nigel: A systemic grammar for text generation.*, Tech. report, University of Southern California Marina Del Rey Information Sciences Inst., 1983.

[107] Philippe Masset, Jean-Philippe Weisskopf, and Mathieu Cossutta, *Wine tasters, ratings, and en primeur prices*, Journal of Wine Economics **10** (2015), no. 1, 75–107.

[108] David R Mayhew, *Realignment: The theory that changed the way we think about American politics*, Journal of Interdisciplinary History **35** (2004), no. 2, 321–322.

[109] Julian McAuley and Jure Leskovec, *Hidden factors and hidden topics: understanding rating dimensions with review text*, Proceedings of the 7th ACM conference on Recommender systems, ACM, 2013, pp. 165–172.

[110] Julian McAuley, Jure Leskovec, and Dan Jurafsky, *Learning attitudes and attributes from multi-aspect reviews*, 2012 IEEE 12th International Conference on Data Mining, IEEE, 2012, pp. 1020–1025.

[111] Neil McIntyre and Mirella Lapata, *Learning to tell tales: A data-driven approach to story generation*, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, Association for Computational Linguistics, 2009, pp. 217–225.

[112] Kathleen McKeown, *Text Generation*, Cambridge University Press, London, 1992.

[113] Thomas Corwin Mendenhall, *The characteristic curves of composition*, Science **9** (1887), no. 214, 237–249.

[114] Lise Menn, *Cross-language data and theories of agrammatism*, Agrammatic Aphasia: A Cross-Language Narrative Sourcebook **2** (1990), 1369–1389.

[115] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher, *Pointer sentinel mixture models*, arXiv preprint arXiv:1609.07843 (2016).

[116] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig, *Linguistic regularities in continuous space word representations*, Proceedings of the 2013 conference of

the north american chapter of the association for computational linguistics: Human language technologies, 2013, pp. 746–751.

[117] Wendy W Moe and David A Schweidel, *Online product opinions: Incidence, evaluation, and evolution*, Marketing Science **31** (2012), no. 3, 372–386.

[118] Wendy W Moe and Michael Trusov, *The value of social dynamics in online product ratings forums*, Journal of Marketing Research **48** (2011), no. 3, 444–456.

[119] Sangkil Moon and Wagner A Kamakura, *A picture is worth a thousand words: Translating product reviews into a product positioning map*, International Journal of Research in Marketing **34** (2017), no. 1, 265–285.

[120] Frederick Mosteller and David L Wallace, *Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers*, Journal of the American Statistical Association **58** (1963), no. 302, 275–309.

[121] ———, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, MA, 1964.

[122] Simone Mueller, Larry Lockshin, Yaelle Saltman, and Jason Blanford, *Message on a bottle: The relative influence of wine back label information on wine choice*, Food Quality and Preference **21** (2010), no. 1, 22–32.

[123] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang, *Abstractive text summarization using sequence-to-sequence RNNs and beyond*, Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning (Berlin, Germany), Association for Computational Linguistics, August 2016, pp. 280–290.

[124] Editors Nature, *Next chapter in artificial writing*, Nat. Mach. Intell. (2020), 419.

[125] Oded Netzer, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko, *Mine your own business: Market-structure surveillance through text mining*, Marketing Science **31** (2012), no. 3, 521–543.

[126] Chester A Newland, *Personal assistants to Supreme Court Justices: The law clerks*, Or. L. Rev. **40** (1960), 299.

[127] Franz Josef Och, *Minimum error rate training in statistical machine translation*, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, Association for Computational Linguistics, 2003, pp. 160–167.

[128] Richard Pacelle, *The Transformation of the Supreme Court's Agenda: From the New Deal to the Reagan Administration*, Routledge, 2019.

[129] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, *Bleu: a method for automatic evaluation of machine translation*, Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[130] Wendy V Parr, *Exploring the nature of wine expertise*, Australian and New Zealand Wine Industry Journal **17** (2002), no. 1, 32–36.

[131] Joann Peck and Terry L Childers, *To have and to hold: The influence of haptic information on product judgments*, Journal of Marketing **67** (2003), no. 2, 35–48.

[132] Todd C Peppers, *Courtiers of the marble palace: The rise and influence of the Supreme Court law clerk*, Stanford University Press, Stanford, CA, 2006.

[133] Pierre Perron, *Dealing with structural breaks, in "palgrave handbook of econometrics", vol. 1: Econometric theory, k. patterson and tc mills*, Palgrave Macmillan **278** (2006), 352.

[134] Richard A Posner, *Law and literature: A relation reargued*, Virginia Law Review (1986), 1351–1392.

[135] ______, *Judges' writing styles (and do they matter)*, U. Chi. L. Rev. **62** (1995), 1421.

[136] Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek V. Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri, *Neural paraphrase generation with stacked residual LSTM networks*, CoRR **abs/1610.03098** (2016).

[137] Davide Proserpio and Georgios Zervas, *Online reputation management: Estimating the impact of management responses on consumer reviews*, Marketing Science **36** (2017), no. 5, 645–665.

[138] Chris Quirk, Chris Brockett, and William B Dolan, *Monolingual machine translation for paraphrase generation.*, EMNLP, 2004, pp. 142–149.

[139] Alejandro Ramos-Soto, Alberto Jose Bugarin, Senén Barro, and Juan Taboada, *Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data*, IEEE Transactions on Fuzzy Systems **23** (2015), no. 1, 44–57.

[140] Sudha Rao and Joel Tetreault, *Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer*, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long

Papers) (New Orleans, Louisiana), Association for Computational Linguistics, June 2018, pp. 129–140.

[141] William H Rehnquist, *Who writes decisions of the Supreme Court*, Brief **53** (1957), 89.

[142] ______, *Another view: Clerks might 'influence' some actions*, US News and World Report (1958).

[143] David A Reinstein and Christopher M Snyder, *The influence of expert reviews on consumer demand for experience goods: A case study of movie critics*, The Journal of Industrial Economics **53** (2005), no. 1, 27–51.

[144] William D Rogers, *Do law clerks wield power in Supreme Court cases*, Brief **53** (1957), 182.

[145] Jeffrey S Rosenthal and Albert H Yoon, *Detecting multiple authorship of united states supreme court legal decisions using function words*, The Annals of Applied Statistics (2011), 283–308.

[146] Mike Schuster and Kuldip K Paliwal, *Bidirectional recurrent neural networks*, IEEE Transactions on Signal Processing **45** (1997), no. 11, 2673–2681.

[147] Rico Sennrich, Barry Haddow, and Alexandra Birch, *Neural machine translation of rare words with subword units*, CoRR **abs/1508.07909** (2015).

[148] Rico Sennrich, Barry Haddow, and Alexandra Birch, *Controlling politeness in neural machine translation via side constraints.*, HLT-NAACL, 2016, pp. 35–40.

[149] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola, *Style transfer from non-parallel text by cross-alignment*, Advances in Neural Information Processing Systems, 2017, pp. 6833–6844.

[150] Michael Siegrist and Marie-Eve Cousin, *Expectations influence sensory experience in a wine tasting*, Appetite **52** (2009), no. 3, 762–765.

[151] Marina Solnyshkina, Radif Zamaletdinov, Ludmila Gorodetskaya, and Azat Gabitov, *Evaluating text complexity and Flesch-Kincaid grade level*, Journal of Social Studies Education Research **8** (2017), no. 3, 238–248.

[152] Lucia Specia, *Translating from complex to simplified sentences*, Computational Processing of the Portuguese Language (2010), 30–39.

[153] Eric T Stuen, Jon R Miller, and Robert W Stone, *An analysis of wine critic consensus: A study of washington and california wines*, Journal of Wine Economics **10** (2015), no. 1, 47.

[154] John Suler, *The online disinhibition effect*, Cyberpsychology & Behavior **7** (2004), no. 3, 321–326.

[155] Monic Sun, *How does the variance of product ratings matter?*, Management Science **58** (2012), no. 4, 696–707.

[156] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, *Sequence to sequence learning with neural networks*, CoRR **abs/1409.3215** (2014).

[157] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, *Sequence to sequence learning with neural networks*, Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.

[158] Philip E Tetlock, Jane Bernzweig, and Jack L Gallant, *Supreme Court decision making: Cognitive style as a predictor of ideological consistency of voting*, Journal of Personality and Social Psychology **48** (1985), no. 5, 1227.

[159] Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan P. Yamshchikov, *Style transfer for texts: Retrain, report errors, compare with rewrites*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong, China), Association for Computational Linguistics, November 2019, pp. 3936–3945.

[160] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li, *Modeling coverage for neural machine translation*, arXiv preprint arXiv:1601.04811 (2016).

[161] Prasad Vana and Anja Lambrecht, *The effect of individual online reviews on purchase likelihood*, Tuck School of Business Working Paper (2020), no. 3108086.

[162] Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, et al., *Tensor2tensor for neural machine translation*, arXiv preprint arXiv:1803.07416 (2018).

[163] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[164] Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey E. Hinton, *Grammar as a foreign language*, CoRR **abs/1412.7449** (2014).

[165] Ulrike Von Luxburg, *A tutorial on spectral clustering*, Statistics and Computing **17** (2007), no. 4, 395–416.

[166] Paul J Wahlbeck, James F Spriggs, and Lee Sigelman, *Ghostwriters on the Court? A stylistic analysis of US Supreme Court opinion drafts*, American Politics Research **30** (2002), no. 2, 166–192.

[167] Tong Wang, Ping Chen, Kevin Michael Amaral, and Jipeng Qiang, *An experimental study of LSTM encoder-decoder model for text simplification*, CoRR **abs/1609.03663** (2016).

[168] Artemus Ward and David L Weiden, *Sorcerers' Apprentices: 100 Years of Law Clerks at the United States Supreme Court*, NYU Press, 2006.

[169] Richard H Weisberg, *Law, literature and Cardozo's judicial poetics*, Cardozo L. Rev. **1** (1979), 283.

[170] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al., *Google's neural machine translation system: Bridging the gap between human and machine translation*, arXiv preprint arXiv:1609.08144 (2016).

[171] Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer, *Paraphrase generation as monolingual translation: Data and evaluation*, Proceedings of the 6th International Natural Language Generation Conference, Association for Computational Linguistics, 2010, pp. 203–207.

[172] ______, *Sentence simplification by monolingual machine translation*, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Association for Computational Linguistics, 2012, pp. 1015–1024.

[173] Wei Xu, Chris Callison-Burch, and Courtney Napoles, *Problems in current text simplification research: New data can help*, Transactions of the Association for Computational Linguistics **3** (2015), 283–297.

[174] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch, *Optimizing statistical machine translation for text simplification*, Transactions of the Association for Computational Linguistics **4** (2016), 401–415.

[175] Wei Xu, Alan Ritter, William B Dolan, Ralph Grishman, and Colin Cherry, *Paraphrasing for style*, 24th International Conference on Computational Linguistics, COLING 2012, 2012.

[176] Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi, *Controlling the voice of a sentence in japanese-to-english neural machine translation*, Proceedings of the 3rd Workshop on Asian Translation (WAT2016), 2016, pp. 203–210.

[177] G Udny Yule, *On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship*, Biometrika **30** (1939), no. 3/4, 363–390.

[178] Hongyu Zang and Xiaojun Wan, *Towards automatic generation of product reviews from aspect-sentiment scores*, Proceedings of the 10th International Conference on Natural Language Generation, 2017, pp. 168–177.

[179] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals, *Recurrent neural network regularization*, CoRR **abs/1409.2329** (2014).

[180] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi, *Defending against neural fake news*, CoRR **abs/1905.12616** (2019).

[181] Yexun Zhang, Ya Zhang, and Wenbin Cai, *Separating style and content for generalized style transfer*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8447–8455.

[182] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin, *Adversarial feature matching for text generation*, Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 4006–4015.

[183] Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen, *Style transfer as unsupervised machine translation*, ICLR 2019; arXiv preprint arXiv:1808.07894 (2018).

[184] Lei Zheng, Vahid Noroozi, and Philip S Yu, *Joint deep modeling of users and items using reviews for recommendation*, Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, ACM, 2017, pp. 425–434.

[185] Feng Zhu and Xiaoquan Zhang, *Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics*, Journal of marketing **74** (2010), no. 2, 133–148.

[186] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych, *A monolingual tree-based translation model for sentence simplification*, Proceedings of the 23rd international conference on computational linguistics, Association for Computational Linguistics, 2010, pp. 1353–1361.