MisInfoWars: A Linguistic Analysis of Deceptive and Credible News

by

Emilie Francis

B.A., University of Victoria, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
Department of Linguistics
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY
Summer 2018

Approval

Name:	**Emilie Francis**
Degree:	**Master of Arts (Linguistics)**
Title:	**MisInfoWars: A Linguistic Analysis of Deceptive and Credible News**
Examining Committee:	**Chair:** Ashley Farris-Trimble Assistant Professor
	Maite Taboada Senior Supervisor Professor
	Trude Heift Supervisor Professor
	Fred Popowich Professor School of Computing Science Simon Fraser University
Date Defended:	**July 31, 2018**

Abstract

Misinformation, bias, and deceit, clandestine or not, are a pervasive and continual problem in media. Real-time mass communication through online media such as news outlets, Twitter, and Facebook, has extended the reach of deceptive information, and increased its impact. The concept of fake news has existed since before print, but has acquired renewed attention due to its perceived influence in the 2016 U.S. Presidential election. Previous studies of fake news have revealed much about why it is produced, how it spreads, and what measures can be taken to combat its rising influence. Despite the continued interest in fake news, current research on the language of deceptive media has been largely superficial. This thesis serves to provide a profound understanding of the stylistic and linguistic features of fake news by comparing it to its credible counterpart. In doing so, it will advocate for differentiation between disingenuous and respectable media based on linguistic variation. With a dataset of approximately 80,000 articles from known fake and legitimate news sources, specific stylistic differences will be examined for saliency and significance. Using multidimensional analysis for discourse variation established by Biber (1988), this thesis will confirm that there exist sufficient textual differences between the articles of fake news and credible news to consider them distinct varieties. Detecting misinformation has not proven to be simple, neither has minimizing its reach. As the ambition of fake news articles is to appear authentic, acquiring knowledge of the subtleties which serve to discriminate realism from fabrication is crucial. A better understanding of the linguistic composition of deception and fabrication in comparison to credibility and veracity will facilitate future attempts at both manual and automatic detection.

Keywords: Fake News, News Text, Corpus Linguistics, Multidimensional Analysis

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recently there has been growing concern around the validity of sources of information and the influence such information has on society. After the 2016 United States presidential election, the effect of fictitious and biased information on public opinion has been called into question. The bulk of this criticism has focused on social media, considering it a major vehicle for malicious misinformation and influencing public opinion (Spinney, 2017). The social media site Facebook has found itself facing the majority of the blame for the prevalence and influence of misinformation during the 2016 electoral cycle. The website has been accused of abetting the spread of misinformation (Isaac, 2016), possibly contributing to the election of Donald Trump (Parkinson, 2016; Read, 2016).

In 2016, 62% of American adults got their news from social media, specifically from Reddit, Facebook, and Twitter (Gottfried & Shearer, 2016). Facebook is reported to reach 67% of American adults, equating the users who consume news through the site to 44% of the population (Gottfried & Shearer, 2016). As these surveys were conducted in 2016, these numbers are likely to be even higher at present. With such a substantial amount of the population regularly interacting with misinformation, the potential for influence is considerable. It is no surprise that there has been an increase in awareness and outrage towards misinformation and social media.

During the 2016 election cycle and into 2017, a number of 'fake news' stories emerged, generating both confusion and skepticism. One example involving former president Obama reached 1.7 million views through Facebook (Roberts, 2016). The story claimed that Obama had cut 2.6 billion in funding from veterans and reallocated it to Syrian refugees, but was later proven complete falsehood based on budget deficit. In 2017, an article calling out a Canadian Imam for refusing shelter to Christian victims of Hurricane Harvey was shared more than 126,000 times (Rannard, 2017). The story was determined to be false when the man later confirmed that he had never been to Texas, and had been in Mecca at the time of the floods. Despite the website responsible for the articles claiming itself satirical, the story was taken seriously by a substantial amount of readers.

Though it has gained recent notoriety, the issue of deceptive news is not unique to the past few years. In 1938, papers reported mass panic after Orson Welles' adaptation of 'War of the Worlds' premiered on radio (A. Chen, 2017). According to the papers, the mass hysteria reportedly induced by the broadcast was discovered to be greatly exaggerated. In 1475, false testimony from a preacher concerning the disappearance of a young boy resulted in the torture and execution of many members of the Jewish community (Soll, 2016). Though misinformation has been used as an instrument for sensationalism and propaganda in print since the invention of the craft, it has become a topic of particular interest and escalating anxiety in the past few years. Social media, and the Internet in general, have allowed for mass communication in real time. This instantaneous access to content hosted online has enabled misinformation to spread like wildfire, possibly having very serious real world consequences.

1.1 Past and Present Research in Deceptive News

Misinformation, bias, and deception, whether intentional or unintentional, have become a significant problem in online discourse. While the creation of deceptive news is not always motivated by a desire to cause misunderstanding, as is the case with satirical news, unexperienced consumers may not recognize false claims when they encounter them.

Previous research has delved into understanding what constitutes deceptive news, how and why it is created, and how it spreads. Through studying the language of deception in both speech and writing, trends of deception in text are able to be observed. Hedging, subjectivity vocabulary, higher word count, and bias markers are all common in deceptive writing (Hancock, Curry, Goorha, & Woodworth, 2007; Horne & Adali, 2017; Volkova, Shaffer, Jang, & Hodas, 2017). Linguistic features such as pronoun usage, negation, superlative, and modal adverbs are also associated with deception in writing (Hancock et al., 2007; Pérez-Rósas & Mihalcea, 2015; Rashkin, Choi, Jang, Volkova, & Choi, 2017). In general, the accepted definition of fake news is such news that is determined to be provably false.

The motivation for creating deceptive news is often financial gain, as revenue from advertisements on such articles is quite lucrative. Google Alphabet's AdSense program presents a unique opportunity, where successful writers can make upwards of $10,000 a month. However, while money may be a major factor in the inspiration behind deceptive news, some articles are created and propagated with the intention of influencing political opinion. In the recent scandal with Cambridge Analytica, it was revealed that the U.K. company ran Donald Trump's digital election campaign by manipulating the flow of information online, using Facebook users' personal data, and planting misinformation (*Cambridge Analytica offices searched over data storage*, 2018; *Exposed: Undercover secrets of Trump's data firm*, 2018). Aside from political and financial incentives, some deceptive news creators simply enjoy the act of deceiving their readers.

The spread of information online has been adequately represented using algorithms designed for modeling hydrodynamics or epidemics. Such methods are useful for predicting how information will diffuse through a network, but do not discuss the specific mechanisms utilized to propel such information. Bots and organic users act as vehicles for distributing and circulating rumors and misinformation, with the relatively small amount of undetected bots responsible for a significant amount of the claims. Authentic users aid in the expansion of information by showing a preference for the spread of unverified information and showing bias toward the spread of information which confirms preexisting beliefs, regardless of the veracity of such information.

Currently, methods of combating deceptive news have been relying on educating readers in various techniques which utilize critical thinking to differentiate between false and verified information, as well as authentic and inauthentic sources. Many websites, such as the Globe and Mail, the BBC, and Buzzfeed, have created quizzes, such as seen in Figure 1.1, which aim to both correct false claims and educate readers on how to recognize misinformation. From this image, an argument can be made that another possible intention of deception and rumor is the sowing of distrust of mainstream media. The author of the tweet is attempting to present misinformation with the intent for framing the original reporting agency as the offender.

Attempts to solve the problem automatically, through classification with machine learning, or by checking sources against databases of known deceptive news sites have also achieved some success. The BS detector, a web browser extension which applies labels such as deceptive news, extreme bias, junk science, and hate group, operates using a list of confirmed unreliable sources to provide warnings to its users (Hagen & Lutzenberger, 2016). Traditional machine learning approaches have shown promise in classifying deceptive news articles when combined with linguistic features and psycholinguistic features drawn from the linguistic inquiry and word count (LIWC) (Pennebaker, Francis, & Booth, 2001). Neural networks, which gain information from reviewing large corpora, have also been employed to classify texts as deceptive or legitimate with reasonable accuracy (Bajaj, 2017; Horne & Adali, 2017; Mihalcea & Strapparava, 2009; Pérez-Rósas, Kleinberg, Lefevre, & Mihalcea, 2017; Rashkin et al., 2017; Rubin, Conroy, Chen, & Cornwell, 2017; Yang, Mukherjee, & Dragut, 2017).

Given the massive influence misinformation can hold, it is important to combat deception online to prevent malicious manipulation of public opinion. Understanding deceptive news better and improving automatic detection techniques will help the average news consumer remain informed, while allowing them to avoid misleading information.

Figure 1.1: An image from a Buzzfeed quiz on misinformation (Lytvynenko, 2017).

1.2 The Economics of Misinformation

The concept of 'fake news' has garnered much attention from both the public and researchers, largely driven by the influence of Donald Trump. In the 2016 election, it was discovered that a large amount of deceptive news articles were written in support of Trump (Allcott & Gentzkow, 2017). There has also been a lot of criticism directed at the president for his obsession with fake news, particularly after his announcement of a 'fake news award', which some have deemed a breach of federal ethics (Kruzel, 2018). Despite the majority of deceptive news focusing on controversial issues within or related to the United States and its politics, it appears that the main motivation for many deceptive news sources is not political.

The opportunity for financial gain seems to be the primary impetus for many creators of deceptive news. A dominant source of deceptive news was the town of Veles, Macedonia,

with over 100 websites hosted during the 2016 election (Graham, 2017; Subramanian, 2017). One member of the community claimed to have earned $16,000 off of his pro-Trump websites (Subramanian, 2017). The owner of Disinfomedia, a company attributed to numerous sites that spread deceptive news during the election, remarked that proprietors could earn as much as $10,000 to $30,000 in a month from these stories (Sydell, 2016). Paul Horner, a kingpin of the deceptive news phenomenon during the 2016 election, has stated that he makes $10,000 a month from Google AdSense (Dewey, 2016).

Although Facebook and Google have made promises to be more proactive in combating misinformation on their respective platforms (Love & Cooke, 2016), it appears that doing so would be counter to their own interests. Despite Google's fame as a search engine, the company gained 88.7% of its wealth from advertising in 2016 (Porat, 2017). In 2017, Alphabet and Google earned a total $73.9 billion in ad revenue, 33% of the revenue from advertisements globally (Graham, 2017). As the producer of an article and Google gain revenue from traffic to the page, a portion of this revenue is invested into promotion of the site's articles on Facebook (Graham, 2017). The writers of one popular source of misinformation, Paris Wade and Ben Goldman, claimed that $3000 of their income was paid to Facebook each month for promotional purposes (Tynan, 2016).

While money may be a compelling incentive to produce misinformation, some writers simply experience enjoyment from deceiving their readers. In an interview, one deceptive news writer claimed that seeing his websites accumulate views was "like an addiction" (Radutzky, 2017). Other writers who create deceptive news do so simply because they believe in what they are reporting. A lawyer from Southern California and deceptive news writer, Michael Cernovich, claimed that he "doesn't say anything he doesn't believe" (Pelley, 2017). Based on such anecdotes, it seems the individuals who produce fabricated news stories do so for a variety of reasons. Since there is great potential for sensationalized headlines to attract clicks and yield ad revenue, one can understand why many people may choose to write fabrications as a means of income. Motivations such as entertainment and sincerity, though more difficult to understand, also have their place in contributing to the issue.

1.3 What is Fake News?

After seeing some examples of deceptive news, as well as discovering the various reasons behind their creation, it is important to discuss what 'fake news' actually is. Commonly, the term 'fake news' is used as an all-encompassing term including the concepts of falsehood, credibility, bias, and satire. Not all cases of 'fake news' are completely untrue, neither are they all intentionally misleading. It is also possible for credible sources to exhibit some elements of 'fake news', such as bias or misinformation. Figure 1.2, which presents an image of a tweet from the Twitter channel of CNN Reporter Oliver Darcy (A. B. Wang, 2017), refers to an incident in which former ABC News journalist Brian Ross incorrectly reported

that Trump had directed Michael Flynn to make contact with Russian officials before the election. ABC issued a clarification through Twitter, which has since been deleted, and through Ross's appearance on "World News Tonight". Even credible news agencies are capable of making mistakes which may have larger consequences.

Figure 1.2: The tweet from Twitter channel of CNN Reporter Oliver Darcy (A. B. Wang, 2017) on the ABC News Scandal.

Satire is a form of political commentary that uses rhetorical devices, such as exaggeration and ridiculousness, to elicit a reaction from its readers. The majority of satirical news is not intended to mislead its readers, though the possibility of misinterpretation remains. Since satirical news attempts to mimic traditional journalism, and often takes inspiration for their articles from real world events, those who are not familiar with the format may be deceived by some of the headlines. Many articles from satirical news sites, such as the Onion or the

Beaverton, are shared over social media, where some readers may be fooled (Rubin et al., 2017).

In biased sources, there is some element of truthfulness to the information. Though containing a kernel of truth, biased information is presented in a manner that portrays some topic favourably or unfavourably according to the opinions of its creator. Bias is not always willful, nor is it a feature specific to 'fake news'. Respectable writers are capable of exhibiting bias, although the presence of bias in sources is correlated with a loss of credibility.

Credibility, or believability, is a perceived quality which is affected by various factors. A survey conducted by Fogg et al. (2001), determined the most persuasive factors in assessing website credibility. From the responses, eight unique scales were outlined, including real-world feel, ease of use, expertise, trustworthiness, tailoring, commercial application, and amateurism. The results showed that real world feel and ease of use had the strongest positive impact on perception of credibility, while commercialism and amateurism had the strongest negative impact. Specific features, such as infrequent updating and challenges in distinguishing advertisement from content were the most closely related to a decrease in perceived credibility. The credibility of a source has a profound effect on one's interpretations of the information. People are more likely to trust sources they are familiar with, or which confirm their preconceived notions of the world (Lazer, Baum, & Mele, 2017).

Given the multiplicity and politically charged nature of 'fake news', many researchers prefer to avoid the term in favor of more specific words such as 'misinformation' and 'disinformation'. The most concise definition of fake news, as defined by Allcott and Gentzkow (2017), is that fake news are "news articles that are intentionally and verifiably false, and could mislead readers". As the research presented here concerns all manner of deceptive information, the term 'deceptive news' will be employed. The label of 'deceptive' avoids the controversy of the colloquial 'fake news', while remaining an accurate descriptor of the data presented in this thesis.

1.4 A Text-based Approach to Understanding Deceptive News

Current approaches to quantifying and detecting misinformation in news have focused on utilizing metadata, such as the source of the information, verdicts from dedicated fact checking authorities, and where, how, and by whom it is shared.

Analysis of the language of deception has revealed cues specific to types of satire, click-bait, and propaganda, as well as general features used in false claims. Research on diffusion has put effort into mapping the spread of rumors, hoaxes, and fabrication through social networks by treating users as nodes in hydrodynamic and epidemiological models. Automatic deception detection has made use of large datasets to train and implement useful tools for classifying information as true or false as it is encountered.

In spite of the wealth of previous research, a general concept of what constitutes 'fake' or 'deceptive' news has not yet been achieved. An in-depth analysis of the actual content of deceptive and credible articles for the purpose of comparison and differentiation has been largely neglected. When linguistic analysis has been applied to deceptive information, it has been used to improve automatic detection tasks without any detailed discussion of their purpose in the text, how they are used, or how it can be compared to credible news. Furthermore, when the content of articles has been used, it has been restricted by subtype of deceptive news and has not included text types from all manifestations of misinformation.

Research on diffusion, although useful for curtailing rampant and damaging rumors and hoaxes, provides little insight on the anatomy of deceptive information. Should a false claim fall through the cracks, users will be no less susceptible to its influence.

Automatic deception detection serves to ease the experience of social media users and news consumers by bearing the burden of identifying misinformation. However, these programs are not infallible and risk removing the right of an individual to make their own decision on what content they wish to view. Moreover, particularly in the case of black box machine learning algorithms, they do not contribute toward the understanding of what makes 'fake' news deceptive.

In this exploration of deception, I will provide a new approach to distinguishing deceptive news which utilizes the content of the articles themselves. A detailed analysis of the language of deceptive news and how it compares to credible news will be conducted with the assistance of multidimensional analysis. The variation in writing will be used to argue for a distinction in register between deceptive and credible news.

Support for compartmentalizing deceptive and credible news will come from an investigation of register with the multidimensional analysis framework (Biber, 1988). Two corpora of substantial size have been tagged and analyzed for significant differences in writing. The results of these analyses have been compared with each other, other established types of text, and other varieties of news.

As the following chapters will illustrate, the content of credible news and deceptive news exhibits significant differences. These differences manifest in both general tendencies, such as density of information and narration, and specific linguistic features, such as nouns and sentence relatives. Such differences provide support for the argument that deceptive news and credible news are two distinct varieties of news.

This thesis will contribute a more vivid and generalized understanding of the constitution of deceptive news and confer a completely different lens through which to view deceptive and credible media. The findings of this thesis may be purposed to provide useful features for future classification tasks in machine learning. Testing a successful model built with the assistance of the findings from this analysis on a more general classification task will assess the possibility of generalized deception detection models, as well as encourage further research into misinformation detection. In addition, understanding of the rhetorical devices

and discourse of deceptive writing will supply readers with more knowledge to be used in forming their own informed decisions on information online.

In the following chapters, I will begin by discussing previous research in the language of deception, the patterns of diffusion, machine learning approaches to deception classification tasks, and how these relate to the research presented in this thesis. Then, I will discuss the data used for this study and explain why it is well suited for this research. The methodology behind text type analysis and its previous applications will be discussed. I will also provide a brief description of the analysis process. The results of the analysis will be discussed in detail, including the different dimensions of writing and specific linguistic features. The results for each corpus will then be compared to other common varieties of text. Concluding this thesis, I will summarize the major findings and contribute remarks on the future directions of this research.

Chapter 2

Disinformation: The Language of Lies, Spread of Misinformation, and Deception Detection

Despite deceptive news being as ubiquitous and indistinguishable as it is, the onus has largely been placed on the reader to decide whether or not to believe the contents of an article. Multiple organizations, including universities such as Simon Fraser University, University of British Columbia, and University of Toronto, have provided helpful guides for identifying deceptive news, as illustrated in Figure 2.1. These guidelines are available in multiple languages and involve a series of tips for determining deceptive news. Common tips involve checking the sources and the author and considering biases. However, distinguishing deceptive news from authentic need not be so burdensome for readers.

In this chapter I will introduce the current understanding of fabrication and deception in speech and writing achieved from previous research. The most successful models for dissemination will be presented in the context of how they have been useful for mapping the spread of falsehood versus reliable information. Finally, common approaches in machine learning for classification will be introduced, along with examples of these approaches in practice. Through this discussion, it will be demonstrated that the idiosyncrasies of deceptive writing and credible writing are distinct, and that deceptive news and credible news are not identical.

2.1 Deception in Speech and Writing

In speech, higher frequency and longer duration of hedging words, such as 'um' or 'like', can indicate that someone is being truthful (Arciuli, Mallard, & Villar, 2010). In a study of online interactions, use of first person personal pronouns and affect words were higher amongst deceivers, while negations and words of cognition were used more frequently among truth tellers (Ho et al., 2015).

Pérez-Rósas and Mihalcea (2015) studied deception detection through a dataset composed of short texts exclusive of context and domain with the aim of classifying age and gender in deceptive writing. Analysis showed that spontaneous lies made use of negation, certainty, and second person pronouns, which indicated an attempt at reinforcement of lies through stronger wording. In general, individuals were less likely to lie when describing their family, religion and positive experiences. Men typically employed more references to friends and others when lying, while women tended to prefer references to money and the future. Female truth-tellers used metaphor, whereas males made references to sports and music. Older deceivers generally made more references to anxiety, money, and motion, while younger liars used anger, negation, and vocabulary associated with death (Pérez-Rósas & Mihalcea, 2015).

It seems sensible to assume that deceptive news would also display cues related to deceitful language. Volkova et al. (2017) identified a number of cues derived from posts on Twitter pertinent to satire, clickbait, and hoax, which included bias, subjectivity, psycholinguistic, and moral foundation cues. Inquiry into the cues revealed that verified news contained fewer bias markers, hedges and subjective vocabulary. In regard to moral cues, they presented fewer instances of authority and contrast between harm and care, and loyalty and betrayal. Propaganda, hoax, and clickbait were found to be similar based on moral, bias, and subjectivity cues, though propaganda targeted morality and authority more than satire and hoax. Compared to satire, hoax and propaganda exhibited fewer bias markers, such as hedging and implicative verbs. Compared to hoaxes, propaganda and clickbait displayed more factive verbs and bias markers (Volkova et al., 2017).

Horne and Adali (2017) explored the content and titles of deceptive news to determine linguistic and stylistic features of deceptive writing with the use of a dataset consisting of multiple authentic and deceptive news sources. Stylistic features, complexity features measured with readability indexes such as Gunning Fog and Flesh-Kincaid, and psycholinguistic features from the Linguistic Inquiry and Word Count or LIWC, (Pennebaker et al., 2001), were categorized for study. Analysis of the data showed that credible news is typically longer, while deceptive uses fewer technical words, has a shorter average word length, has less punctuation and quotations, and increased lexical redundancy. Results concerning readability indicated that deceptive news necessitates a lower education level for comprehension compared to authentic news. The titles of deceptive news are generally longer, contain simpler words, and have fewer stop words (Horne & Adali, 2017).

Using a corpus for satire, hoax, and propaganda collected from Gigaword, Rashkin et al. (2017), examined the linguistic patterns across different types of articles. It was observed that first and second pronouns were more frequent in deceptive news, as well as subjectivism, superlatives, and modal adverbs. Credible news normally employed more comparatives, numbers, and assertive language, while avoiding hedging. Satire was discovered to display

a high frequency of adverbs, and hoaxes tended to use more comparatives and superlatives (Rashkin et al., 2017).

Despite its intention of providing entertainment, satire is often linked with deceptive news due to its potential to mislead. Experienced readers are easily able to distinguish satire from credible news given the absurdity of its content. However, some sources of satire are quite subtle, rendering it challenging to distinguish them from authentic sources. Rubin et al. (2017), created a dataset of 360 satirical articles for the purpose of analyzing and defining features of satirical news. Five satirical features were identified in the study, including absurdity, humor, grammar, negative affect, and punctuation. Analysis revealed that credible news and satire were found to vary the most in the first and final lines of the articles. Credible news typically reported new information in the first line, while satire normally repeated information presented in the title. In satirical news articles, the final line showed similarity with the 'punch line' of a joke. Satire also exhibited more common application of slang and cursing, as well as packing a generous number of clauses into sentences to enhance comedic effect (Rubin et al., 2017).

2.2 How Misinformation Spreads

Although deceptive news is not a modern invention, social media and other modern conventions, such as search engines, can be easily manipulated to provide a robust space for it to flourish. Understanding how unverified information and false claims flow throughout a social network is essential for prevention and containment.

One of the most common types of reporting on social media is rumor. Rumor, a concept strongly correlated with misinformation, is described as a statement of questionable veracity which stimulates skepticism. Zubiaga, Liakata, Procter, Hoi, and Tolmie (2016) investigated the spread of rumors on social media, specifically on Twitter. A collection of 330 conversation threads discussing nine different newsworthy topics were utilized to study rumor diffusion and patterns of support and denial on social media. Tweets which were marked by a high number of retweets were selected for veracity assessment by professional journalists. 2,695 tweets in English, German, and French were annotated for truthfulness. Results showed that the median for resolution of true rumors was approximately two hours, while the median for false rumors was over fourteen. Tweets which include unverified rumors are more widely dispersed, with the majority of retweets occurring within the first few minutes. While unverified rumors spark the most retweets, those which are later confirmed receive more retweets than false rumors. However, once a rumor has been verified as true or false, the rate of diffusion decreases dramatically. A concerning discovery noted that reputable users, such as news agencies, have a tendency to support rumors, regardless of their being eventually confirmed or denied.

The diffusion of assertions, both rumor and news, throughout Twitter was also studied by Vosoughi, Roy, and Aral (2018). Using a dataset of 126,000 rumour cascades tweeted more than 4.5 million times, claims were examined for veracity and labeled true, false, and mixed. The true and false claim cascades were then quantified for depth. The results from this analysis confirmed the findings of Zubiaga et al. (2016), with cascades of false rumors diffusing further, deeper, more quickly, and more extensively than true rumors. False political news was found to have the most prominent reach, spreading deeper and more broadly than other categories of disinformation. The novelty of information was also tested due to its possible influence on diffusion. Rumors were found to be significantly more novel than their truthful counterparts, with assessment of emotional content within replies confirming that users respond with more surprise to rumors than truthful information. In general, the results from this analysis reinforced the findings of Zubiaga et al. (2016), revealing that rumors and false information are supported more by the community and experience a more expansive spread.

Hoaxes, a form of misinformation similar to rumor, tend to display the opposite in terms of life expectancy. Kumar, West, and Leskovec (2016) studied misinformation in the context of Wikipedia hoaxes, with over 20,000 hoaxes articles flagged by Wikipedia's patrolling process. It was discovered that when a hoax is created, 90% of the time it is flagged within one hour. Most hoaxes are ephemeral, with the detection rate of hoaxes steadily increasing over time, from 92% after a day, 94% at a week, and up to 96% after a month, with only one in a hundred surviving over a year. It was also revealed that most hoaxes are rarely viewed, with a median of three views per day. Likely, the transitory nature of hoaxes on Wikipedia is due to an efficient moderation and flagging system utilized by the site. Hoaxes, when distributed through less regulated channels, are more likely to behave like rumors given their similar nature.

Although there is an expectation for distinguished news organizations to uphold journalistic integrity by avoiding spreading unsubstantiated claims, competitiveness and pressure may drive journalists to publish information as soon as they receive it. One method of countering deceptive news is to hinder its diffusion, particularly through social networks.

Three roles are necessary for information to spread throughout a network, the spreaders who initiate, the spreaders who propagate, and the receivers (X. Hu, 2016). Epidemiological models, such as SIR, are effective at predicting diffusion patterns by classifying nodes as either susceptible to infection, infected, or in recovery (X. Hu, 2016). Within the context of deceptive news, susceptible individuals are those who have not yet come in contact, while infected individuals are users who possess the opportunity to spread the misinformation. Individuals in recovery are said to be users who have encountered the information, but are not spreading it. Under this assumption, within time, all users in a network will become infected. However, despite mass dissemination, some nodes persist unaffected.

To explain this phenomenon, X. Hu (2016), identified two different types of users in a network, forceful and regular. When regular users interact, they influence each other. In an interaction between a regular and a forceful user, only the regular user is affected and the forceful user preserves previous assumptions. If a social network is not well connected, the extent of the misinformation diffusion is enhanced. In disengaged communities, the forceful nodes act as a bridge and allow information to aggregate.

Another method of modeling information diffusion in online social network makes use of hydrodynamics. A problem with epidemiological and hydrodynamic models is the ignorance of variables such as influential users and social platform. Y. Hu, Song, and Chen (2017), improve upon a previous hydrodynamic model by combining the physical model of hydrodynamics with online social networks, considering factors such as influence of the user and popularity of the information. A dataset of 6,500 video tweets from Sina-weibo were collected and analyzed to identify influence, popularity of information, and the platform of diffusion. The model performed at 76.7% accuracy, but increased to 83.05% accuracy with the superimposition of factors related to information diffusion (Y. Hu et al., 2017).

A common method of spreading misinformation is through the use of bots. Shao, Ciampaglia, Varol, Flammini, and Menczer (2017), collected 15,053 articles containing approximately 390,000 debunked and unverified claims from known deceptive news articles, as well as over 1,130,000 posts containing fact-check URLs and over 13,615,000 posts linking to claims from Twitter. From the accounts posting a high number of claims, a sample of 915 were analyzed with Botometer to compute the likelihood that the account was controlled by a program. Only 8% of the 915 accounts were labeled as bots, but they were responsible for 36% of claims. Bots were found to be most active in sharing links within seconds after they are posted, as well as exploiting influential users by mentioning them in tweets (Shao et al., 2017).

Using A collection of roughly 270,000 posts from 73 Facebook pages, Bessi et al. (2015), studied trends among users and shares of pages associated with science and conspiracy. Results showed that conspiracy articles receive more likes and shares than science. Conspiracy consumers were also less likely to trust non-conspiracy sources of information. Users polarized on conspiracy news, number of likes with respect to total like activity being 95% on one category, interact within their own community, while users polarized on science commented slightly more outside. Conspiracy page followers were also found to be more prone to interaction with false claims (Bessi et al., 2015).

2.3 Deceptive News and Machine Learning

In order to ebb the tide of misinformation, a great deal of effort has been poured into the identification of deceptive news as it arises. Wikipedia and Opensources (Hagen, 2017), as well as many independent researchers, have maintained lists of known deceptive news

sources. Many fact-checking organizations, including Politifact and Snopes, have also been attentively combating misinformation online since well before the renewed concern toward deceptive news. Despite extensive documentation on sources of misinformation and expeditious correction, deceptive news remains a pervasive issue.

Once misinformation has been absorbed, correcting it does not always effectively change the consumer's beliefs (Lazer et al., 2017). Due to familiarity biases, repetition may be harmful and enhance the effect of misinformation. Preventing the negative effects of misinformation requires more than simply correcting narratives as they arise. As the sensitivity to deception and misinformation increases, researchers and other interested individuals have turned to automatic detection methods as a means of combating deceptive news.

2.3.1 A Soft Introduction to Machine Learning Approaches Used in Deception Classification

Advancements in machine learning have made the task of detecting misinformation by classifying large amounts of data more approachable. Approaches to machine learning take one of two forms, supervised or unsupervised. Unsupervised approaches have no associated label and are best used in clustering tasks. In an unsupervised system, the algorithm identifies patterns which it uses to organize the data. Supervised approaches, such as Support Vector Machines (SVM), Naive-Bayes, Logistic Regression, and Decision Trees, learn a mapping from an input x to an output y from a large amount of data, which is then applied to novel data to predict its output. The learning is considered supervised as the answers are already known, which allows it to adjust itself and make better predictions though continued iterations. Supervised learning is used for classification tasks, and is the method most often employed in deception detection.

Some approaches, such as SVM and Naive-Bayes, benefit from the inclusion of linguistic and paralinguistic features. A SVM classifier plots each data point in n-dimensional space, with n being the number of features, then attempts to draw a line that best categorizes the two classes. These features may be linguistic, such as syntactic features or parts of speech, or paralinguistic such as article length or n-grams.

The Naive-Bayes classifier is based on Bayes theorem, working with conditional probability. It functions on the premise of the probability of x occurring, given that y has already occurred. Probability of membership to a class is predicted based on the probability that a given data point belongs to a particular class. Logistic Regression, another probabilistic model, is based on the assumption that the input space can be categorized into two distinct regions with a linear boundary.

Random forests, based on Decision Trees, devise a set of rules from a given dataset and features. As input traverses down through the decision trees, it gets put into smaller sets. Typically, the more trees in the forest, the better the results will be.

Maximum Entropy classifiers, based on the principals of uniformity or maximum entropy, begin as a uniform distribution when no data is observed. As data is introduced, the classifier attempts to explain the data and thus deviates from maximum entropy. After data has been explained, the classifier will then return to maximizing the entropy on remaining unseen data.

Other algorithms, mainly neural networks, act as a black box where data is supplied and the network independently decides which features are most indicative. A Recurrent Neural Network (RNN) makes use of sequential information to make predictions about an input. What differentiates a RNN from a traditional neural network is that it does away with the assumption that all inputs and outputs are independent. It is said that a RNN has memory, which allows it to perceive information about previous calculations.

The Long Short Term Memory network (LSTM) is a variation of the RNN which is capable of learning long-term dependencies. All variations of a RNN are formed like a conveyor belt of repeated components of information. A LSTM network is also based on this structure, but is expanded into more layers.

Gated Recurrent Units (GRUs) may be considered a simplification of the LSTM. What differentiates a GRU from a LSTM is the presence of an update gate and a reset gate. These gates act as regulators and decide what information should be passed through to the output.

Convolutional Neural Networks (CNNs), commonly used for image classification, have also had much success with sentence classification tasks. CNNs differ from RNNs through the use of a filter which slides over sections of the input and does mathematical convolutions to learn mappings of features.

Hierarchical Attention Networks (HANs) are similar to GRUs, but have introduced the concept of an attention mechanism. The attention mechanism is used to extract words from a sentence in the input which are deemed salient to the meaning of the sentence. This attention mechanism is also performed on sentences which are considered useful for classifying a document correctly.

Another useful tool commonly used as an addition to feature-based machine learning networks is term frequency - inverse document frequency (tf-idf). The term frequency (tf) is a statistic which measures the frequency of a term in the given text, while the inverse document frequency (idf) measures how common a word is in all documents.

This section has sought to provide a brief introduction to the most popular methods utilized in the task of classification. The following section will provide examples for ways in which these methods are applied to real classification tasks for the purpose of deception detection.

2.3.2 Automatic Deception Detection

Previous research has delved into the problem of detecting misinformation, rumor, conspiracy, and satire throughout social media and independent news. Using a manually labeled dataset comprised of approximately 12,386 short statements from the fact-checking website Politifact, W. Y. Wang (2017) created and tested multiple machine learning models for deceptive news detection. The models were either standalone, or combined with metadata, such as the subject, speaker, or context. While the results of these tests were unremarkable, the CNN was found to perform the best. Without the use of additional features, it received a score of 27% in testing. When combined with metadata, the CNN's test score rose to 27.4% (W. Y. Wang, 2017).

Mihalcea and Strapparava (2009) used Naive-Bayes and SVM classifiers to detect deceptive language in text. A novel dataset using opinions on abortion, the death penalty, and one's best friend was created through Mechanical Turk (AMT), yielding 100 true and 100 false statements for each topic. Contributors were instructed to write one statement, of four to five sentences, expressing their true opinion, and one statement expressing the opposite. Both the Naive-Bayes classifier and the SVM performed reasonably well, at 70.8% and 70.1% accuracy respectively. LIWC was used to determine dominant word classes in deceptive text. It was found that in both truthful and dishonest text, the top dominant classes were human related. Deceptive texts included classes which expressed detachment from the writing, whereas truthful texts included classes related to the self and friendship (Mihalcea & Strapparava, 2009).

Pérez-Rósas et al. (2017) built a SVM classifier using data from legitimate news sources, such as USAToday and CNN, fabrications from AMT crowdsourcing, and celebrity gossip. Features, including n-grams, punctuation, psycholinguistic features from LIWC, readability and syntax, were combined with the SVM and tested on both the celebrity and AMT datasets. The features which performed the best for the AMT dataset were readability, with 78% accuracy, and the use of all features, with 74% accuracy. The best performing features in the celebrity dataset were a combination of all features, 73% accuracy, and punctuation at 70% accuracy (Pérez-Rósas et al., 2017).

Bajaj (2017) developed multiple machine learning models, including Logistic Regression, RNN, LSTM, and CNN, which were tested on a dataset of 63,000 articles obtained from Kaggle and Signal Media News. CNNs with max pooling and attention achieved the best performance on precision, though GRUs performed the best for both the f-score and recall. The author notes that high precision scores were seen for all models, as the data was heavily skewed with credible news articles outnumbering inauthentic ones to a great degree (Bajaj, 2017).

By identifying suspicious accounts on Twitter with PropOrNot, Volkova et al. (2017) were able to obtain a dataset of 130,000 tweets from 174 suspicious and 252 verified news

accounts. A RNN model and CNN model with late fusion were combined with cues related to bias, subjectivity, LIWC, and moral foundation, then applied to a classification task and compared with a logistic regression model. Both the CNN and RNN outperformed the logistic regression model, with 95% accuracy (Volkova et al., 2017).

Based on a corpus of nearly 14,000 documents from Gigaword containing satire, hoax, and propaganda, Rashkin et al. (2017) constructed a Max Entropy classifier and a LSTM model with output concatenated with LIWC feature vectors. Politifact was used to provide truthfulness distinction labels, establishing a six-point classification from true to pants-on-fire. As results for six-way classification were abysmal, classification was reduced to binary, which increased the performance of the classifier from approximately 20% to 50% (Rashkin et al., 2017).

In the study by Kumar et al. (2016) on Wikipedia hoaxes, Random Forests were used to answer the questions of whether an article is a hoax, how long it will survive, and whether a flagged article is truly a hoax. An accuracy of 92% was achieved when determining whether an article was a hoax, and 76% when concluding whether a flagged article deserved its hoax status (Kumar et al., 2016).

Using a BuzzFeed, US politics, and satire dataset, Horne and Adali (2017) studied the content and titles of deceptive news to identify stylistic features. A SVM was developed for classification between fake and real, satire and real, and satire and fake. Classification between the content of satire and real news achieved the best performance, with 91% accuracy, and classification between satire and fake news performed the worst, with only 55% accuracy on the title and 67% on the body (Horne & Adali, 2017).

Rubin et al. (2017), investigated classification of satire with a total of 360 news articles from two sites based in the US and Canada and two credible news sources, organized into four domains with three topics each. A trained linguist paired satirical articles with an equivalent credible article, and identified five satirical features, absurdity, humor, grammar, negative affect, and punctuation. An SVM with topic, semantic, and feature selection based on absurdity and humor was developed for classification. The baseline achieved 82% accuracy and increased to 87% with the inclusion of absurdity features.

Using a dataset collected from 14 different satirical news sites, Yang et al. (2017) implemented a wide variety of machine learning models, such as SVM, GRU, and 4-level HAN, combined with psycholinguistic, stylistic, readability, and structural features. Both the SVM with tf-idf and linguistic features and the 4-level HAN with paragraph and document level linguistic features attained an accuracy score of approximately 98% on testing, though the results of other models also performed well with none scoring below 90% in accuracy (Yang et al., 2017).

2.4 Human Deception Detection

The previous section has shown how adept computers are at detecting deception in writing. What remains to be discussed is the accuracy of actual humans in detecting deception in speech and writing. Assuming the characteristics of deceptive language, discussed in Section 2.1, one may assume that this task would not be overly difficult. However, the bulk of research suggests that this is not usually the case. In general, people perform only marginally better than chance with approximately 54% accuracy (Ekman, O'Sullivan, & Frank, 1999; Frank, Menasco, & O'Sullivan, 2008).

There is an inherent bias to presume honesty due to either failing to recognize the possibility of deceit or as a fallback state after failing to obtain sufficient evidence for deceit (Levine, 2014). This theory of a default presumption of truthfulness, the truth default theory (TDT) (Levine, 2014), makes a number of propositions, such as that most people believe most of what is communicated by other people most of the time. That said, certain conditions may improve ones ability to detect lies in speech. The higher the stakes of the lie, the more likely the liar is to express emotional states through facial expressions or other nonverbal cues (Frank et al., 2008; Levine, 2014). Training in lie detection, even if the techniques are mediocre, also has some positive effect on detecting deceit (Frank et al., 2008).

Certain professions and experienced individuals may perform more accurately on lie detection tasks, up to 67-73% (Ekman et al., 1999). Federal agents, trained police interrogators, and psychologists with an interest in deception were found to be the most accurate in their detection of lies, with up to 80% accuracy (Ekman et al., 1999). However, their detection of the truth was not vastly superior, with a maximum of 66% accuracy for federal agents. Certain personality factors, such as openness and agreeableness, may also improve one's ability, though not enormously (Enos et al., 2006). Openness, described as the degree to which an individual is open to experience and flexible in their viewpoints, is correlated with intelligence. Agreeableness, often associated with dependent personality disorder, may be expressed as an intense "attention to the affective state of others" (Enos et al., 2006).

In writing, human accuracy of deception detection is similar. Ott, Choi, Cardie, and Hancock (2011) conducted a study using truthful and deceptive positive reviews for hotels on trip advisor, where three human annotators were asked to make judgments on a subset of the reviews. The highest reported accuracy was 60% for one judge, with the agreement between judges being quite low. A Fleiss' kappa of 0.11 was calculated, indicating only slight agreement.

As previous research concludes, people are generally not that proficient at detecting lies, written or otherwise. While there are some ways to improve ones accuracy, very few individuals likely have the time and motivation to train themselves on lie detection techniques

or become federal agents and psychologists. The advancement of automatic detection is therefore vital to ensuring that misinformation is curbed before it is allowed to propagate.

2.5 Harmonizing Previous Research and the Current Research

The previous sections of this chapter have introduced various ways the phenomenon of disinformation has been approached in the past. In this section, I will discuss how each of these approaches to analyzing deception relates to the research presented in this thesis.

Existing research on features of misleading writing has shown that there is some basis for differentiation based on lexical diversity. In terms of evaluating for register or text type, linguistic features provide characteristics which form the basis for separation of discourse. The shortcoming of many of these previous studies is the small amount of data analyzed. The findings of this research serve to atone for the weaknesses of previous research by analyzing a significantly greater sample. In doing so, a more representative and generalized result will be produced.

Studies assessing the differences in diffusion between truthful and false information help to further differentiate deceptive news from credible news. Confirming varied behaviour in dissemination lends support to the notion of deceptive news and credible news as distinct. While much research has already concluded that diffusion of rumors, false claims, and hoaxes through social media channels differ substantially, less has been done to conclude whether the content of such stories is as varied in its content. In analyzing the lexical diversity of the articles, differences between the two categories will provide further evidence of distinction.

The adoption of machine learning for deception detection has allowed researchers to classify news efficiently and effectively. Unfortunately, many of these algorithms act as a black box and provide little to no insight on what the program is actually gleaning from the input. The research presented in this thesis will contribute much toward the knowledge of what differences these mysterious programs may be focusing on, as well as supply a more robust set of features for supplementing feature-based machine learning. Furthermore, analyzing dimensions of text will introduce a novel set of features for machine learning application.

In this thesis, I will present evidence with multidimensional analysis that deceptive news and credible news are distinct linguistically. Following the approach to variation in register (Biber, 1988; Nini, 2015), I will show that deceptive news and credible news vary in enough discrete dimensions to constitute exclusive varieties. Taking inspiration from the work on deceptive language as examined in Horne and Adali (2017), Volkova et al. (2017), and Rashkin et al. (2017), I will extract linguistic cues in the content of both deceptive and credible news sources. These cues will provide support for demarcating the text types of deceptive and credible news.

Figure 2.1: Guidelines created by the International Federation of Library Associations and Institutions (IFLA) based on a FactCheck.org articles (Kiely & Robertson, 2016).

Chapter 3

Methodology and Data

The primary goal of the exploration of deceptive news presented in this thesis is to ascertain whether or not deceptive news constitutes a register unto itself, distinct from recognized legitimate hard news.

In this chapter the concept of multidimensional analysis as it applies to variation in register will be discussed in detail. The fundamentals of the multidimensional analysis framework as devised by Biber (1988) and previous applications of the approach will also be examined. The data used for the analysis will also be introduced along with the analysis process.

3.1 Multidimensional Analysis and Varieties of Text

The multidimensional analysis approach (MDA) to register analysis was originally employed by Biber (1988). It was used to measure textual variation in several domains based on a wide variety of linguistic features over 481 texts. Six dimensions, with each dimension being associated with underlying communicative functions, have been established by Biber and used to associate a text with similar texts within a collection of registers called a 'text type'.

3.1.1 Register, Genre, Style, and Text Type

Before discussing multidimensional analysis as it applies to form and function of discourse, it is necessary to provide a clear definition for register, genre, and style.

From the perspective of Eggins (2004), register refers specifically to the style of language appropriate for a certain situation, while genre is characterized by obligatory elements which compose its structure.

According to Biber and Conrad (2009), register, genre, and style represent three perspectives of text analysis. Register combines analysis of linguistic variables which are common in a text type with analysis of the context. Linguistic features are considered functional, with some features being more commonly associated with the communicative purposes and situational context of certain texts. The driving force behind register is the dependence on

variation in the relative frequency of parts of speech, which allows this perspective to be machine measured.

Genre is similar to register in its inclusion of the communicative purposes and situational context of a text variety, but the linguistic analysis of genre targets the structures used to establish a complete text within the variety, as in the format of the beginning and conclusion of an e-mail.

Style is also similar to register in its analysis of the core linguistic features distributed throughout texts from a variety. However, the use of these features is not affected by the situational context. Features of style communicate aesthetic preferences of the authors or period in history.

These classifications of discourse all present similar but distinct perspectives to text analysis based on how linguistic variables are perceived, whether or not there should be consideration of context. The multidimensional analysis approach to text analysis also relies on these core linguistic features to classify texts. Although the appropriate terminology is still under debate, the terms used by Biber in his analysis are 'register' and 'text type' or 'text variety'. In this thesis, these terms will also be adopted.

3.1.2 Using Linguistic Variables to Determine Dimensions of Text

MDA analyzes a text by reviewing the linguistic tags within it and calculating a factor score for each. A setting of 400 words is used to calculate the type-token ratio, where these 400 tokens represent the maximum number of tokens analyzed for each text. As the length of a text varies, including the whole text introduces the possibility of distorting the results of the analysis. After experimentation, 400 was determined to be the most effective and was set as the standard in Biber (1988) and Nini (2015). With this, only the first 400 tokens will be analyzed as a means of balancing the texts. Therefore, in a text of 600 tokens, only the initial 400 would be considered in scoring. Texts which are fewer than 400 tokens in length are disregarded. This type-token ratio is used to measure the linguistic diversity within a text, as it presumed that the majority of linguistic variety in a text will occur within this space.

Factor scores represent a grouping of linguistic features which are observed to have high co-occurrence with each other. They are computed by summing the number of occurrences of each linguistic feature bearing a significant loading in a factor, characterizing texts in relation to that factor. In Nini (2015), the factor scores have been replaced with z-scores. Factor scores and z-scores are interchangeable in regard to function.

Some examples of linguistic features included by Biber in the analyses are analytic and synthetic negation, stranded prepositions, and phrasal coordination. The features would be measured with factor analysis, where the frequencies of linguistic features are reduced to a small set of variables, called factors. A table with the full list of linguistic features, the z-scores, and their associated dimensions has been provided in Appendix A.1. Some

features may have more than one associated dimension, as they are associated with either the positive or negative side of the scale. Associations were translated from the factors in Biber (1988). Features which do not have have an associated dimension were either from a defunct factor, or were not explicitly mentioned in Biber (1988).

In the 1988 analysis, results of this factor analysis showed that synthetic negation hovered around a score of 2.5 for most varieties, but revealed a lower score of 0.7 for personal letters. Analytic negation showed more variability among text varieties, with the highest being a score of 18.5 for face-to-face conversations, and the lowest being 3.4 for official documents.

The factor score for each of the linguistic variables is computed for the corpus in the corpus, with factor scores showing a magnitude higher than two being flagged for significance. Statistics from the corpus reveal the frequency per 100 tokens for all linguistic variables found within each document of the input corpus.

In a corpus, the dimension scores for each text are calculated using the factor scores for variables with a mean greater than $|0.40|$. The average of these scores is then used to produce the dimension scores for an entire corpus.

3.1.3 Dimensions and Text Types

The six dimensions outlined in Biber (1988) are largely based on contrasting styles, such as informational or involved production, narrative or non-narrative, and explicit or situation-dependent reference. These dimensions provide the basis for comparison of corpora in MDA. As the dimension scores will be discussed in greater detail in later chapters, this section will only serve as a brief introduction to the core concepts used for analysis.

Dimension one represents the contrast between involved and informational discourse, where low scores indicate information density with higher frequencies of nouns and adjectives. An example of a text which would receive a low score on this dimension is an academic journal, while a personal journal would receive a high score.

Dimension two is the opposition between narrative and non-narrative, where narrative texts will display higher frequencies of third person pronouns and past tense. Low scores indicate that a text is non-narrative, while high scores entail the opposite. A novel would receive a higher score on this dimension, whereas a math textbook would receive a very low score.

Dimension three presents the contrast between context dependent and context independent discourse, where a high score in this dimension would indicate a text that is not dependent on context and shows a higher frequency of nominalization, such as academic prose. The fast-paced commentary of a World Cup soccer announcer is an example of discourse which would score very low on this dimension, whereas top-secret documents from the CIA would be scored high.

Dimension four measures a text's overt expression of persuasion, where higher scores assert that the author explicitly marks their own point of view and assessment of certainty by incorporating more modal verbs in the discourse. Opinion articles on news sites would likely receive a higher score on this dimension, whereas an instruction manual would receive a low score.

Dimension five indicates the contrast between abstract and non-abstract information, where high scores indicate that information is presented in a technical and abstract manner. High scores in dimension five are associated with the use of passive clauses and conjuncts. A book on the topic of blackholes would be scored high for this dimension, while an online blog post on the antics of one's cat would probably be scored low.

The final dimension, dimension six, is not typically relevant for most corpora. It measures on-line informational elaboration, where a high score would indicate informational discourse produced under time constraints, such as debates. Texts with a high score in dimension six demonstrate increased usage in postmodifications of noun phrases.

In addition to the six dimensions of text, eight text types, which include multiple varieties of comparably structured texts, were established by Biber (1988) based on the dimension scores of a corpus. These eight text types, as well as examples of associated registers, are presented in Table 3.1.

Text Type	Dimension Scores	Examples
Intimate Interpersonal Interaction	High score D1 Low score D3, D5	Telephone conversations, Intimate conversation
Informational Interaction	High score D1 Low score D3, D5	Interaction in person, Personal letter
Scientific Exposition	High score D3, D5 Low score D1	Academic writing, Official documents
Learned Exposition	High score D3, D5 Low score D1	Academic writing, Press review
Imaginative Narrative	High score D2 Low score D3	Fiction, Prepared speech
General Narrative Exposition	High score D1 Low score D2	Editorial, Biography
Situated Reportage	Low score D3, D4	Broadcast
Involved Persuasion	High score D4	Interviews, Spontaneous speeches

Table 3.1: Biber's eight text types, their typical dimension scores, and examples.

The first text type 'intimate interpersonal interaction' is characterized by high scores on dimension one and low scores on dimensions three and five. Texts of this text type are generally interactions with interpersonal chatacterisitics, such as conversations between acquaintances or friends.

‘Informational interaction’, also characterized by high scores on dimension one and low scores on dimensions three and five, are typically spoken intimate interactions focused on conveying information.

‘Scientific Exposition’, marked by low scores on dimension one and high scores on dimensions three and five, are formal and informational compositions focused on conveying information in a technical manner. As the name suggests, academic research is a prime example of an associated register.

‘Learned Exposition’, which also has low scores on dimension one and high scores on dimensions three and five, is very similar to scientific exposition. Learned exposition focuses on conveying information formally, but lacks the technical aspect of the previous text type.

‘Imaginative Narrative’, which displays high scores on dimension two and low scores on dimension three, is indicative of texts which are highly narrative. This is the text type associated with most varieties of fiction.

‘General Narrative Exposition’, marked by low scores on dimension one and high scores on dimension two, uses narrative elements to present information. Registers which are narrative, but non-fiction, are typically associated with this text type.

‘Situated Reportage’, characterized by low scores on dimensions three and four, relates to live commentaries of events that are currently taking place at the time of utterance. This is the text type with the most limited application, as it only applies to transcripts of live broadcasts.

The final text type ‘Involved Persuasion’ only shows high scores on dimension four. Discourse within this text type is persuasive and argumentative, typical of the language used in speeches or debate.

3.1.4 MDA in Application

As MDA is used to measure variation among registers, it is a beneficial tool for corpus linguistics. While the previous section explained the fundamentals of MDA and how it is calculated, it is more practical to observe how this approach is applied to specific discourse-based tasks.

In later research, Biber and Egbert (2016) broadened the MDA framework to apply to texts found on the searchable web. As an improvement on the previous framework, hybrid registers were included. The classifications of registers within the expanded framework were general, 2-way hybrid, or 3-way hybrid. The general registers included in the analysis were named narrative, informational description, opinion, interactive discussion, how-to instructional, informational persuasion, or lyrical/spoken, with major and non-major hybrid registers being a combination or two or three of these general registers. Major hybrid registers are described as those which have dimension scores intermediate between the general registers. With a corpus comprised of texts from roughly 53,424 URLs, this new MDA framework was employed to assess differences in the discourse of these web-based registers.

Analysis of the corpus showed that 70% of the documents fell within one general register, with 11% classified as a two way hybrid, and 14% classified as a three way hybrid. The remaining 5% of the documents were classified as non-major hybrids.

The MDA approach has also been used in research on World Englishes. Using five corpora from the International Corpus of English (ICE), Great Britain (ICE-GB), Hong Kong (ICE-HK), India (ICE-IN), the Philippines (ICE-PH), and Singapore (ICE-SG), Xiao (2009) tested an enhanced MDA model in the research of World Englishes. The elaborated MDA employed semantic components with the inclusion of the University Centre for Computer Corpus Research on Language (UCREL) semantic tagger. Categorization of variation was also expanded to twelve registers reflecting the varieties of English represented in the ICE. Nine factors, corresponding to the dimensions of Biber (1988), were devised based on groupings of salient variables. These factors are to be differentiated from the factor scores used to determine the dimensions in Biber (1988), as they are used to be the equivalent of the six dimensions. Analysis showed that the English varieties used in Southeast Asia behave similarly in factor one. In factor two, British English exhibits a generally higher score than other varieties of English. Factor three did not present any significant variation among the varieties of English. Factor four showed that Indian English and British English presented similarly. For factor five, British English received the highest score in written registers, non-printed correspondence, and private conversation, while Indian English showed the lowest scores in general.

The multidimensional analysis tagger (MAT) developed by Nini (2015), expands upon the original six dimensions defined by Biber (1988) by relabeling the traditional linguistic features and incorporating additional features, such as indefinite pronouns and quantifiers. As with the original MDA framework, Nini (2015) has also employed the six dimensions representing different styles of discourse and demonstrating correlation with certain linguistic features. While the original Biber MDA framework utilized factor scores to determine the dimension score of a document, Nini uses z-scores for each linguistic feature to establish a relationship with the dimensions. The dimension scores for the inputted corpus are then compared using euclidean distance with the original text types determined by Biber (1988) to measure the text type it most closely resembles. With these, the user is able to thoroughly examine and decipher the corpus, inspecting the results for any significant variation. The tool developed by Nini (2015), the multidimensional analysis tagger, will be the one which is utilized in this thesis.

3.2 Data

The exploration of variation in register and text type between authentic and deceptive sources requires an appropriately robust dataset. Previous datasets tackling the topic of deceptive news were found to be either incomplete or missing information, such as the

Kaggle dataset 'Getting Real About Fake News' (Risdal, 2016), or unsuitable for this task, such as the 'Liar, Liar Pants on Fire' dataset (W. Y. Wang, 2017) which made use of fact checked statements instead of full articles. A novel dataset of significant size for both deceptive and credible news was created.

A selection of deceitful and reputable sources were scraped and cleaned to provide information including article content, headline, author, date, and tags, depending on availability. All articles are written in English, containing content pertaining to North America, particularly the United States of America and Canada. All of the articles are fairly recent, having been written within the time period of 2013 to 2017. Articles were preprocessed to remove links to Facebook and Twitter, non-text characters such as emoji, as well as to disambiguate punctuation and correct line spacing.

Four sources were included in the corpora for reputable news. Articles were taken from the magazine Vox (Zheleva, 2017), the New York Times, USA Today, and the New Yorker. The deceptive sources were selected from a curated list of biased, satirical, and false news sources and based on which sites were the most recently active, or which had the largest available content (Hagen, 2017). Example 1 is an excerpt from an article from USA Today, included within the credible news corpus:

(1) "*The Obama administration imposed sanctions Thursday on Hamza bin Laden, a son of Sept. 11 mastermind Osama bin Laden, saying he poses a risk to national security in the United States. The State Department said in a statement that the younger bin Laden was added to its Specially Designated Global Terrorist list after he was determined to have committed, or pose a serious risk of committing, acts of terrorism that threaten the security of U.S. nationals or the national security. The sanctions will deny him access to the U.S. financial system, the State Department said. Hamza bin Laden was officially named an al-Qaeda member in 2014 by his father's successor, Ayman al-Zawahiri. In an audio message in 2016, Hamza bin Laden threatened revenge against the U.S. and warned Americans they would be targeted at home and abroad. In a 2015 audio message from al-Zawahiri, Hamza bin Laden called for acts of terrorism in Western capitals, according to the State Department. Hamza bin Laden also has called for lone wolf attacks against the United States, France and Israel.*" –**USA Today (Durando, 2017)**

Many deceptive news sites tend to become defunct in a relatively short time period, with a great number of websites active during the 2016 U.S. presidential election already being out of commission. Due to the ephemeral nature of deceptive news sources, it was not possible to acquire a sufficiently large dataset from any individual source. To ensure that the corpora were relatively balanced, it was necessary to include a greater variety of sources in the deceptive news corpus.

Three sources, the Beaverton, the Borowitz Report, and Clickhole, are well known satirical news sites. Six sources, Liberty Writers, Empire News, If You Only News, Mad World,

Viralactions, and Free Thought, claim to be satirical, but are not well known and do not state this fact obviously or clearly. The remaining source, Breitbart, is a right wing biased media source. Example 2 is an excerpt from a Beaverton article, included within the deceptive news corpus:

(2) "BRASILIA - *Amidst a presidential impeachment, numerous public health crises, incomplete infrastructure, and a wildfire that rages throughout the country, Brazilian officials assured the world that Rio will be 'only moderately on fire' by the time they host the Olympic Games in August. Acting president Michel Temer, who assumed office last month after the Brazilian Senate impeached president Dilma Rousseff for manipulating government figures, assured reporters that "curtailing this raging fire will be very near to the top of our 'to-do list'. Obviously we cannot hold a world class sporting event while the nation of Brazil is wholly engulfed in flames" he conceded, actively working to avoid the currently burning podium and floor.*" – **the Beaverton (MacIntyre, 2016)**

In total, over 81,000 articles were collected. The total number of articles collected for the credible news corpus was 46,165, and a total of 35,478 for the deceptive news corpus. This is currently the largest collection of deceptive news articles and one of the largest collections of credible news articles. In comparison to datasets used in previous research, the dataset collected for this thesis is more robust considering both size and scope. In terms of size, this dataset is larger than all of the three datasets collected from Buzzfeed, political news stories, and Burfoot and Baldwin (2009) used in Horne and Adali (2017). It is also much larger than those used in Rashkin et al. (2017) and Pérez-Rósas et al. (2017). The closest datasets in size are those of Shao et al. (2017) and Bajaj (2017), with the latter containing 63,000 articles. Considering scope, this dataset is more inclusive than those which only focus on data from Twitter, such as Volkova et al. (2017) and Zubiaga et al. (2016). It is also more substantial than datasets based on fact checking of short statements, as seen with the datasets utilized in Pérez-Rósas and Mihalcea (2015) and W. Y. Wang (2017). Tables 3.2 and 3.3 illustrate the break down of articles for each source in the two corpora.

Source	Corpus	Articles
The New Yorker	Credible	10,035
The New York Times	Credible	11,211
USA Today	Credible	1,907
Vox	Credible	23,012
Total		46,165

Table 3.2: List of sources used in the credible news corpus, the number of articles per source, and total.

Source	Corpus	Articles
The Beaverton	Deceptive	1,589
The Borowitz Report	Deceptive	749
Breitbart	Deceptive	15,303
Clickhole	Deceptive	883
Free Thought	Deceptive	1,136
Liberty Writers	Deceptive	4,583
Mad World	Deceptive	2,970
Viralactions	Deceptive	2,048
If You Only News	Deceptive	5,689
Empire News	Deceptive	928
Total		35,878

Table 3.3: List of sources used in the deceptive news corpus, the number of articles per source, and total.

To make each the corpora equal in size for fair comparison, articles from the larger dataset were randomly removed until they matched the size of the smaller. As with any sufficiently large dataset, outliers are inevitable. After tagging each corpus, the results for both dimension scores and z-scores were investigated for problematic outliers. Most of the outliers were due to incorrectly labeled articles, such as videos with captions being included in hard news, or to javascript which escaped cleaning. Any articles which resulted in unwarranted scores were removed from the analysis prior to statistical testing so as not to affect the results. Articles of considerable brevity, due to their tendency to produce inflated scores, were also removed. After removing erroneous data and short articles from both corpora, articles from the larger corpus were once again randomly removed to ensure equality. Ultimately, 60,000 articles were used in the analysis. These adjustments put this corpus on a similar scale as that used in Bajaj (2017). However, the corpus used by Bajaj (2017) is not balanced, with only 13,000 fake articles compared to 50,000 credible articles.

3.3 Analysis

The extent to which misinformation in media differentiates itself from reputable sources, and whether or not such differences can be measured statistically and classified automatically are questions which will be addressed by this research.

The analysis in this thesis was conducted with the results from the MAT tagger, which is based on the original framework established in Biber (1988) and developed by Nini (2015). To account for articles deemed inadequate for analysis, being either too short or improperly tagged by the website, and instrument error, the full 81,000 articles were tagged.

The tagging was conducted with the Windows version of the MAT software, version 1.3. The software builds on the preliminary grammatical analysis and part of speech tagging of the Stanford Tagger (2013) by adding additional features from Biber (1988) (D. Chen &

Manning, 2014; De Marneffe, MacCartney, Manning, et al., 2006). Tagging both corpora took a total of 48 hours. The tagger takes a corpus as input, then iterates throughout the documents in that corpus. Once the corpus has been tagged, it can be analyzed to produce dimension scores. A final 24 hours was required for the multidimensional analysis, totaling 72 hours for the entire process.

After tagging and removal of outliers due to instrument error and inadequacy, the credible news corpus and deceptive news corpus were made equal. This was done by randomly eliminating articles from the larger corpus, credible news, to match the number of articles in the smaller corpus, deceptive news. Ultimately, 60,000 articles were used for computing the scores and determining significance.

Example 3 is the first sentence from a Beaverton article. It shows what a single tagged sentence from a document looks like after being tagged for linguistic features, such as part of speech.

(3) `Amidst_PIN a_DT presidential_JJ impeachment_NOMZ ,_, numerous_JJ public_JJ health_NN crises_NN ,_, incomplete_JJ infrastructure_NN ,_, and_ANDC a_DT wildfire_NN that_TSUB rages_VPRT throughout_PIN the_DT country_NN ,_, Brazilian_JJ officials_NN assured_VBD the_DT world_NN that_TOBJ Rio_NN will_PRMD be_VB '_' only_RB moderately_RB on_PIN fire_NN '._.'`

The results from both corpora have been tested for statistical significance in the differences between dimension scores and the differences between z-scores of linguistic variables with a Wilcoxon signed-rank test. Differences between z-scores have also been examined with least significant difference (LSD), using results from the LSD threshold matrix to determine inclusion in the discussion.

An analysis of correlations between variables has also been conducted using Pearson's correlation coefficient, represented with correlation matrices, to determine any major dissimilarity or similarity in behaviour between linguistic features across corpora.

In the following chapter, the results from the MDA analysis of both corpora will be analyzed and discussed in detail. The results of each corpus will be compared with each other on several fronts. First, the corpora will be compared over five of Biber's dimensions of text. Secondly, they will be compared based on salient linguistic variables. Finally, the corpora will be compared with established varieties of text, and with other news types.

Chapter 4

Results and Discussion

In this chapter, I will present the results from comparing the discourse of deceptive news and credible news using MDA and assess correlations between salient linguistic variables with Pearson's correlation. The corpora will also be compared to other varieties of text, as well as with other types of news.

First, the results of comparison between the five dimension scores for each corpus will be discussed. Correlations between dimension scores have also been examined for each corpus, revealing that the dimensions were mostly independent of each other for both corpora. This independence reveals that no dimension had any strong influence over any other dimension, positive or negative, for both corpora.

In the second part of this chapter, both significant and interesting z-scores were compared, demonstrating variation in the application of linguistic variables. Correlations between linguistic cues have also been inspected for intriguing relationships within each corpus, revealing patterns of usage unique to each register. Finally, each corpus was compared to the text types described in Section 3.1.3.

The results and discussion presented in this chapter will show that significant differences exist between the discourse of deceptive and credible news. These differences will provide evidence to promote a conclusion calling for separation of these news types based on their being distinct registers. Before delving into in-depth analysis of the results, there are some general observations that are important to note.

4.1 General Findings

Although the standard deviation for both corpora was high, it was much higher for the deceptive news corpus across all dimension scores. A high score on the standard deviation indicates a greater degree of variability within the corpus, which is unsurprising considering the corpus for deceptive news required more sources than its credible counterpart. Articles from credible news sources are typically simple to obtain; the sites experience a long lifespan

with a rich article history, often have an accessible API useful for collecting content, and boast a reputation for respectability which makes them easily identifiable as credible sources.

Sources for deceptive news do not usually possess the same merits. With the exception of some satirical sources, they are often short-lived, practice incoherent web encoding, and are ordinarily only known and visited by a small user base. On a site such as The New York Times, one can acquire more than 10,000 articles on politics alone, whereas from a site such as Empire News, the number is limited to fewer than 1,000.

4.1.1 Statistical Significance of the Differences between Dimension Scores and Z-Scores

An important detail to include in this section is the result of significance testing on the differences between the mean dimension scores across corpora and mean z-scores across corpora. The differences between the means of both corpora for all dimension scores and the differences between the z-scores for all linguistic features for both corpora were tested for significance with a Wilcoxon signed-rank test and Least Significant Difference (LSD) Student's t.

The p-values returned for all significance tests in the differences between the dimension score means and the z-scores means between the credible and deceptive news corpus were $p < 0.0001$, allowing for a rejection of the null hypothesis. The null hypothesis assumes that no significant difference exists between specified populations, any observed difference being due to sampling or experimental error. Rejection of the null hypothesis indicates that there are grounds for the belief that a relationship between two phenomena is present and not due to random chance. This test was applied to all differences presented within the results section of this thesis.

Given the size of the dataset, a small p-value is expected. With any sample of considerable size, even the slightest difference between the samples is likely to be considered significant. As the results across all categories are identical, $p<0.0001$, there is no need to go into greater detail. As this chapter progresses, it is essential to bear in mind that the differences presented between dimension scores for the corpora and the z-scores for each variable between corpora retain a meaningful difference.

4.2 Explanation of Scoring

Prior to discussing the results of this analysis, it is important to explain how these scores are calculated and how they can be interpreted. Calculation of the dimension scores will be explained, followed by the calculation of the z-scores.

4.2.1 Dimension Scores

Dimensions are described by Biber (1988) as "bundles of linguistic features which co-occur in texts because they work together to perform some underlying function". These co-occurring features are referred to as 'factors', which are identified through factor-analysis. For example, necessity modals and predictive modals are factors of dimension four, meaning their frequencies are used in the calculation of the dimension score for this dimension. In the current analysis, these are represented by z-scores. The dimension scores are computed by summing the frequencies of features with positive loadings and subtracting the frequencies of features with negative loadings. Positive loadings are described in Biber (1988) and Biber (1995) as factors which receive a score of $|0.4|$ or greater, with factors outside this range being considered of little importance whether or not they are significant. This cutoff was based on the mathematics of factor analysis, with the standard being 0.7, but often reduced to 0.4 based on the researchers judgment (Gorsuch, 1983).

There are six dimensions outlines in Biber and MAT, though only five of these are applicable to this analysis. The sixth dimension, which measures discourse produced under time constraints, is only appropriate for speeches and debate. As the element of time is not really a factor in written news, it will not be included in the discussion.

4.2.2 Z-Scores

There are 67 linguistics variables included in this analysis. Of the linguistic variables, 64 were initially established by Biber (1988), while the remaining three tags were introduced by Nini (2015) to enhance the original variables. The score for split infinitives has not been included in this analysis due to instrument error, making the total variable count 66.

The z-scores produced by MAT are the equivalent of the factors referenced in the original Biber. In statistical terms, z-scores are the number of standard deviations from the mean value of the reference population, where positive scores indicate a result above the mean and negative scores indicate a result below the mean. Any z-scores with a magnitude greater than 1.96 are considered variables of interest, and are flagged. Normally, 95% of z-scores will fall between -1.96 and 1.96, signifying that scores of greater magnitude are of particular importance. The absolute value of the z-score reveals how many standard deviations the variable is from the mean. If a z-score is equal to 0, it is equal to the mean. For example, a z-score of 0.5 is considered above the average, while -0.5 would be below. The p-value associated with a 95% confidence level is 0.05. If the z-score is between -1.96 and 1.96, its p-value will be larger than 0.05, and the null hypothsis cannot be rejected. While a score between -1.96 and 1.96 is not as far from the mean, it may still be worth discussing. Within the context of the multidimensional analysis tagger, z-scores are used to signify the importance of a variable in a text. They consider the frequency of the variable relative to the number of words in the entire body. These z-scores are also used to calculate the dimension

scores, which makes them particularly relevant to the discussion. Furthermore, variation above or below the average and the differences in this variation between corpora provide an idea of how these linguistic features are used. While the feature may not be significant to the general corpus, its patterns of usage may paint some picture of the writing of deceptive news as opposed to credible news.

4.3 Dimension Score Results

The following sections present the results for each dimension, and discuss the implications of these results while exploring possible explanations. The mean dimension score for dimensions one through five are included in Table 4.1, organized by corpus.

Dimension	Deceptive	Credible
D1: Involved vs. Informational	-9.7	-10.4
D2: Narrative vs. Non-narrative	1.3	0.3
D3: Context Dependency	2.6	2.8
D4: Overt Expression of Persuasion	-0.8	-1.1
D5: Abstractness	0.3	-0.1

Table 4.1: Mean dimension scores for dimension one through five. These numbers represent the average dimension score across all articles within the corpus.

As previously mentioned, statistical analysis has shown that the difference in the mean scores for each dimension presented in 4.1 are significant, with $p < 0.0001$. The largest difference between these means is found in dimension two, which measures narrative discourse. The smallest difference is found in dimension three, which measures context dependency. From these preliminary scores, it can be inferred that the average credible news article is more informational, less narrative, slightly less dependent on context, less opinionated, and less abstract in comparison to the average deceptive news article.

4.3.1 Dimension One: Involved vs. Informational

Dimension one measures the opposition between involved and informational discourse. Informational density is indicated by negative scores in this dimension. Positive scores indicate affective and interactive text, similar to conversation. Information for dimension one, including the dimension scores and salient features are presented in Table 4.2. The dimension scores are calculated as described in Section 4.2.1, while the z-scores are calculated as described in Section 4.2.2.

Both the deceptive news and credible news corpus received substantially low scores in dimension one, revealing that both corpora contain an abundance of information. This is similar to mediums such as academic prose, press reportage, and science fiction within the archetypes of general narrative exposition and scientific exposition proposed by Biber

	Deceptive	Credible
Dimension Score	-9.7	-10.4
Nouns	3.9	3.1
Attributive Adjectives	-0.1	0.5
Predicative Adjectives	0.3	0.9
Average Word Length	0.6	0.8

Table 4.2: Mean dimension score per corpus and mean z-scores per corpus of features associated with dimension one.

(1988) and described in Nini (2015). However, the mean dimension score for the credible news corpus, -10.4, is noticeably lower than the deceptive news corpus, -9.7. In comparison to the scores for press reportage and editorial reported in Biber (1988), these scores fall within the normal range for dimension one. This difference is both visible and proven with significance testing, as mentioned in Section 4.1. Unsurprisingly, this suggests that credible news articles display more informational density than deceptive articles.

Low scores in dimension one indicate a higher frequency of nouns, adjectives, and long words. In the case of nouns, there is a moderate difference between the frequency of general nouns in the deceptive and credible news corpus. While z-scores revealed that nouns were a prominent feature of both corpora, a z-score of 3.9 for deceptive news and 3.1 for credible news implies that noun usage is more excessive in deceptive news. Referring back to Biber (1988), a high frequency of noun is expected, as it appears that nouns are also frequent in press editorial, press review and broadcast. A possible explanation for this result is that deceptive news focuses on the actions and relationships between individuals, often repeating proper nouns or including them when unnecessary. This preservation of proper nouns is also seen in passage 5, where 'Trump' is repeated several times within the short passage. As the tagger classified all nouns, including proper nouns, as one type and pronouns separately, repetition of proper nouns may lead to a higher noun score. The following passage, which received a z-score of 6.3, is an example from an article which has been marked for a high frequency of total nouns.

(4) "*HALIFAX - Four hours after a striking a massive iceberg, the Province of Newfoundland has sunk off the coast of Labrador leaving an unknown number of survivors. All contact was lost with Canada's most-easterly province early this morning. Skipper Dwight Ball, who was reportedly at the helm in St. John's, tried to steer Newfoundland hard to starboard when they came too close to a mountain of ice, but it was too late. The 520,000 residents on-board the large landmass were awoken by a large crash at 11:49 PM NT and the land quickly took on water. A distress signal was sent to neighbouring PEI, but everyone was in bed by that time.*" –**the Beaverton (Huntley, 2017)**

Regarding the use of adjectives, credible news appears to be more colourful in its writing. Both attributive and predicative adjectives are employed at a grander scale in the writing of credible articles compared to deceptive news. A z-score of 0.5 for attributive and 0.9 for predicative adjectives in credible news articles confirms that the feature is somewhat salient, whereas a negative z-score of -0.1 for attributive and a lower z-score of 0.3 for predicative adjective reveals that the language of fabricated articles is less adjectivally rich. In contrast, scores for press editorial and reportage in Biber (1988) suggest the opposite, that attributive adjectives are more frequent than predicative. This may reflect a shift from 1988 to 2017 in the use of adjectives in news discourse, or it may be a characteristic of online news in comparison to print. Passage 5, which received a z-score of -3.2 for attributive adjectives, is an example from an article which has been marked for low frequency of attributive adjectives.

(5) "*President Trump is definitely one of a kind. His love of country is unlike anything we have ever witnessed in history. But Trump took it a step further today while speaking in Orlando. Right in the middle of his speech, Trump spotted a supporter who he recognized from being interviewed in the media. The man, being identified as Gene Huber, arrived at the venue at 4am and waited all day to meet our president. Many thought that Trump was just going to give him a wave, but not this president. To the surprise of everyone, especially the Secret Service, president Trump invited the supporter to the stage to speak to the people. The supporter was so shook up and even looked to be on the verge of tears. He came on stage, hugged our president and was so grateful! You could see the love in his eyes!*" – **Liberty Writers (Lindsey, 2017)**

Concerning average word length, there is a marginal difference between deceptive news and legitimate news. Although the saliency for average word length for both corpora was positive, indicating that the feature was not underused in either corpus, the credible news corpus received a z-score of 0.8, while the deceptive news corpus received 0.6. In comparison to Biber (1988), these scores are expected. Word length was one of the least volatile features across corpora, with the largest range for scores being 0.3 points. The z-score for average word length is slightly higher in credible news than it is to deceptive news, suggesting that the average word length of articles produced by credible sources is greater. It is likely that deceptive news is deliberately opting for simplistic word usage in effort to be more approachable to a larger reader base and is asserting itself as unpretentious in opposition to traditional media. Readability measures conducted on deceptive news, including misinformation and satirical, showed that deceptive writing is more easily comprehended than writing from credible and truthful sources (Afroz, Brennan, & Greenstadt, 2012; Frank et al., 2008; Yang et al., 2017).

Dimension one has revealed that the content of reliable news exhibits more density of information compared to deceptive news. Upon further inspection into linguistic features

pertinent to the dimension, the presence of nouns is more defining of deceptive articles, while generous use of adjectives and longer average word length is more frequent in the content of legitimate news. From these results, it is suggested that deceptive news introduces and makes references to more entities, whereas credible news is more descriptive of the entities it has introduced.

4.3.2 Dimension Two: Narrative vs. Non-narrative

Dimension two represents the contrast between narrative and non-narrative discourse, where high scores are indicative of text which is narrative in nature. Information for dimension two, including the dimension scores and salient features are presented in Table 4.3.

	Deceptive	Credible
Dimension Score	1.3	0.3
Past Tense	0.2	-0.2
Present Tense	-0.6	-0.6
'Seem' and 'appear'	-0.1	0.3
Third Person Pronoun	-0.1	-0.4

Table 4.3: Mean dimension score per corpus and the mean z-scores of features associated with dimension two.

The mean score in dimension two shows the highest degree of disparity between the two corpora. The deceptive news corpus produced a positive mean dimension score of 1.3, compared to the credible news corpus which yielded only a slightly positive mean score of 0.3. In comparison to Biber (1988), the credible news corpus is scored similarly to press reportage, while deceptive news is scored more similarly to general fiction. The implication is that the element of narrative is more distinctly associated with deceptive news. As fabrication is fiction, there is some aspect of storytelling involved in weaving deceptive information into a narrative. With legitimate news sources, the facts guide the narrative and leave less leeway for improvisation. The passage is an example of a satirical article with a very high degree of narration, as evidenced by its dimension score of 13.6 on dimension two.

(6) "*WASHINGTON (The Borowitz Report) - Former Secretary of State Hillary Clinton spent several hours at the United States Capitol on Thursday compiling a mental list of people she will destroy at a later date, an aide to Clinton has confirmed. Clinton gave no outward appearance of compiling such a list as she answered questions relating to her tenure as Secretary of State, the aide said, but was busy assembling the list nonetheless. "This is the kind of multitasking that she is very good at," he said. "Believe me, the entire time she was talking, she was working very hard on that list." In response to reporters' questions, the aide said that there was "no firm timetable" for Clinton to destroy the people on her list. "She will wait for the appropriate time," the aide said, "and she will crush them." The aide would neither confirm nor deny that Representative Trey Gowdy, Republican of South Carolina, was at the top of Clinton's list. "All I can*

say is that, after this morning, there are a lot of people on that list," he said. "A lot of people"."– **the Borowitz Report (Borowitz, 2017)**

Passage 7 is an example of a hard news article from the same publisher as the previous passage with a low degree of narration, as shown by its dimension score of -6.7 on dimension two.

(7) "*Sixteen years, nine months, and five days. That's how long it took advocates of same-sex marriage to defeat the Federal Defense of Marriage Act, from the day it was signed into law, by President Bill Clinton, in 1996, until Wednesday, when the Supreme Court ruled, by a vote of 5-4, that its key section is unconstitutional and that the federal government must recognize the state-sanctioned marriages of same-sex couples. The Court also put aside another case, Hollingsworth v. Perry, a decision that will have the effect of bringing same-sex marriage to California.*"– **The New Yorker, 2017**

The linguistic features most closely associated with this dimension are the past tense and third person pronoun. Past tense is a common feature in narrative, as suggested by the high scores for fiction related genres in Biber (1988). As is consistent with the dimension score, the deceptive news corpus revealed a more prominent usage of the past tense with a z-score of 0.2, compared to the -0.2 scored on the credible news corpus. In comparison to press reviews and editorial in Biber (1988), this is unexpected. Credible news reported the past tense as slightly infrequent, which is inconsistent with Biber. This may be due to differences in online media compared to print, or indicative of a shift in discourse patterns, as the present tense displays the same phenomenon of being reportedly underused in these corpora compared to Biber. Despite the past tense being a slightly underused variable in the legitimate articles, the present tense is not more frequent in contrast, as both the deceptive and credible corpus acquired a z-score of -0.6. In a quest to discover which verbal construction is more noticeable in the legitimate corpus, 'gerunds', the passive, 'by' passive, perfect, main verb 'be', and the verbs 'seem' and 'appear' were also considered. If information is not conveyed through either the present tense or past tense, then perhaps it is conveyed through one of these structures. However, most of these other constructions were also slightly underused in the corpora. In the credible news corpus, the exceptional case was with the verbs 'seem' and 'appear'. The aforementioned raising verbs received the only positive z-score, 0.3, out of all verbal construction for credible news, while the deceptive corpus received a negative score of -0.1.

The third person pronoun was not a very prominent variable in either of the corpora, though slightly more common in deceptive news articles than credible news articles. Third person pronouns received a z-score of -0.4 based on the content of reliable articles, but -0.1 with the content of deceptive news. As seen in dimension one, nouns were found to

have a higher frequency in deceptive articles. An explanation posited for this finding was the repetition of proper nouns. If this explanation is assumed, then one would expect a lower frequency of third person pronouns in comparison to credible news. However, as seen in Examples 5 and 6, deceptive news often replaces proper nouns with nouns indicating profession or position, such as 'the president' or 'the aide'.

Dimension two has illustrated that deceptive news embodies more of the stereotypical characteristics of narrative. The usage of past tense suggests that the content of deceptive articles tends to present information with past certainty, compared to reputable articles which present information with the expression of uncertainty. Reporting on events currently in progress may also contribute to a higher frequency of present tense in credible articles, as fabricated events typically do not occur in real time. The passages below present examples of the difference in the usage of entities and pronouns between deceptive and credible news. Example 8 is a deceptive article which contains a high frequency of third person pronouns, having received a z-score of 3.6. Example 9, in contrast to the deceptive article, received a z-score of -1.3. The more prevalent utilization of the third person pronoun conveys more of a preference for pronoun usage in deceptive news in comparison to credible news, although the use of pronouns seems to be generally infrequent in news discourse.

(8) "*Wisconsin child has accused her father of choking her and ordering her to eat cockroaches and a dead mouse. The 9-year-old Madison area girl testified Wednesday during her father's child abuse trial in Dane County that she was starved at home and ordered by her father to also eat other insects. The girl also said her father put a large plastic bag on her naked body and used duct tape around her waist before she was able to free herself. The Wisconsin State Journal reports the girl said her father choked her at least twice. The 39-year-old Town of Madison man is charged with child abuse and sexual assault. He's pleaded not guilty. Her mother pleaded no contest last month to six counts of child abuse.*"– **Viralactions, 2016**

(9) "*Sen. Lindsey Graham, the latest Republican to join all the cool kids running for president, wants you to know that the world is a very scary place and that is why he should be president. "I want to be president to protect our nation that we all love so much from all threats foreign and domestic," Graham said in his announcement. Graham is one of the Senate's most distinguished hawks, but he isn't an outlier here. The entire Republican field, Rand Paul excepted, is trying to convince you that the world is dark and full of terrors: the United States is under threats at every turn, that our fundamental national security is at risk, and that the world is, as Graham once said, "literally about to blow up." The exact opposite is true. Today, the United States is actually enjoying a time of extraordinary safety: threats to the homeland are few and very far between. And while it's true that there are lots of bad things in the world, it's conflict-free as it's ever been, at nearly any point in human history. On this, the Republican candidates are just wrong.*"–**Vox News (Beauchamp, 2015)**

4.3.3 Dimension Three: Context Dependency

Dimension three evaluates the dependency of the text on context. Low scores mark texts which are dependent on the content of the discourse, whereas high scores demonstrate texts which stand alone. The dimension scores and salient features of dimension three are presented in Table 4.4.

	Deceptive	Credible
Dimension Score	2.6	2.8
Nominalization	0.1	0.2

Table 4.4: Mean dimension score per corpus and the mean z-score of the feature most closely associated with dimension three.

Both the deceptive news and authentic news corpus received similarly high scores in this dimension. The difference between the mean dimension score for the deceptive news corpus, 2.6, and the authentic news corpus, 2.8, was only 0.2 units. Although the two corpora are similar in their independence, a slightly higher score for the credible news corpus suggests a marginally wider separation from context in its content.

The linguistic variable associated with high scoring texts in this dimension is nominalization. When discussing legal and political concepts, nominalization is a common feature. Nominalizations, particularly when incorporated in passive structures, present a degree of bureaucracy and mystification (Fowler, 2013). They manipulate the interpretation of information by avoiding the inclusion of agency, such as using the word 'accusations' instead of 'x accused y'. While both corpora presented a positive z-score, the authentic news corpus was modestly higher, at 0.2, compared to the deceptive news corpus, at 0.1. Although this difference is not large, as mentioned in Section 4.1.1, it is significant. In comparison to Biber (1988), these scores are normal. Both press editorial and press review revealed a higher frequency of nominalization. The frequency of nominalization within each corpus contrasts with the frequency of total nouns seen in 4.3.1, where deceptive news exhibited more noun usage in comparison to credible news. As nominalization is a common grammatical tool for creating lengthy noun phrase subjects and reiteration of previous information, it is a pervasive feature of academic and scientific discourse (Martin, 1991; Schleppegrell, 2001). As demonstrated through the previous two dimensions, credible news tends to be more informational and less narrative in its discourse. The presence of nominalizations indicates a more impersonal and academic tone in the discussion of current events, and suggests the reiteration of information through nominalized forms. The following short article, which produced a z-score of 6.3, is an example of a text from the credible news corpus which has been marked for a high frequency of nominalizations.

(10) "*The Justice Department is expected to announce the findings of a 20-month investigation into the Cleveland Police Department's use of deadly force, training, discipline, recruitment, and relations with minority communities. The news comes less than two weeks after a Cleveland Police officer shot and killed Tamir Rice, a black 12-year-old who was carrying an airsoft gun, a toy, at the time of his death.*" –**Vox News (Lopez, 2014)**

The results from this dimension suggest that the discourse of both credible and deceptive news does not rely on context. Context independence seems to be a common feature of news discourse, which is exemplified by the positive score attributed to both corpora. Considering the quality of impartiality, credible news establishes a slightly impersonal tone in comparison to deceptive news.

4.3.4 Dimension Four: Overt Expression of Persuasion

Dimension four assesses a text's expression of persuasion. High scores in this dimension explicitly convey the author's opinion on certainty and point of view, as in the discourse of interviews or professional letters. Of the text types conceived by Biber (1988), high scores in this dimension are unique to involved persuasion. The dimension score for each corpus and salient features of dimension four are presented in Table 4.5.

	Deceptive	Credible
Dimension Score	-0.8	-1.1
Modals of Necessity	-0.5	-0.5
Modals of Possibility	-0.5	-0.1
Predictive Modals	-0.2	-0.1

Table 4.5: Mean dimension score per corpus and the mean z-scores of the features most closely associated with dimension four.

The results from the analysis of both corpora for dimension four were quite close. Both corpora received a negative mean dimension score, with -0.8 for the deceptive news corpus and -1.1 for the credible news corpus. This result is consistent with press review in Biber (1988). Although both corpora are relatively unmarked for persuasiveness and personal opinion in the content of their articles, credible news displays appreciably less expression of persuasion and opinion.

Modal verbs are categorized by function as necessity modals, possibility modals, or predictive modals. Upon inspecting the individual features, a difference was found between two of the three categories of modal. Modals of necessity, such as 'ought' or 'must', were the same for both corpora. A considerably negative z-score of -0.5 suggests that this is not a commonly used variable in press reportage in general.

Modals of possibility, such as 'can' or 'may', showed the greatest difference of the three between the corpora. The deceptive news corpus received a z-score of -0.5, while the credible news corpus received a score of -0.1. Although both corpora reported this as a less frequently observed variable, the content of deceptive news appears to present information with less speculation on the consequences and outcomes of events than that of credible news articles.

Predictive modals, such as 'will', 'would', or 'shall', exhibited a minor difference between the two corpora. While the scores for both corpora were negative, the deceptive news corpus received a z-score of -0.2, which conveys a lower frequency of the variable compared to the -0.1 of the credible corpus. It appears that the content of deceptive news is less likely to use language which implies speculation and possibility.

In comparison to the genres of press review and editorial in Biber (1988), lower frequency of all sub-categories of modal verbs is expected. In this regard, both deceptive news and credible news are consistent.

Although dimension four has shown that both corpora are generally lacking persuasive language, deceptive news seems to display mildly more expression of opinion and certainty. The content of deceptive news may communicate more personal opinion in an effort to convince readers of its content. Despite the lower frequency of modal usage in deceptive texts, the author's point of view is expressed. Concerning the findings of Rashkin et al. (2017), this is contradictory. Rashkin et al. (2017) noted that modal adverbs are more frequent in deceptive news, but it appears that they are not as pervasive as originally discovered. This expression of persuasion may be present in other variables. Given the nature of propaganda, characteristics related to manipulative writing are not unexpected. This measure of speculation and persuasion may be a factor worth considering in automatic detection tasks.

4.3.5 Dimension Five: Abstractness

Dimension five quantifies the abstractness of the information within a text. High scores in this dimension suggest that the text presents information in a technical or abstract manner, as in scientific prose and official documents. The dimension score and salient features of dimension five for each corpus are presented in Table 4.6.

	Deceptive	Credible
Dimension Score	0.3	-0.1
By-passive	0.1	0.0
Agentless Passive	-0.2	-0.5
Conjuncts	1.4	0.2

Table 4.6: Mean dimension score per corpus and the mean z-scores of the features most closely associated with dimension five.

Surprisingly, the deceptive news corpus received a higher mean score in this dimension, with 0.3, while the credible news corpus received a negative mean score of -0.1. Deceptive news articles are more analytical in their presentation, providing information with formality through the usage of passivization. Compared to deceptive news, the credible news sources presented information more colloquially. An explanation for this may be the intended target audience. Studies suggest that the elderly and those with decreased cognitive ability are more susceptible to deceptive news (Hambrick, 2018; Marzilli, 2018). As a traditional media style may be more familiar to this audience, deceptive news may adhere more strictly to this formality than credible news.

High scores in this dimension indicate substantial usage of passive clauses and conjuncts. As previously noted in the discussion of dimension two, passive clauses were more frequently employed by deceptive news articles. With by-passives, deceptive news was moderately more liberal in its usage with a z-score of 0.1, compared to the credible news corpus which received an unmarked score of 0. Once again, this difference is not large despite its significance. From these results, it can be assumed that deceptive news is slightly more likely to refer to the logical subject in passive clauses.

Both deceptive news and credible news were frugal with the employment of agentless passive clauses. Passive clauses were noticeably rarer in the credible news corpus, with a z-score of -0.5. The deceptive news corpus, while also demonstrating conservativeness with a negative z-score of -0.2, featured a more commonplace usage of the passive in comparison to credible articles. This lower frequency of agentless passives is also similar to press review in Biber.

Conjuncts include words such as 'altogether' or 'alternatively'. These conjuncts may appear as a device to connect two clauses, or sentence initially. The presence of conjuncts in the corpus presented the most striking disparity between the corpora. While both corpora received a positive score, implying a well utilized variable, the deceptive news corpus produced a score which was remarkably greater than the credible corpus. The saliency of conjuncts in the deceptive news corpus, with a z-score of 1.4, was more prominent than the credible news corpus, a score of 0.2. Inspecting the articles revealed that conjunctions were predominantly used sentence initially to connect a sentence with the previous one. As a consequence of the immense difference between the scores, it can reasonably be concluded that a higher frequency of conjuncts is a quality of deceptive news. The number of articles within the deceptive news corpus which received a high score for conjuncts was more numerous than the credible news corpus, with the amount of articles receiving low scores being minimal. The short article in Example 11 is a text from the deceptive news corpus which has been marked for a high frequency of conjunctions, receiving a score of 10.1.

(11) "*New York, NY - A disturbing video has surfaced which shows an NYPD cop kneeling over a man's lifeless body while puffing on a cigar. The video was uploaded to YouTube*

this week and it shows multiple NYPD cops standing around an unconscious man laid out on the pavement. Several moments pass before any of the officers attempt to tend to the man. The Free Thought Project reached out to both the NYPD and the video uploader. ***However****, we have yet to hear anything back from the uploader. The NYPD said they will look into the video,* ***however****, they were unable to confirm when and where this took place. It is unclear as to how or why the man is laying on the ground.* ***However****, he appears to need medical assistance and is receiving none. Aside from the sadistic nature of enjoying a cigar kneeling over the lifeless body of a man, NYPD policy prohibits smoking while in uniform. This policy appears to be entirely ineffective at preventing this officer's cigar-smoking,* ***however****."–***Free Thought (Agorist, 2017)**

On the opposite ends of the spectrum, articles from the credible news corpus received much lower scores, with the number of articles with high scores being few and lower in comparison to the deceptive corpus.

Dimension five has established that deceptive news articles possess a higher degree of formality, whereas credible articles maintain an element of intimacy. Although passives are not used superfluously in either corpus, deceptive news is more likely to mention the logical agent of passive clauses. However, conjuncts are emphatically employed in deceptive news. The exceptional usage of conjuncts in the deceptive corpus could be a form of emphasized imitation. When drawing caricatures, or doing impersonations, it is common to identify one of two distinct characteristics of the subject and accentuate them. Compared to the findings of Horne and Adali (2017), which noted fewer technical words in the writing of deceptive news, the higher degree of abstractness in the deceptive news corpus is unexpected. As the aim of satirical and fabricated news is to imitate the style of original, conjuncts may have been one of the linguistic features emphasized to promote authenticity in the imitation.

4.4 Correlations Amongst Dimension Scores

To examine the relations between dimensions further, the correlations between dimension in general and within each corpus were inspected. Using multivariate methods, a linear correlation measure was calculated to discover any curious relationships between dimension scores. The relationships between the variables are captured by the slope of the line and ellipse in Figures 4.1 and 4.2, where the density ellipse also contains the mass of the points determined by a 95% confidence interval. The axes are scaled linearly. Columns for each variable run horizontally and vertically, with the correlation score noted in each column where two unique variables intersect.

4.4.1 Dimension Correlations within the Deceptive News Corpus

The dimensions were largely independent from each other, bearing little to no correlation. The exceptional cases for the deceptive news corpus were the relationship between dimension one and dimension three, which represents the relationship between informational discourse and context dependency, and dimension one and dimension four, which represents the relationship between informational discourse and expression of persuasion.

Figure 4.1 displays the correlation between each of the dimensions in the deceptive news corpus. The r score and width of the ellipsis indicate the strength of the correlation, while the line of fit signifies the directionality. In the figure, the intersection of dimensions one and three shows a negative correlation score of r=-0.34, which signifies a moderately downhill correlation between the two dimensions. The downhill relationship is further marked by the downward slope of the line and density ellipse in the intersecting cells for dimension one and three of the matrix.

The implication of this relationship is that informational density and context dependency are slightly divergent in the discourse of deceptive news. As dimension three distinguishes texts which utilize explicit reference, opposed to situation dependent reference, this it not surprising. Dimension three can also be used to classify expository or situation-dependent discourse (Leistyna & Meyer, 2003). Expository discourse, for example scientific debate, includes more nominalization. Situation-dependent discourse displays a higher frequency of adverbs, particularly adverbs of time and place. As previously discussed, deceptive news has a very minimally positive nominalization score. In Appendix A.1, adverbials of time and place received a much lower score or were unmarked. These results indicate that deceptive news shows more similarity with expository discourse. As presenting information is the primary purpose of exposition, it is expected that as dependence on context rises, informational density should decrease. The divergent relationship seen between dimensions one and three is therefore completely expected for a text conveying information. Given this assumption, the same divergent relationship should also be present in the credible news corpus. As will be discussed in the following section, this is the case.

The correlation between dimension one and four revealed a relationship of the opposite nature. In Figure 4.1, the intersection of dimensions one and four shows a positive correlation score of r=+0.23, which connotes a meager uphill correlation. The uphill relationship is also visible in the upward slope of the line and density ellipse in the intersecting cells for dimension one and four of the matrix. This suggests that density of information and expression of persuasion are linked in the discourse of deceptive news articles. Language which implies conviction and certainty is likely a tactic which lends a tone of confidence to the information presented.

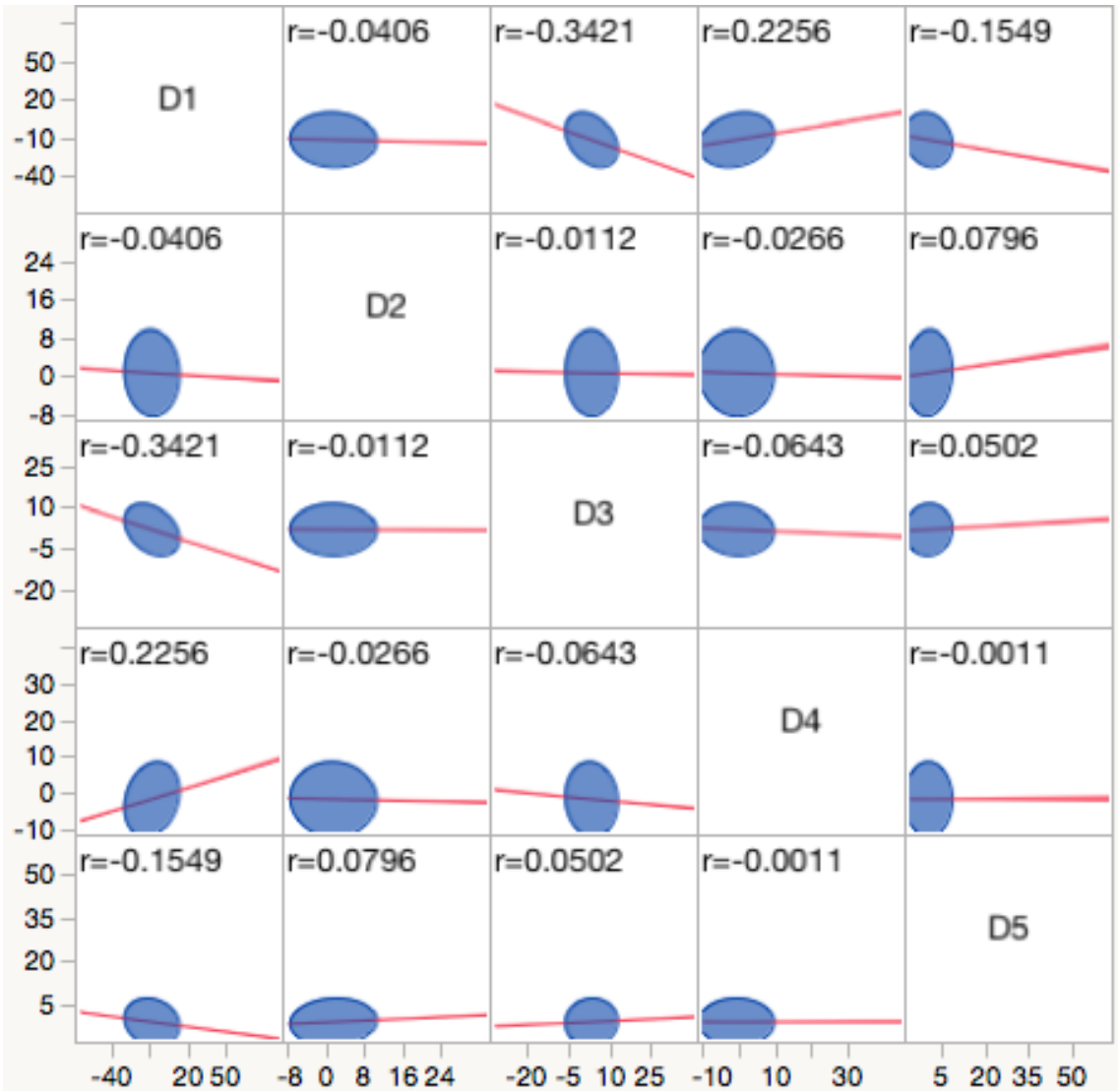

Figure 4.1: The results for the deceptive news corpus from the r correlation measure displayed by dimension.

4.4.2 Dimension Correlations within the Credible News Corpus

The correlation results for each dimension for the credible news corpus are displayed in Figure 4.2. Columns for each variable run horizontally and vertically, with the correlation score noted in each column where two unique variables intersect.

Within the credible news corpus, a few weak correlations exist between dimension one and other dimensions. It seems that there is also a slight downhill correlation between dimension one and three at r=-0.26, similar to the deceptive news corpus. The downhill relationship is also visible in the downward slope of the line and ellipse in the intersecting cells for dimension one and three of the matrix in 4.2. A slightly weaker score for this interaction suggests that context dependency and informational density are less contradictory in the

discourse of credible news. As credible news sources rely on real events, both those which have already occurred or are in progress, it is not surprising that these two dimensions would bear a weaker negative relationship compared to deceptive news.

Dimensions one and two also displayed a weak negative relationship with r=-0.23, indicating that informational density and the element of narrative are lightly disassociated in the discourse of credible news. The downhill relationship is also visible in the downward slope of the line and ellipse in the intersecting cells for dimension one and three of the matrix in Figure 4.2. This suggests that when information is conveyed in credible news, it tends to do so without narrative devices. This relationship is expected for credible news, as it received both a lower score on dimension one, indicating higher informational density, and dimension two, indicating less narration, compared to deceptive news.

Somewhat surprisingly, there is a very minimal uphill correlation between dimension one and dimension four in the credible news corpus. The correlation score, r=0.24, is slightly higher than the score for these dimensions seen in Figure 4.1 for deceptive news. The uphill relationship is also visible in the upward slope of the line and ellipse in the intersecting cells for dimension one and four of the matrix in Figure 4.2. There seems to be a weak link between persuasion and density of information in both the credible news corpus and deceptive news corpus. It appears that both deceptive and credible news have a slight tendency to express information through persuasive language. It is likely that this slight tone of assertiveness is a common trait of news discourse. As a journalist is confident in their research, or attempting to appear so, persuasive language is an expected result. Furthermore, news is often reported from an angle, usually motivated by external political and economical interests, so some element of opinionated language is not surprising (Fowler, 2013).

Altogether, correlations between dimension scores for both corpora were unremarkable. The majority of relationships were void of any strong relationship. The weak uphill correlation found in both corpora between dimensions one and four and the the downhill relationship found between dimensions one and three show that these relationships are associated with both deceptive and credible news. It is possible that the absence of situational dependence in informational reporting and a slight timbre of persuasiveness are traits which are shared between credible and deceptive news. If these propensities were initially commonplace in legitimate news discourse, then it is normal for deceptive news to include them when emulating credible news.

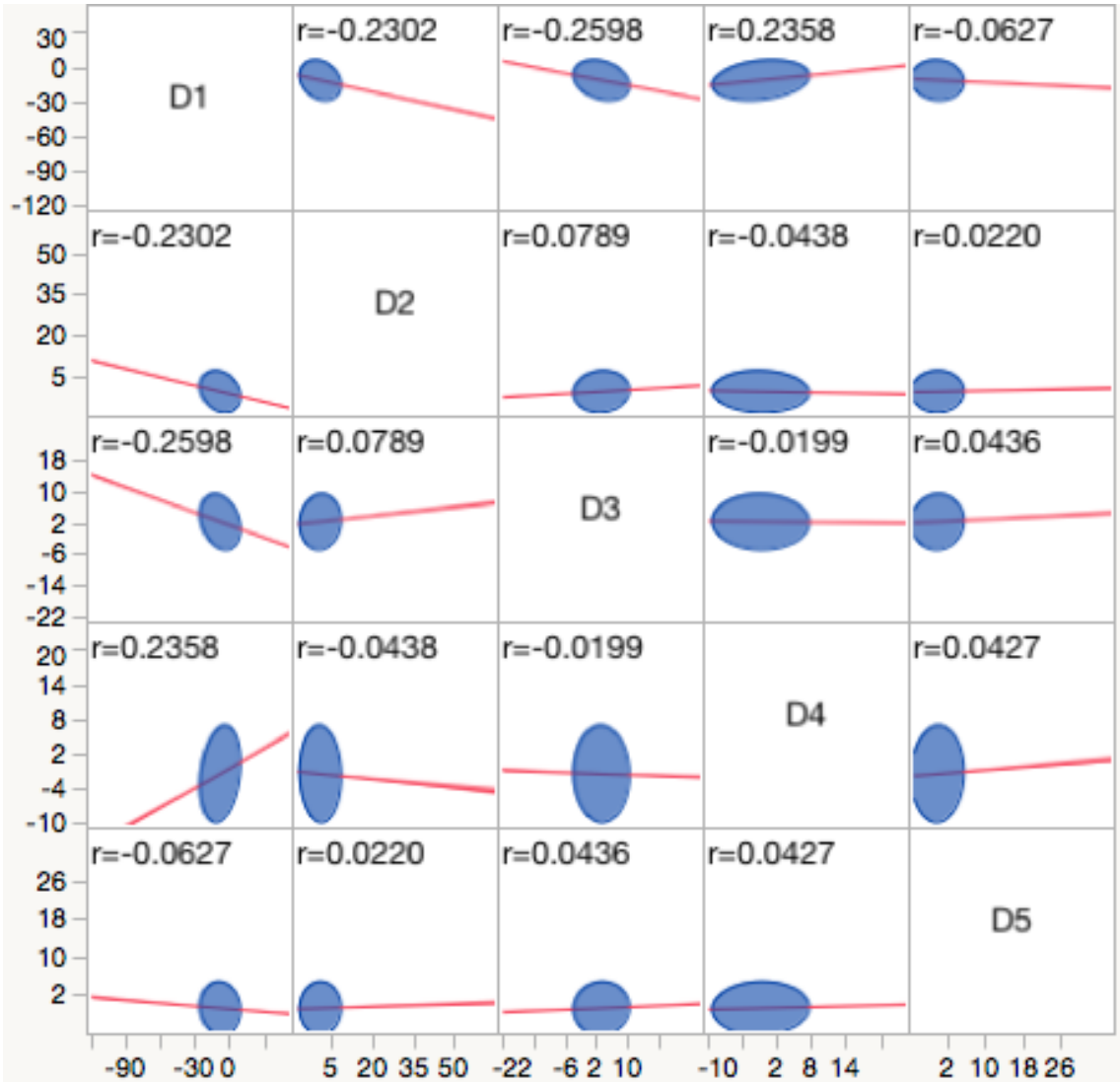

Figure 4.2: The results for the credible news corpus from the r correlation measure displayed by dimension.

4.5 Z-Scores Results

The following sections will present the z-score results for linguistic features. As seen in Sections 4.2.1 and 4.3, these features are used to determine the dimension score for a text. Here, they will be examined as they are. First, the discussion will focus on variables flagged as important within the corpus. Second, scores which have not been flagged as variables of interest, but show a considerable difference between the two corpora, will also be examined. Finally, the scores will be analyzed for any interesting correlations between the variables.

As the original MDA framework includes a total of 67 features, it is not reasonable to discuss all of these in detail within the limits of this thesis. For this purpose, a least significant difference (LSD) Student's t of 0.5 was selected for the minimum difference

between linguistic variables. All other linguistic variables received a score much smaller than 0.5, indicating that they are less distinct. The mean z-scores for all variables, including those not discussed in the following sections, are included in Appendix A.1.

4.5.1 Salient Linguistic Variables

The z-scores outputted for the corpus include a short list of linguistic variables expressed as either underused or overused. These variables received a score outside the range for -1.96 to 1.96, indicating observably high or low usage in the corpus. Both the corpora displayed similar overused variables, with general nouns and sentence relatives being most common. Unlike the deceptive news corpus, credible news did not report any significant underused variables. Table 4.7 displays the z-scores outputted for each of the linguistic variables identified as those of importance. These scores are within the range of -1.96 to 1.96 indicating statistical significance within the corpus.

Variable	Deceptive	Credible
General Nouns	3.9	3.1
Sentence Relatives	2.3	3.6
'That' relative clauses on subject	1.9	2.8
Adverbs	-2.2	-1.8

Table 4.7: The average scores per salient linguistic variable for each corpus.

The most prominent feature of both corpora was general nouns. Other than typical nouns, named entities, possessives, and plurals are also included in this label. General nouns were noted as being rather liberally employed in both datasets, a fact which is not surprising as the goal of news reportage is to report on entities and events. While both corpora reported high scores for this variable, the deceptive news corpus produced a higher overall z-score of 3.9. The credible news corpus produced a slightly lower average z-score of 3.1, signifying a more pronounced usage of nouns in the discourse of deceptive news. Once again, as mentioned in section 4.1.1, these results have all been tested for significance and returned a $p<0.0001$ allowing for the rejection of the null hypothesis. This result applies for all scores presented. Nouns are typically used to add a new relation or detail to the text, or to connect with other nouns through collocation or compounding. Liars may use more words and a include a higher frequency of questions in their communication (Hancock et al., 2007), possibly in an effort to persuade by fleshing out their fabrication. Nouns are also noted as a commonly distributed feature in academic prose (Biber, Conrad, & Cortes, 2004). It is likely that by including more nouns, thus more references, deceptive news is attempting to subtly persuade its readers into accepting the information it presents.

The second most utilized variable was the sentence relative, which included punctuation marks followed by the word 'which'. It was the credible news corpus which received the higher score for this feature, a z-score of 3.6. The deceptive news corpus, similarly embossed

with this feature but to less of an extreme degree, received a score of z-2.3. The difference between the two corpora is sizable, demonstrating that sentence relatives are more a defining characteristic of credible news than deceptive news. The primary usage of this variable was for the purpose of elaboration, as seen in the examples below. Example 12 from the credible news corpus is an excerpt from an article which received a z-score of 49.5 for sentential relatives.

(12) "*President Donald Trump's tax returns were perpetually sought and never fully published during the 2016 presidential campaign - and that hasn't changed. But on Tuesday night, MSNBC's Rachel Maddow shared Trump's 2005 1040 form,* ***which*** *was anonymously mailed to reporter David Cay Johnston. The return reveals that in 2005, Trump paid a cumulative $38 million in taxes on $150 million in income, an effective tax rate of about 25 percent. Most of that came from the alternative minimum tax,* ***which*** *applies to wealthy households that have taken a lot of tax deductions. The two-page document,* ***which*** *Maddow introduced with significant speculation about what the president's full tax returns might contain, doesn't answer all the questions Trump's more recent full returns might reveal, such as where the president's debts are owed or all the sources of his income. Nor does it include all of Trump's itemized deductions,* ***which*** *would include things like his charitable donations.*"–**Vox News (Nelson, 2017)**

The final overused variable was 'that' relative clauses on subject position. This variable marked occurrences of 'that' which were preceded by a noun and followed by a verb or an auxiliary verb. 'That' relative clauses on subject position were noted as a vital feature of the credible news corpus with a z-score of 2.8. The deceptive news corpus did not exhibit the same affinity for the structure, as it received an average score of 1.9. While 1.9 is not statistically significant, it is still a remarkable score. As the content of credible news displayed more outstanding usage of 'that' relatives on subject position, it can be concluded that it is a feature closely associated with the writing of credible news sources. Relative clauses on subject position are another tool, similar to sentence relatives, which function to provide elaboration and detail.

Previous research has indicated that dependent clauses are typical of written prose (Akinnaso, 1982), while Biber et al. (2004) has shown that they are also typical of classroom speech. As dependent clauses serve a referential function, lexical bundles including them should be more common in elaboration. Biber et al. (2004) also shows that, while such clauses are more common in classroom speech, they are also modestly distributed in academic prose and textbooks. As previously discussed, credible and deceptive news show some similarity with academic prose in their low scores on dimension one. The article presented below in Example 13 received a high score of 40.8 for this variable, illustrating the context of relatives on subject within press reportage.

(13) "*Arrowheads and stone ballista balls fired by catapults were uncovered on the main road* ***that*** *ascended from the city's gates and the Pool of Siloam to the Jewish Temple, the IAA said Thursday. The artifacts, excavated with the financial support of the City of David Society, tell the story of the final battle between the Roman army and Jewish rebels* ***that*** *ended with the destruction of the Second Temple and the rest of ancient Jerusalem, events* ***that*** *are famously described by historian Flavius Josephus.*"–**Breitbart, 2017**

The only variable recorded as underused was the adverb. Unlike the Stanford Tagger, which identifies unique labels for different forms, all adverbs are combined under one label in this system. Adverbs were only recognized as underused within the deceptive news corpus, which received a score of -2.2. Despite also receiving a solidly negative score of -1.8 for adverbial presence, it was not viewed as an underused variable within the content of credible articles. Considering the difference between the scores is only 0.4, and the significance of only one of the variables, it seems adverbials may be uncommon in press reportage. This assumption conforms to the findings of Biber (1988), which reported similarly low frequencies for adverbial usage in press reportage. However, given the disparity in scores, the general scarcity of adverbials appears to be more of a defining of deceptive news articles than of credible news.

As the under and overused variables identified by the tagger have shown, there is compelling evidence to assume variation in the discourse of deceptive news and credible news. While credible news expands upon its entities and assertions with relativization, deceptive news presents more referents with copious nouns. Despite a slight gap between the average score of the credible and deceptive corpus, the application of adverbs seems to be generally uncommon in both corpora. As low adverbial frequency is also seen in Biber's previous analyses, it appears that this is not a productive feature of news text (Biber, 1988, 1995).

4.5.2 Small Numbers Big Differences

The linguistic variables presented in this section were not flagged as significant by MAT, but are nonetheless worthy of mention due to a considerable disparity between the scores of each corpus. Table 4.8 shows the z-scores of variables showing observable inequality which had not been flagged as salient features by the system. As previously mentioned, to avoid a discussion of egregious length which would detract from the focus of distinguishing news types based on observable differences, only features which varied by an LSD of 0.5 or greater were included in this section. The motivation for this cutoff is partially attributed to Biber (1988), which only included features with a score of 0.4 or greater in the computation of factor scores. It was noted in both Biber (1995) and Biber (1988) that scores less than 0.4 were of little importance, regardless of significance. This attribution of importance refers back to the original mathematical description of factor analysis, which specifies a standard

for loadings being 0.7, where 0.7 corresponds to half of the variance (Gorsuch, 1983). This value is often lowered to 0.4 for most research, as the standard value of 0.7 is considered by many to be too high. Gorsuch (1983) mentions that the cutoff is mathematically arbitrary and is largely left to the discretion of the researcher, based on their unique theoretical stance. I have increased the cutoff to 0.5 in this analysis to optimize the discussion by eliminating variables of lesser importance and to accommodate for a difference in scale between feature scores in Biber and LSD. The z-scores for the corpora presented in this thesis represent the mean score for the entire corpus, either deceptive news or credible news. Scores specified for individual examples represent the z-score for that example alone.

Variable	Deceptive	Credible
Independent clause coordination	0.1	0.9
Emphatic	0.0	0.6
Attributive adjectives	-0.1	0.5
Predicative adjectives	0.3	0.9
Type-token ratio	-1.6	0.1
Public verbs	1.0	0.0
Past Participle Clauses	0.8	1.3

Table 4.8: Scores of variables with an LSD threshold of ≥ 0.5 not flagged by MAT.

The first variable of interest was independent clause coordination. Independent clause coordination is assigned when the word 'and' meets one of three conditions. The first is cases where it is preceded by a comma and followed by either an omitted subject, or 'it', 'so', 'then', 'you', or 'there' in conjunction with 'be' as a main verb, with a demonstrative, or with the nominal form of personal pronouns. The following sentence 14 provides an example of this condition.

(14) "*The New Yorker has been selected as a finalist for a 2011 National Magazine Award for General Excellence,* ***and*** *received eight additional nominations in seven categories.*"–**New Yorker, 2017**

The second condition is when it is preceded by punctuation, as seen in Example 15.

(15) "*I had no way of reaching him. It still breaks my heart.* ***And*** *it breaks it more every time he says something harsh against the gay community.*"–**Empire News, 2017**

The third condition is when it is followed by any wh-word or pronoun, an adverbial subordinator, a discourse particle, or a conjunct. Sentence 16 is an example of this condition with the use of a pronoun.

(16) "*We apologize to our readers,* ***and*** *we promise more stringent oversight in the future.*"–**Clickhole, 2014**

When comparing the z-scores of both corpora on the prevalence of independent clause coordination, it was found that this variable had a disparity of 0.8. The deceptive news corpus received a lower score of 0.1 on this variable, indicating that independent clause coordination is not a commonly used feature in deceptive news discourse. Contrastingly, the credible news corpus received a much higher score of 0.9. Unlike deceptive news, independent clause coordination appears to be a frequent device in the discourse of credible news.

The second variable which showed considerable difference was the emphatic. Words which are included in this category are 'just', 'really', 'most', 'real' and 'so' when combined with an adjective, or any form of 'do' followed by a verb. Sentence 17 is one example of the usage of the emphatic.

(17) "*Hopefully this is* ***just*** *some viral marketing for a horror movie that's coming out soon.*"–**Viralactions, 2015**

Usage of the emphatic proved to be more conspicuous in the credible news corpus, with a z-score of 0.6, compared to the deceptive news corpus. The emphatic was unmarked in deceptive news discourse, receiving a score of 0. This discovery is somewhat counter to expectations one may have of the discourse of credible news sources. However, it appears that the employment of emphatic devices has a deeper association with the discourse of credible sources compared to deceptive sources. The following passage in Example 18 belongs to an article from a credible source. It received a score of 9.8 for the use of the emphatic.

(18) "*If it is left up to the prosecutors alone, they might have a* ***more*** *jaundiced view of how a jury would hear a witness than does a commander-again, no longer the unit commander, and no longer alone. And part has to do with the changing culture of the military: McCaskill and others have fought hard to make commanders* ***more*** *responsible for addressing the crisis of sexual assault in their ranks, not less so. (McCaskill and Kelly Ayotte, a Republican, also co-sponsored another bill to strengthen the N.D.A.A., which moved forward easily on Thursday.) The idea here is that it has to be an element in what we consider a good officer, a part of promotions and evaluations. A commander shouldn't* ***just*** *shrug and guess that, if a young soldier* ***really*** *has a problem, he or she will go to someone else's office.*"–**New Yorker (Sorkin, 2017)**

Adjectives, both attributive and predicative, also displayed a rather sizable disparity in saliency between the two corpora. Attributive adjectives encompass all adjectives tagged

by the Stanford Tagger as adjective, superlative adjective, or comparative adjective. The credible news corpus presented more active employment of attributive adjectives, yielding a z-score of 0.5. In opposition, the deceptive news corpus reported attributive adjectives as a slightly underused variable with a score of -0.1. Referring back to the findings of Rashkin et al. (2017), this is somewhat contrastive. Rashkin et al. (2017) noted that deceptive news had a higher frequency of superlatives in comparison to credible news. This may still be the case, as superlatives are included along with other adjectives in this analysis and not measured separately. However, adjectives in general, as suggested by these results, are less frequent in the discourse of deceptive news. The presence of adjectives is linked with low scores in dimension one, as discussed in section 4.3.1.

Predicative adjectives are tagged as such in cases where any form of 'be' is followed by an adjective, a word that is not another adjective, an adverb, or a noun. Neither intervening adverbs nor negation disrupt the tagging, and predicative adjectives combined through the use of a phrasal coordination are included within this category as well. Once again, the credible news corpus received the higher score on this variable, a score of 0.9. While predicative adjectives were not considered underused in the deceptive news corpus, it received a much lower score of 0.3. From these results, it is evident that the discourse of credible sources is more richly embossed with modifiers and descriptives than that of deceptive news. Passage 19 contains both predicative and attributive adjectives, taken from an article which received a score of 0.6 for predicative adjectives and 13.2 for attributive.

(19) "*While speaking at Paris Fashion Week on Monday, fashion designer Stella McCartney introduced her* ***new*** *collection with a* ***heaping*** *side of gender norms. She explained backstage that her* ***new*** *collection welcomed the "fragility" of women and displayed their "softness." "Strength on its own in a woman is quite* ***aggressive*** *and not terribly* ***attractive*** *all the time," McCartney said. "This collection is really celebrating the* ***gentle*** *side." It's* ***possible*** *that McCartney meant that the clothing in her collection is supposed to help women feel* ***softer*** *and appreciate their* ***gentle*** *sides. That, in and of itself, is* ***fine****. But implying* ***strong*** *women aren't "terribly* ***attractive*** *all the time" is more* ***problematic****. McCartney's comments are a reminder that the way a woman is supposed to look is dictated by society. Be* ***thin****. Be* ***fragile****. Be* ***soft****, social constructions say, but don't be* ***strong*** *because being* ***strong*** *is* ***dominating*** *and "****aggressive****," and women are supposed to be* ***passive****."*–**Vox News (McKinney, 2014)**

Example 20 is a passage from a deceptive news article contains both predicative and attributive adjectives. The article received a score of 1.0 for predicative adjectives and 2.7 for attributive.

(20) "*Although the recently revealed allegations against Donald Trump are* ***unverified****, they've already had a* ***major*** *effect on the country's internet search history: "****Golden*** *showers" has skyrocketed from being the most* ***Googled*** *term in the United States to being far and away the most Googled term.* ***Erotic*** *urination has long been* ***front*** *and* ***center*** *in the* ***national*** *zeitgeist, but the unsubstantiated report about Trump paying prostitutes to pee on each other has brought way more attention to the* ***wet sexual*** *fetish Americans were already obsessed with.*"–**Clickhole, 2017**

Public verbs, as defined by the Oxford Dictionary of English grammar (Chalker, 2003), are verbs which denote spoken discourse and are often used in combination with 'that'-clauses to express veritable propositions. Examples of public verbs are 'acknowledge', 'confirm', and 'deny'. For this variable, it was the deceptive news corpus which displayed the most frequent usage with a score of 1.0. Contrarily, the credible news corpus reported public verbs as unmarked with a score of 0. The copious usage of public verbs suggests that deceptive news makes more appeals to authority or witness statements, likely in an effort to reinforce an impression of factuality. Example 21 is a passage from an article retrieved from a deceptive source received a score of 12.1 for the frequency of public verbs. The public verbs used in this passage clearly show its use as a tool to provide testimony from official figures and organizations for both parties.

(21) "*"The launch, from the eastern area of Wonsan, may have failed" Kyodo* ***said****,* ***adding*** *that the type of missiles involved was unknown. Chief Cabinet Secretary Yoshihide Suga* ***said*** *that the government hasn't* ***confirmed*** *North Korea launched missiles towards Japan and that it was always gathering and analyzing information on the country. South Korea's defense ministry* ***said*** *it was trying to* ***confirm*** *the report of the missile launches.*"–**Breitbart, 2017**

The second to last variable to be discussed presented the largest variance between the two corpora. Type-token ratio, referred to as TTR in Biber (1988), Biber and Egbert (2016), Nini (2015), measures the number of different types of tokens found within the first x number of words of the text. It is a means of representing the lexical diversity of a text. Texts longer than the number of tokens set by the ratio are only considered up to that maximum number of tokens, the remaining tokens are not considered in the analysis. It is an appraisal of the linguistic diversity within discourse, where x is set by the analyst to accommodate smaller or larger texts by equalizing the lengths. As a convention, applied by both Biber and Nini, the number of tokens used to calculate this score is the first 400 of the text. It is assumed that frequent words within a text will reappear more often after these initial 400 tokens, thus will not contribute further to the lexical diversity.

The disparity between the two corpora was extensive, with a difference of 1.7. The credible news corpus received a positive score of 0.1 for this variable, indicating a higher degree

of linguistic diversity. The score received by the deceptive news corpus, a remarkably low score of -1.6, has revealed the opposite about the discourse of fictitious news. In comparison to credible news discourse, the linguistic diversity of deceptive news is much lower. This may be a general trait of deceptive news, as suggested by the higher degree of lexical redundancy and lower readability grade noted by Horne and Adali (2017). However, it may also be the case that deceptive articles repeat more information within the first 400 tokens of an article, leading to a lower diversity score.

The final variable in this discussion is past participle clauses. Past participial clauses include all cases in which a punctuation mark is followed by the past participle form of a verb, which is then followed by a preposition or an adverb. The following sentence is an example from a credible news article.

(22) "*The survey of 1,000 people,* ***taken*** *by landline and cell phone Wednesday through Monday, has a margin of error of plus or minus 3 percentage points*"–**USA Today**

The difference between corpora for this variable was 0.5. The credible news corpus displayed more frequent usage with a score of 1.3. In contrast, the deceptive news corpus reported past participial clauses as only minimally frequent with a score of 0.8. Past participial clauses have been found in higher frequency in written language (Akinnaso, 1982), and may serve as a "guidepost" for flow of information (Chafe, 1984). As credible news is less abstract than deceptive news, as shown in section 4.3.5, it may be using past participial clauses more frequently to convey information more explicitly in order to emphasize information which must be clearly understood.

While these features do not fall outside the range of -1.96 to 1.96 indicating statistical significance, the additional features discussed in this section have assisted in providing more detailed insight into the characteristics of the discourse found within credible and deceptive news. The discourse of deceptive news exhibits a higher frequency of assertions and propositions introduced through public verbs, but it is decidedly lacking in linguistic variety when considering other variables. As evidenced by the profoundly low score reported for type-token ratio, deceptive news does not present a great variety of unique linguistic features withing the first 400 tokens. Dependable news, through increased usage of independent clause coordination, emphatic, and adjectives, has revealed a more vibrant linguistic tapestry in its expression.

4.6 A Note on Correlation

The strength of relationships between certain linguistic variables within a text can provide some insight on the specific stylistic tendencies of different corpora, motivating the discernment of each corpus as a distinct register. While the majority of correlations found within

both the deceptive news and credible news corpus were negligible, a few stood out as having a weak to moderate relationship. When interpreting a correlation coefficient, R, only values of $\geq |0.4|$ are considered to have any important relationship, therefore only correlations between variables that meet this standard will be discussed in this section. Results are presented in a scatterplot matrix, where the r score and width of the ellipsis indicate the strength of the correlation and the line of fit signifies the directionality.

4.6.1 Correlations Attributed to Grammatical Patterns

Most of the relationships bearing moderate or strong correlation were easily attributed to grammatical conventions. The correlation between predicate nouns and 'be' as a main verb was found to be moderately uphill in both corpora, with an r score of 0.6 in both the deceptive news and credible news corpus. This is unsurprising given the prevalence of the structure 'x is y' in common discourse. If a 'be' verb is present within a text, it is also very likely that a predicate noun will be too.

Another correlation that was mildly positive in both corpora was the use of public verbs with subordinator that deletion, with both the deceptive news and credible news corpus finding an r score of 0.4. As subordinator that deletion is identified by the tagger in numerous cases where a public verb is followed by another variable, such as a demonstrative pronoun or a noun, it is understandable that these two features would often appear within the same text. Grammatical structures similar to 'x confirms y' are likely to be commonplace in news discourse, regardless of veracity.

A final obvious correlation was the weak downhill relationship discovered between adverbs and nouns, with the deceptive news corpus showing an r score of -0.4 and the credible corpus a slightly lower score of -0.5. It is unlikely that nouns and adverbs will be strongly linked as, with the exception of special cases such as gerunds or 'very', adverbs ordinarily do not co-occur with nouns. In articles featuring a high frequency of nouns, adverbs may therefore become less frequent. As seen in section 4.5.1, adverbs are reportedly underused in the deceptive news corpus, and very infrequent in the credible news corpus, while nouns were reported as overused for both corpora.

4.6.2 Correlations within the Deceptive News Corpus

The variable which interacted the most actively with other variables was the average word length. Horne and Adali (2017) observed in their research that deceptive news tends to have a shorter average word length. This is also noticed in this analysis with a z-score of 0.6 for the deceptive news corpus and 0.8 for the credible news corpus, included in Appendix A.1. A minimally positive correlation was observed for the relationships with attributive adjectives and prepositional phrases, indicating that these features are a modest factor in increased average word length. As seen in Figure 4.3, the intersection of average word length and attributive adjectives shows a positive correlation score of r=0.4, which signifies

a moderately uphill correlation between the two variables. The uphill relationship is further marked by the upward slope of the line and density of the ellipse. This uphill correlation is also seen in the intersection of average word length and prepositional phrases, with the same positive score of r=0.4. These relationships are not seen in the credible news corpus. In contrast to the deceptive news corpus, the intersection between average word length and attributive adjectives and average word length and prepositional phrases seen in Figure 4.4 shows that these relations are unmarked in credible news.

The relationship between average word length and nominalization was more profound in the deceptive news corpus. As seen in Figure 4.3, the intersection of average word length and nominalization shows a positive correlation of r=0.6, which signifies a moderately uphill correlation between the two variables. The uphill relationship is further marked by the upward slope of the line and density of the ellipse. The close association between average word length and nominalization suggests that nominalization is the device most responsible for increasing the average word length in deceptive articles. As the frequency of nominalization in a deceptive article increases, so does the average word length. Despite this correlation, deceptive news has a shorter average word length compared to credible news, where deceptive news received a score of 0.6 and credible news received a score of 0.8. These scores are presented in Appendix A.1. Figure 4.4 shows the relationship between these same variables in the credible news corpus. The intersection between average word length and nominalization shows that this relationship is relatively unmarked. Since there was no notable correlation present within the credible news corpus, this argues that this relationship is unique to the discourse of deceptive news. As credible news articles exhibit a greater average word length, it can be deduced that the average word length is increased by all linguistic variables relatively equally in credible articles, while it is mainly nominalization which increases the average word length in deceptive articles.

In contrast, present tense verbs and average word length displayed a mildly negative correlation, indicating that verbs in the present tense do not contribute to an increase in average word length. As seen in Figure 4.3, the intersection of average word length and present tense verbs shows a negative correlation score of r=-0.4, which signifies a moderately downhill correlation between the two variables. The downhill relationship is further marked by the downward slope of the line and density of the ellipse. This relationship indicates that present tense verbs in the discourse of deceptive news reduce average word length, as evidenced by the moderately negative correlation. No meaningful correlation was found within the credible news corpus, as shown in the intersection between present tense verbs and average word length seen in Figure 4.4.

A final notable relationship found within the deceptive news corpus was the correlation between prepositional phrases and present tense verbs. A weak downhill correlation was found between these two variables, signaling that they are not commonly used together in deceptive news articles. As seen in Figure 4.3, the intersection of prepositional phrases and

present tense verbs shows a negative correlation score of r=-0.4, which signifies a moderately downhill correlation between the two variables. The downhill relationship is further marked by the downward slope of the line and density of the ellipse. Although near negligible and considerably less than the deceptive news corpus, the credible news corpus did also display a weak downhill correlation between the two variables. The intersection of prepositional phrases and present tense verbs in Figure 4.4 shows a negative correlation score of r=-0.3, which signifies a mildly downhill correlation between the two variables in the credible news corpus. The link between prepositional phrases and present tense verbs is substantially negative in both registers, which suggests that it is naturally an uncommon pattern of co-occurrence in news discourse. Previous research on written discourse in comparison to oral discourse has noted that prepositional phrases are more frequent in formal written language (Chafe, 1982). In an analysis of lexical bundles, Biber et al. (2004) remarked that time, place, and text-deixis bundles, for example "as shown in x" are more common in written discourse. Prepositions of this sort may co-occur more often with the past tense or other verb types over the present tense, resulting in a mild downhill relationship between the variables.

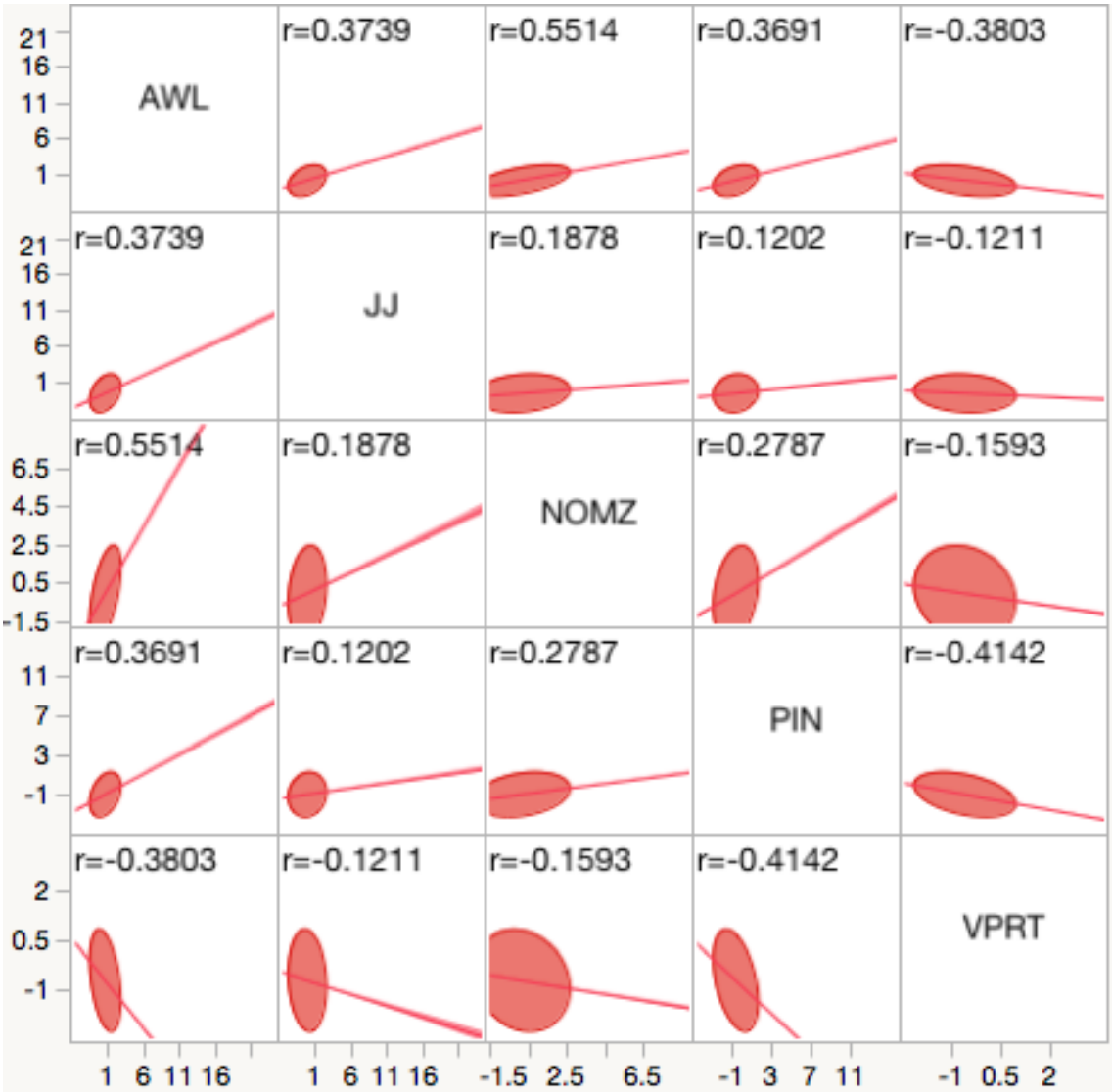

Figure 4.3: Scatterplot matrix showing correlations within the deceptive news corpus. The legend for the variables: AWL - average word length, JJ - attributive adjectives, NOMZ - nominalization, PIN - prepositional phrases, VPRT - present tense verbs.

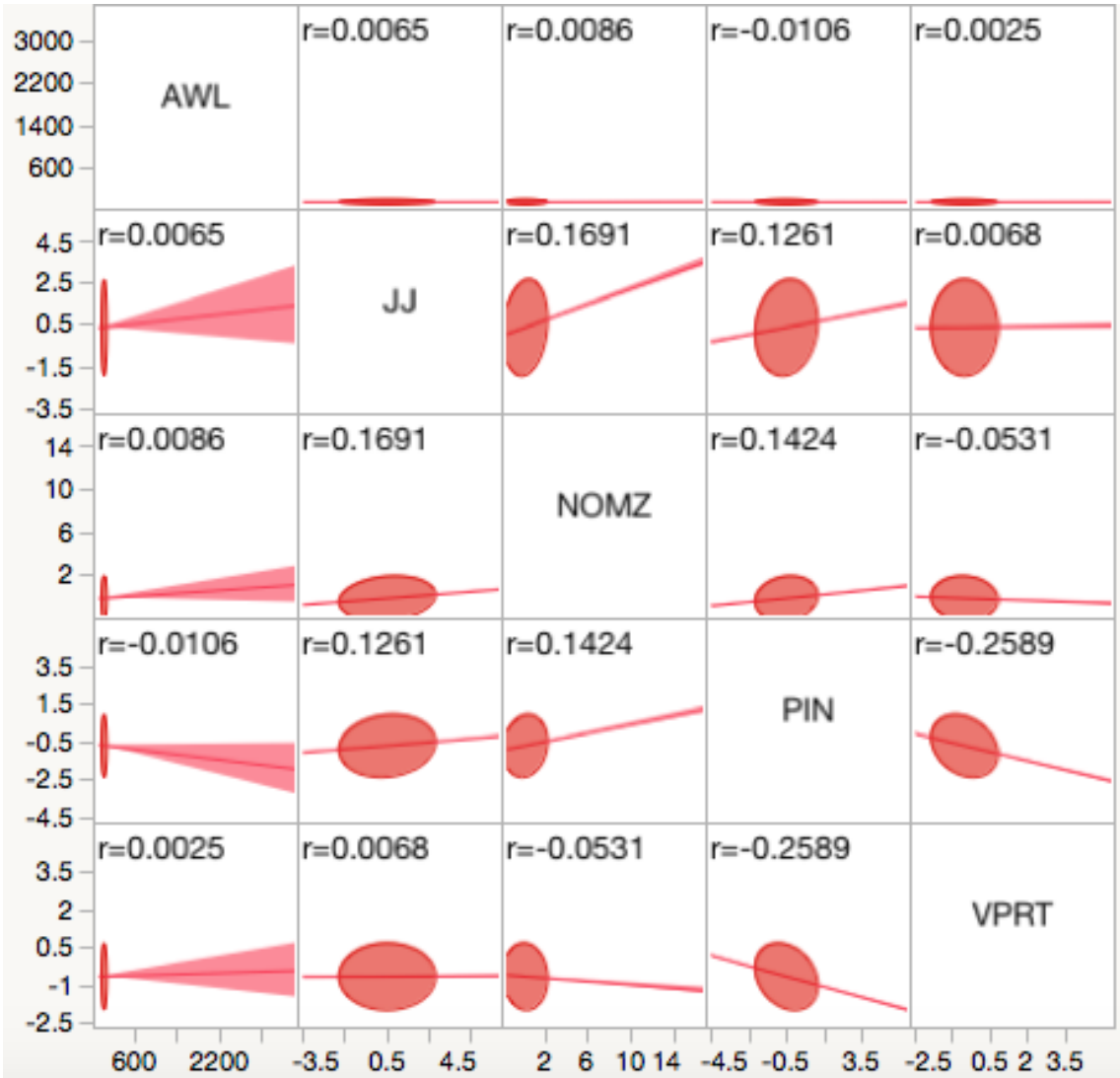

Figure 4.4: Scatterplot matrix showing the same variables in Figure 4.3 for the credible news corpus in contrast.

4.6.3 Correlations within the Credible News Corpus

While robust correlations found within the deceptive news corpus largely focused on the variable for average word length, the credible news corpus presented more variation among correlations. Past tense verbs and third person pronouns were found to be positively linked. The intersection of past tense verbs and third person pronouns shows a positive correlation score of r=0.4, which signifies a moderate uphill correlation between the two variables. This is seen in Figure 4.5, where the upward slope of the line and angle of the density ellipse also indicate this relationship. The deceptive news corpus also showed a slight positive correlation between these variables, with a score of r=0.3. Figure 4.6 displays how the correlations found noteworthy within the credible news corpus interact within the deceptive corpus. While the

usage of third person pronouns with the past tense is doubtlessly a universal structure in news discourse, as reference to past events and actors is an essential element of reporting, it appears slightly more prominent in the discourse of credible news. This is possibly due to credible articles referring to actual past events with witnesses and expert opinion.

The strongest correlation within the credible news corpus was discovered between analytic negation and contractions. Negation includes both the expanded form, 'not', and its contraction, 'n't'. The contracted form is also included in the category of contractions, leading to an overlap in contraction and negation where 'n't' receives both labels. The intersection of contraction and negation shows a positive correlation score of r=0.6, which indicates a moderately uphill correlation between the two variables. This relationship is shown in Figure 4.5, where the upward slope of the line and angle of the density ellipse further illustrate this relationship. Contrary to expectation, the deceptive news corpus did not present any correlation between these two features. Figure 4.6 shows the interaction between negation and contraction in the deceptive news corpus. The intersection between these two variables in the deceptive corpus shows a negligible correlation of r=0.1. From these observations, it appears that contraction in conjunction with negation is more common in credible discourse. Appendix A.1 shows the z-scores for both negation and contraction. Credible news discourse has a higher frequency of contraction compared to deceptive news discourse, -0.3 in the credible corpus opposed to -0.7 in the deceptive. Credible news discourse also displays a higher frequency of negation compared to deceptive news discourse, once again -0.3 opposed to -0.7. Although these variables were not flagged, this difference is interesting. This shows that both contraction and negation are more common in credible news, thus it is more likely they will show correlation. It appears that negation is likely to surface through contraction in the discourse of credible news based on the correlation findings, while deceptive news is more likely to express negation through the use of the expanded form. This observation seems counter to intuitions about deceptive news, where one might expect more informal discourse with higher a frequency of contraction. It is possible that the lower frequency of contraction is due to hyper-awareness of this variable as a feature of informal discourse. As deceptive news typically emulates the discourse of credible sources, use of contraction may be avoided purposefully to establish a sense of formality and sophistication.

Analytic negation was also found to have a minimal negative association with nouns. The intersection between nouns and negation shows a negative correlation score of r=-0.4, which indicates a moderate downhill relationship between the two variables. This relationship is shown in Figure 4.5, where the line's downward slope and density ellipse's angle also depict the dissociation between the variables. The link between these two variables was reported as trivial within the deceptive news corpus. Figure 4.6 shows the interrelationship between nouns and negation in the deceptive corpus, where a clearly insignificant r=0.1 is presented. Despite a higher frequency of negation and lower frequency of nouns in the credible news

corpus, the negative correlation between these two variables indicates that nouns do not commonly interact with negation in the discourse of credible articles.

Contractions, in addition to previously discussed conditions, are also tagged for any instance of an apostrophe followed by a tagged word. Contractions and nouns were also found to be dissociated within the credible news corpus. The intersection between nouns and contractions in Figure 4.5 shows a negative correlation score of r=-0.4, which indicates a moderate downhill relationship between the two variables. The downward slope of the line in the matrix also depicts the complementary nature of the variables. In contrast, the deceptive news corpus reported these variables as unmarked for correlation. Figure 4.6 shows the intersection between nouns and contractions. The correlation score, lack of slope in the line, and flatness of the ellipse, indicate that there is no obvious correlation between these variables in the deceptive news corpus. Previously, it was observed that negation and contraction are correlated within credible news discourse. Despite the positive correlation found previously between contractions, the negative correlation between nouns and contractions in the discourse of credible articles is not surprising. As contractions of variations of 'be', 'have', or the modal verbs 'will' and 'would' with nouns is a feature of informal or spoken discourse, it is not surprising that highbrow news sources would avoid this form in their writing. While these two variables are not positively correlated in the discourse of deceptive news either, contractions with nouns are not explicitly avoided.

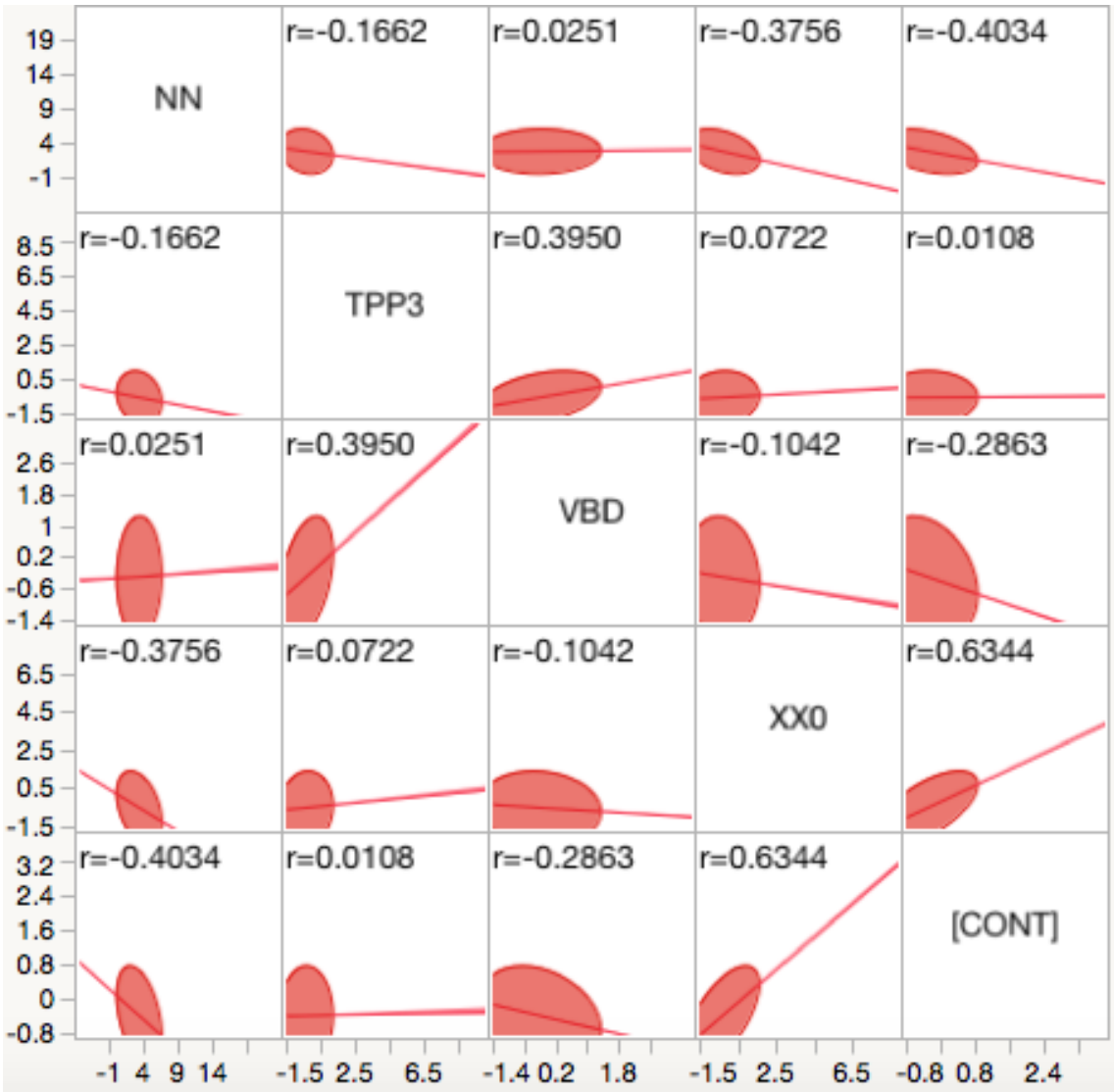

Figure 4.5: Scatterplot matrix showing correlations within the credible news corpus. The legend for the variables: NN - total nouns, TPP3 - third person pronouns, VBD - past tense verbs, XX0 - analytic negation, [CONT] - contractions.

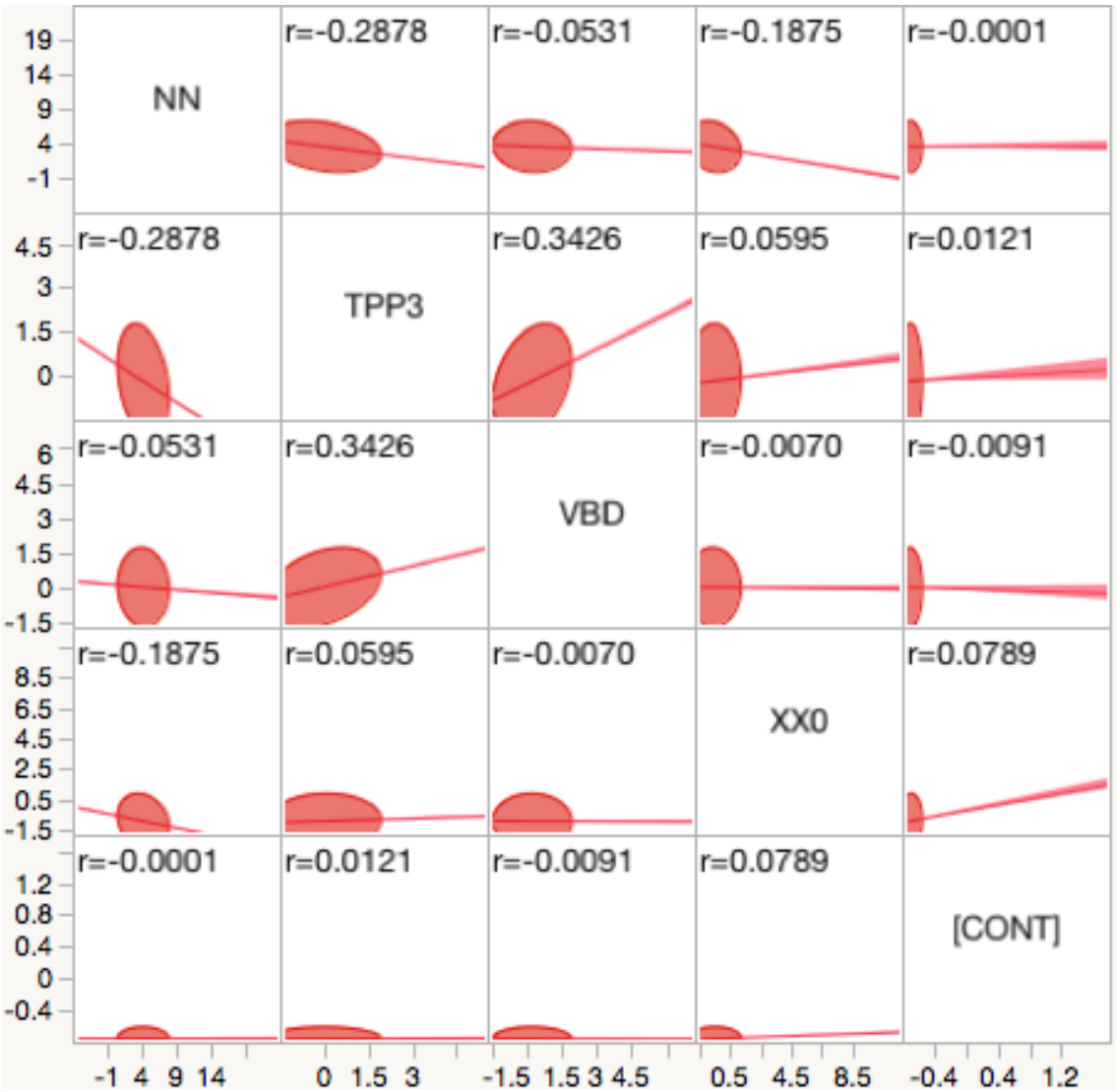

Figure 4.6: Scatterplot matrix showing the same variables as in 4.5 within the deceptive news corpus as contrast.

4.6.4 Summary of Correlations

In this section, it has been revealed that the variable which interacts the most with other variables in the deceptive news corpus is average word length. Several variables have some extent of influence on the average length of words found within the articles of deceptive news, an influence which was not perceived within the credible news corpus. Average word length surged with the usage of nominalization and adjectives, but declined with present tense verbs.

Within the credible news corpus, nouns, negation, and contraction were discovered to interact with each other in varied manners. Contractions interacted with negation progressively, but had an adverse relationship with nouns. The differences between the credible news

and deceptive news corpus presented here provide further support for the consideration of dichotomizing such news types based on dissimilarities in discourse.

4.7 Comparing Corpora with Established Varieties of Text

In addition to examining the variation between the discourse of deceptive news and credible news, the corpora have also been compared to eight text types which had previously been established in Biber (1988). These text types were compared based on the dimension scores from dimension one to five, with one text type from amongst the eight noted as the type most closely related to the corpora under comparison. The results of this comparison are shown in appendices B.1 and B.2. Additionally, the dimension scores of the credible news corpus and deceptive news corpus have been compared to those of subtypes of news discourse, such as 'political press' and 'press review'. The results of these comparisons have been presented in Figures 4.7 through 4.11.

4.7.1 Deceptive News, Credible News, and Common Text Types

As indicated in both Appendix B.2 and Appendix B.1, the text type 'general narrative exposition' was determined to be the most similar type for both corpora. Text types, as described by Biber, refer to groupings of texts which display similarity based on register, genre, and style. 'General narrative exposition' is inclusive of text varieties which rely on narration to convey information, and is composed of registers such as press reportage, editorial, biography, non-sports broadcast, and science fiction. Registers in this text variety have low scores on dimension one and high scores on dimension two.

Appendix B.1 shows the comparison per dimension between the scores of the deceptive news corpus and those of the eight text types, while Appendix B.2 shows the same comparisons between the credible news corpus and the text types. No specific scores were reported by MAT for the text types, therefore the scores reported in this section are approximate. The scores for dimension one were nearly identical for both the deceptive news corpus (-9.7) and the credible news corpus (-10.4), almost completely overlapping the point for general narrative exposition. This suggests that general narrative exposition, deceptive news, and credible news are similarly informational and lower in expression of persuasion compared to other text types. As press reportage is a register within general narrative exposition, it is not surprising that the scores for this dimension have aligned.

Dimension two shows some divergence between the credible news corpus (0.3), the deceptive news corpus (1.3), and general narrative exposition. As previously shown in Section 4.3, the deceptive news corpus received a higher dimension score than credible news, indicating that deceptive news is more narrative than credible content. An example of this was provided in passage 6. Compared to general narrative exposition, both corpora were scored as less narrative. Credible news presented more similarity with informational interaction,

the registers of which tend to display less narration as they are based on communicating information. Considering general narrative exposition also includes registers such as science-fiction and biographies, which tend to be more narrative, this is not too surprising.

Appendices B.1 and B.2 show that there is divergence between general narrative exposition, deceptive news (2.6), and credible news (2.8) in dimensions three. General narrative exposition is characteristically unmarked for dimension three, while both the deceptive news and credible news corpus displays a notably positive score. For this dimension, credible news appears to be midway between involved persuasion and scientific exposition. Registers in scientific exposition, such as academic prose and official documents, are all highly independent from context. In Appendix B.1, deceptive news is closer in space to the involved persuasion text type. Registers in the involved persuasion text type, such as spontaneous speeches or professional letters, are more argumentative and slightly more dependent on context compared to scientific exposition. Compared to general narrative exposition, deceptive and credible news are much more less context dependent, given their highly positive scores.

For dimension four, both the deceptive (-0.8) and credible news corpus (-1.1) align almost exactly with that of general narrative exposition. This approximation clearly demonstrates how general narrative exposition was selected as the closest text variety. As dimension four measures expression of persuasion, it can be concluded that argumentation and persuasiveness is not a strong trait of registers within general narrative exposition.

The final dimension, dimension five, exhibits a slight similarity between general narrative exposition, deceptive news (0.3), and credible news (-0.1). While the scores for this dimension are not wildly similar, they are close enough to warrant the selection of general narrative exposition as the closest text type. The credible news corpus resulted in a relatively unmarked score for this dimension, indicating that it is mildly more abstract than general narrative exposition. The deceptive news is also more abstract than general narrative exposition, with a score dipping into the positive range. Despite the higher scores on this dimension, the closest test type remains as general narrative exposition.

Although general narrative exposition was determined to be the closest text variety for both corpora, the disparity in dimensions two, three, and five demonstrate that they are not identical. However, the closest text type, with the exception of dimension three for both corpora, remained general narrative exposition in spite of these discrepancies.

4.7.2 Comparing News Subtypes

Ten news and editorial subtypes, previously analyzed with MDA in Biber (1988), have been included for comparison with the registers of credible and deceptive news. The comparisons for each dimension have been represented through an error bar plot showing the mean scores for each subtype.

Unfortunately, since the individual scores for texts in the corpus of each subtype introduced in Biber were not able to be acquired, the mean scores presented in the following plots are not completely accurate and are only for the purpose of making approximate comparisons.

Furthermore, as the MDA analysis performed in Biber (1988) utilized factor scores instead of z-scores, the scores of the credible and deceptive corpora and the corpora presented in Biber are not measured or scaled similarly. The use of z-scores in this analysis, while serving the same purpose of factor scores, has resulted in scores with a greater magnitude for shorter texts. These more intense scores appear as a much wider range in comparison to the other news subtypes. As a visual aid, the scores from the deceptive news corpus and credible news corpus have been scaled in half in graphs 4.7 through 4.11 to more clearly show the relationships between news subtypes. With these issues in mind, it is crucial to assert that the plots presented in this section serve only as visual aids for the purpose of comparing subtypes based on proximity.

In Figure 4.7, credible news and deceptive news are very close. In Section 4.3, the analysis showed that credible and deceptive news differed by only 0.7 points. While statistical testing determined that this was significant, the scores remain close in space. When compared to other registers of news, the closest type was 'institutional editorial', followed by 'press editorial'. Although all of the news and editorial subtypes from Biber received negative scores for dimension one, these two registers showed the highest degree of interaction and affect. As editorials are opinionated pieces, this is not surprising. As mentioned previously, while a positive score on dimension one was unanticipated for credible news, both the credible news corpus and deceptive news corpus showed a component of intimacy.

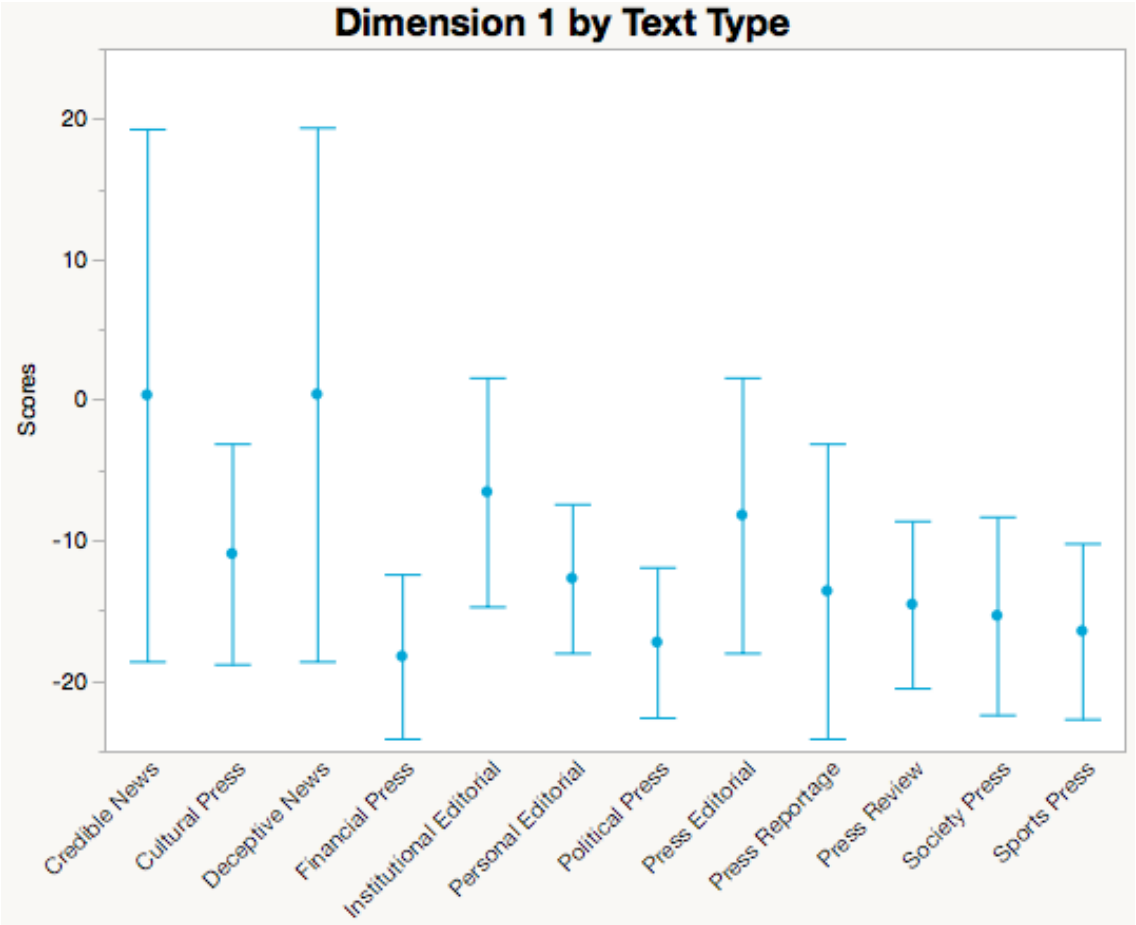

Figure 4.7: Error bar plot showing the maximum and minimum scores for dimension one for subtypes of news and editorial.

Figure 4.8 shows that there is not a lot of similarity between other types of news, the credible news corpus, and the deceptive news corpus. In terms of proximity, the closest type of text is 'press reportage'. As dimension two measures density of information, it is not surprising that types such as 'cultural press' and 'personal editorial' displayed little resemblance to hard news such as 'press reportage' or credible news.

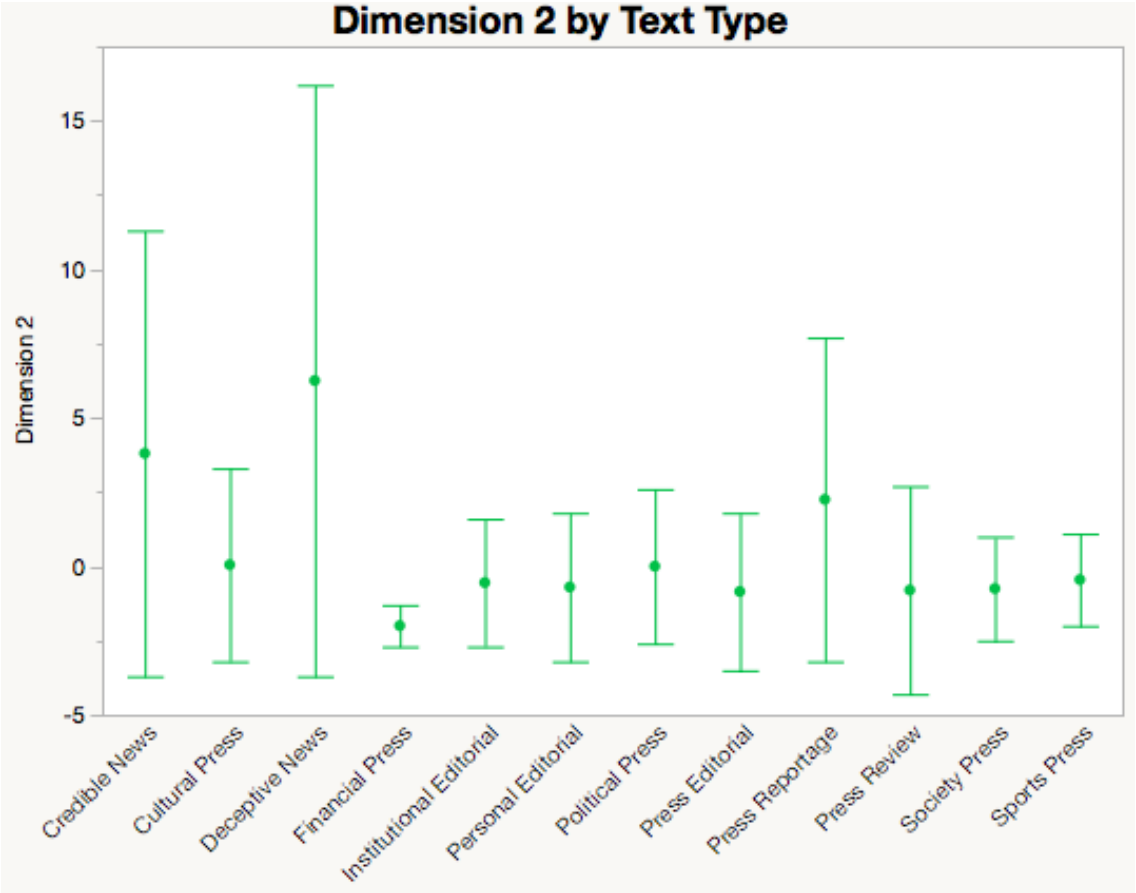

Figure 4.8: Error bar plot showing the maximum and minimum scores for dimension two for subtypes of news and editorial.

Dimension three, presented in Figure 4.9, shows that credible news, deceptive news, and the other ten subtypes are similarly independent of context. 'Personal editorial' and 'press review' presented the most deviant scores, resting slightly more above 0 in comparison to the other news types. As seen in Section 4.3, the mean scores for both credible news and deceptive news were nearly identical. As the other subtypes presented in Figure 4.9 seem to confirm, dependence on context is not a distinct characteristic of news discourse. In this respect, the credible news corpus and deceptive news corpus present results which align with other varieties of news.

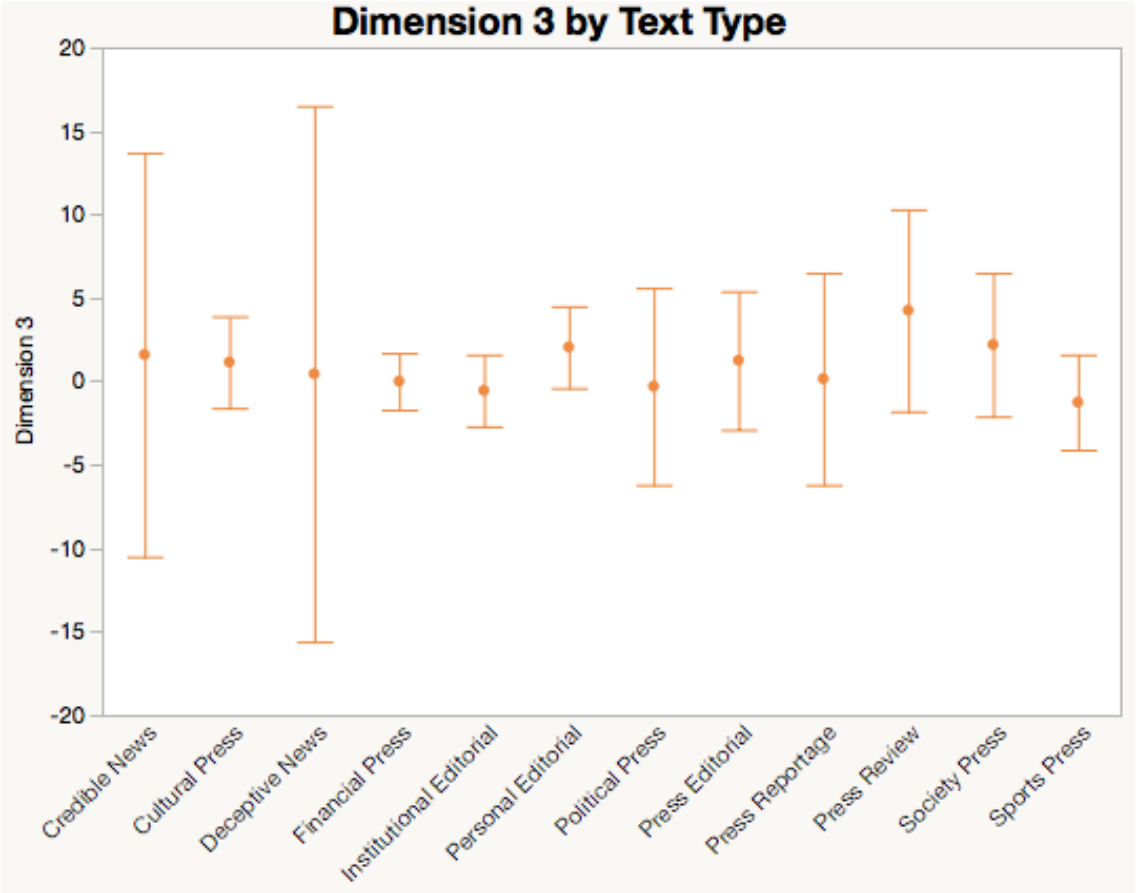

Figure 4.9: Error bar plot showing the maximum and minimum scores for dimension three.

For dimension four, the news subtype which showed the most similarity with the credible and deceptive news corpora was 'personal editorial'. Amongst the news subtypes, personal editorial showed the most substantial degree of personal expression and persuasion. For the majority of new varieties, expression of persuasion is not strong. With the exception of 'personal editorial', 'press editorial', and 'institutional editorial', the text varieties showed negative scores for argumentativeness and expression of opinion. In Section 4.3.4, it was shown that this was also the case for credible and deceptive news discourse, although credible news discourse demonstrated less of this trait compared to deceptive news.

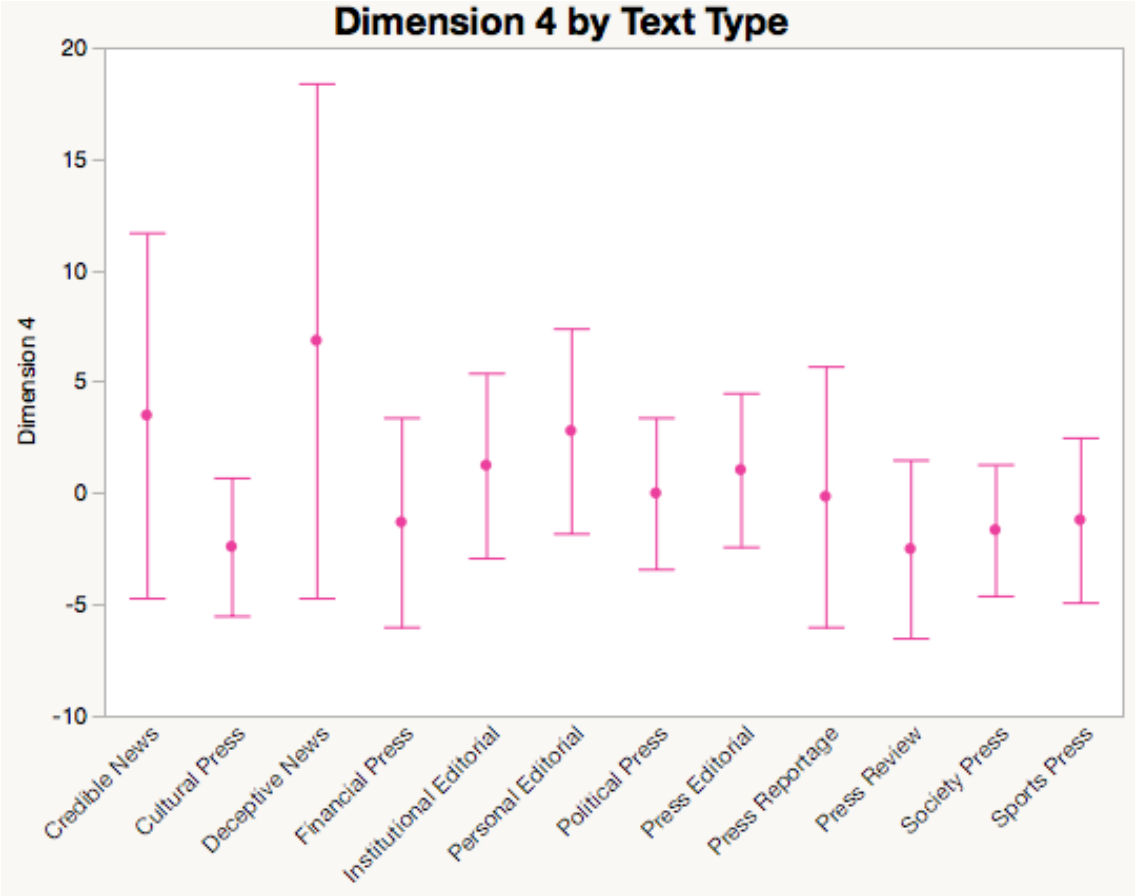

Figure 4.10: Error bar plot showing the maximum, minimum, and mean scores for dimension four.

As discussed in Section 4.3.5, the deceptive news corpus showed a higher degree of formality in its writing compared to the credible news corpus. This was surprising, as one may have expected reputable sources of news to be the more technical of the two. In comparison to other types of news, 'financial press' and 'press editorial' were the closest in space. From Figure 4.11, it can be observed that abstractness is a visible quality of news discourse. The majority of news varieties presented a positive scores for this dimension, with only two being unmarked.

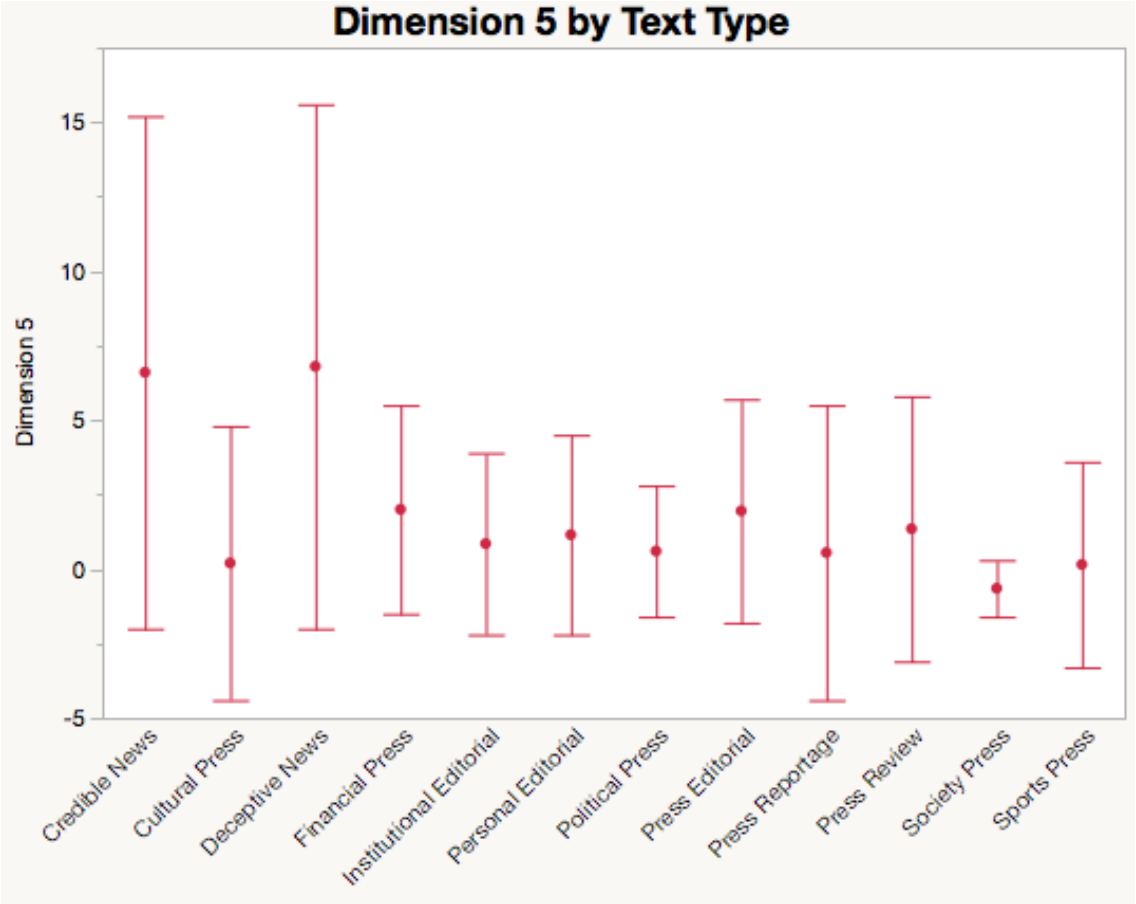

Figure 4.11: An error bar plot showing the maximum, minimum, and mean scores for dimension five for ten subtypes of news and editorial.

Based on dimensions one, four, and five, the deceptive news corpus and credible news corpus exhibited the most closeness with editorial registers. While no opinion articles were included in either of the corpora, the discourse of these registers suggest that they possess some element of folksiness. However, as seen in Figures 4.7 through 4.11, the credible news corpus and deceptive news corpus are obviously distinct from other types of news. Unfortunately, this analysis was limited due to insufficient access to existing data. Once again, it is important to remember that the mean scores in each graph are not representative of the actual mean. In the future, it will be beneficial to recreate this study with more data and the inclusion of subtypes within deceptive news.

Chapter 5

General Discussion and Conclusion

The purpose of this thesis was to provide evidence in support of the claim that deceptive news and credible news are discrete registers. Using multidimensional analysis (MDA) in conjunction with statistical examination, I revealed that there exists significant linguistic variation between the articles of deceptive news and credible news. This final chapter provides a summary of the most striking discoveries from the results of MDA and statistical analysis. A discussion on the contributions these findings will grant to future research will be given, followed by concluding remarks.

5.1 Summary of Results

The use of MDA to analyze texts for linguistic variation, through measuring five dimensions representing different aspect of discourse, has allowed observation of differences in the writing of deceptive and credible articles.

The density of information in the content of credible news was found to be more robust in comparison to deceptive news. It was also demonstrated that deceptive sources exhibit more characteristics of narrative discourse compared to credible sources. Deceptive news was shown to contain more language of persuasion and opinion in comparison to credible news. Finally, it was established that deceptive articles possess a higher degree of formality and technicality in their writing, while credible articles were less so. This final result was the most surprising considering preconceptions of credible and deceptive news. As previous research involving readability metrics have proposed, deceptive news tends to require lower grade levels to comprehend (Afroz et al., 2012; Frank et al., 2008; Horne & Adali, 2017; Pérez-Rósas et al., 2017; Yang et al., 2017). Given, perhaps biased, preconceptions of deceptive news, we conjure an image of a 'fake news' article which is informal, amateur, and inerudite. If all 'fake news' conformed to this archetype, classification would be a lot simpler. While some articles may conform to this preconception, others are more resourceful. Some writers may pick up on characteristics related to formal language and exploit them, possibly because it is believed that to that complexity is what news readers have become accustomed.

Investigating individual underused and overused variables revealed that credible news employs sentence relatives to expand on ideas and concepts, as in Example 12 in Section 4.5.1. Deceptive news introduces more entities through a profusion of nouns. While both deceptive news and credible news avoid the modification of verbs through adverbs, the use of adverbs was particularly infrequent in deceptive news.

Examining additional features of interest has permitted more detailed observations of discourse features within dependable and deceitful news. Credible news discourse demonstrated a more elaborative language, with productive usage of intensifiers and adjectives. It also showed more complexity and diversity, with more independent clausal coordination and a higher type-token ratio. In comparison, misinformation displayed a higher frequency of public verbs to convey assertion and commentary, as seen in Example 21 in Section 4.5.2. However, the type-token ratio of deceptive news revealed that it was lacking in lexical diversity.

Investigating correlations among z-scores revealed the susceptibility of average word length in deceptive news articles to factors such as nominalization and attributive adjectives. In deceptive articles, nominalization and attributive adjectives contribute the most to increasing the average word length of an article. It was also shown that prepositional phrases often appear with the present tense in deceptive news articles. Articles from reliable sources showed that nouns and analytic negation had a positive impact on each other's frequency, suggesting that credible news tends to discuss entities and events with the presence of negation. Nouns were also found to be uncommon in conjunction with contraction, indicating that combinations such as 'congress'll' and 'the president'd' are less likely to occur and that texts showing a high frequency of nouns will be less likely to show a high frequency of contraction. The absence of these constructions is an expected characteristic of sophisticated writing. Lastly, past tense verbs were frequently used with the third person pronoun. As credible news reports on real events, organizations and individuals, it is natural that these would be discussed in the past and referred to in the third person.

While credible news incorporates more information into its articles compared to deceptive news, it does so through using language which makes it accessible to a larger audience. Despite using being less technical than deceptive news, it does not wholly forfeit formality. The prominence of relative clauses in the articles of credible news is demonstrative of this formality, as these linguistic features are typical of formal written prose (Akinnaso, 1982). In contrast, deceptive news uses narrative language to construct a story, which it attempts to enforce by using technical language. It uses an authoritative tone, which it expresses through public verbs and conjuncts, to appear more credible to its audience. While nouns a regular feature of academic prose and formal writing (Biber et al., 2004), and is an important variable in both corpora, the results showed that nouns are more frequent in deceptive news in comparison to credible news.

5.2 Future Work

In this thesis the various ways in which deceptive news and credible news differ according to linguistic variation in discourse has been examined through the lens of MDA. The findings of the current study can be utilized and expanded upon in four ways.

First, the individual sources within each corpus may be analyzed with MDA to investigate significant differences within both varieties of news which may not have been identifiable in a combined corpus. This may help to reveal specific journalistic styles associated with each source which may influence the larger corpus. Further breaking down the deceptive news corpus into smaller corpora for satire, bias, propaganda, and misinformation may also reveal differences between deceptive news types.

It is also reasonable to assume that the headlines of deceptive news articles are distinct from credible articles based on sensationalism and vagueness, as previous research has suggested (Horne & Adali, 2017). Alongside the content of news articles, the headlines of deceptive news and credible news articles may also be inspected to confirm or disprove this hypothesis on a broader scale than previous studies. The titles of deceptive and credible news may also be distinct enough to distinguish credible articles from deceptive articles, which might allow readers to make judgments before being exposed to the contents of the articles themselves.

There is a likelihood that external circumstances may affect the discourse of both credible and deceptive news. In the case of deceptive news, the intention of the author may subtly influence how a deceptive news article is written. A deceptive story written for the purpose of propaganda may present quite differently than one written to generate ad revenue. Research into the motivation of news writers may also provide a new angle for distinguishing deceptive texts. Unfortunately, without explicit confirmation from the authors themselves, it is difficult to conclude what the author's intention was with certainty.

Another factor which may influence writing is the linguistic background of the authors. As mentioned in Section 1.2, many deceptive news writers are based out of the town of Veles in Macedonia. It is possible that other creators of deceptive news also identify as English as a Second Language (ESL) or even as a third or fourth language. Both deceptive and credible news sources include authors whose first language is not English. However, credible ESL journalists are more likely to live in the English speaking country their news agency is based out of, giving them more exposure and access to English language resources. Referring back to the work on multidimensional analysis of World Englishes by Xiao (2009), variation in English language usage from different language backgrounds is observable. The variation in writing between credible ESL journalists and deceptive ESL journalists is another angle which may prove to have some value in deception detection.

Linguistic variables of MDA can be combined with psycholinguistic features, such as those from LIWC (Pennebaker et al., 2001), to improve performance in classification tasks

with machine learning. While previous research has employed this approach using the Stanford tagger (Mihalcea & Strapparava, 2009; Pérez-Rósas et al., 2017), the linguistic features identified by MDA are more comprehensive and representative of the linguistic makeup of the text content. Furthermore, a classifier which takes dimension scores from MDA as features may also prove to have some value as a tool for automatic identification.

The comparisons conducted in Section 4.7 can also be recreated for subtypes within the deceptive and credible news registers. A shortcoming of the subtype analysis conducted in this thesis was the inability to retrieve the full subtype analysis scores from Biber (1988). Conducting this comparison with scores for all texts within a corpus and the results scaled appropriately would be valuable for determining variation among subtypes between both registers.

Finally, the principles of MDA can be applied to deception detection cross-linguistically and cross culturally. Misinformation is not an issue unique to North American English media. The threat of 'fake news' is prevalent in many countries, particularly in the Philippines and Brazil. In the Philippines, misinformation supporting President Duterte's administration has made rounds through Facebook, with the social media giant being blamed for its lack of urgent response (Chandran, 2018). Brazil has also expressed concerns over the effect of 'fake news' on elections (Londoño, 2018). Currently, research focusing on disinformation in languages and cultures other than those in the Anglo-media has been lacking.

5.3 Concluding Remarks

For consumers of news, this thesis introduced a variety of qualities characteristic of deceptive news which may assist readers in determining whether or not to trust a source.

Generally, deceptive news is narrative. It aims to convince by painting a realistic portrait of a scenario with information presented with an element of narration and inflated formality. Formality, technicality, and precise language are mechanisms for crafting a story of some event that is misleading, overblown, or may not have even occurred.

Linguistically, the use of relative clauses and excessive nouns assist in developing the fabrication by enhancing the details. The writers further assert their veracity with the use of witness or expert statements through higher usage of public verbs. It includes such statements from witnesses or experts to reinforce its narrative, whether or not these individuals made such comments or even exist. As convincing as a deceptive article may be, it is not trustworthy. Often, these articles are written by non-native speakers, trolls, or political fanatics, all of which lack the funds and expertise of recognized credible sources. This is visible in the lexical diversity of deceptive content, which is substantially lacking in comparison to its legitimate counterpart.

However, rather than memorizing fine details of deceptive discourse, or researching the history of authors and sources, it is most important to be aware. As many deceivers are

attempting to manipulate their readers into accepting their viewpoint, looking for manipulation strategies in text may provide some protection. Deceivers often use many details in order to convince both the audience and themselves (de Becker, 1997). Even before the sobriquet of 'fake news', it was commonly understood that no individual newspaper was to be trusted entirely. Referring to multiple sources and forming one's own informed opinion on a matter was, and still is, the most sensible option. As Virginia Woolf wrote of newspapers in *Three Guineas*, "Compare the views, make allowance for the distortions, and then judge for yourself" (Woolf, 1938).

In the current political climate, the distinction of what constitutes 'fake news' is critical. Politicians and the devoutly political are often quick to dismiss even verified information as false when it does not suit them or conform to their view of the world. Donald Trump and his supporters have been twisting words to cast doubt on facts, defame groups, and manipulate the media (Lakoff & Duran, 2018). Diminishing the truth in favor of one's biases in this way may further the partition within a population, where all sides are left with inadequate understanding. With this is mind, it is important to have a clear definition and understanding of what deceptive news is and how it is different from credible news.

The resolution toward compartmentalizing deceptive and trustworthy news is based on the discrepancies observed through multidimensional analysis of the discourse of both genres. The variation in discourse between reliable news and fake news corpora presented in this thesis offers sufficient foundation for the distinction of two separate genres in online press reportage, fake news and reliable news.

References

Afroz, S., Brennan, M., & Greenstadt, R. (2012, May). Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In *2012 IEEE Symposium on Security and Privacy* (pp. 461–475).

Agorist, M. (2017). *Disturbing Video Surfaces Showing Cop Smoking Cigar Kneeling Over Lifeless Body.* The Free Thought Project. Retrieved from `http://thefreethoughtproject.com/disturbing-video-surfaces-showing-cop-smoking-cigar-next-to-lifeless-body/`

Akinnaso, F. N. (1982). On the Differences Between Spoken and Written Language. *Language and Speech*, *25*(2), 97–125.

Allcott, H., & Gentzkow, M. (2017). *Social Media and Fake News in the 2016 Election* (Tech. Rep. No. w23089). National Bureau of Economic Research. Retrieved from `http://www.nber.org/papers/w23089`

Arciuli, J., Mallard, D., & Villar, G. (2010). "Um, I can Tell You're Lying": Linguistic Markers of Deception versus Truth-telling in Speech. *Applied Psycholinguistics*, *31*(3), 397–411. Retrieved from `https://www.cambridge.org/core/journals/applied-psycholinguistics/article/um-i-can-tell-youre-lying-linguistic-markers-of-deception-versus-truthtelling-in-speech/F1290722AC7CAD345C541988D9C1DBBD`

Bajaj, S. (2017). "The Pope Has a New Baby!" Fake News Detection Using Deep Learning. *Stanford University.*

Beauchamp, Z. (2015). *America's Never Been Safer. So Why are Republicans Convinced It's in Mortal Peril?* Vox. Retrieved from `http://www.vox.com/2015/6/3/8716261/gop-primary-threats`

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs Conspiracy: Collective Narratives in the Age of Misinformation. *PLoS One*, *10*(2), e0118093. Retrieved from `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118093`

Biber, D. (1988). *Variation Across Speech and Writing.* New York, NY: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison.* New York, NY: Cambridge University Press.

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style.* New York, NY: Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical Bundles in University Teaching and Textbooks. *Applied linguistics*, *25*(3), 371–405.

Biber, D., & Egbert, J. (2016). Register Variation on the Searchable Web. *Journal of English Linguistics*. Retrieved from `http://journals.sagepub.com/doi/abs/10.1177/0075424216628955`

Borowitz, A. (2017). *Clinton Compiles Mental List of People to Destroy.* The New Yorker. Retrieved from `https://www.newyorker.com/humor/borowitz-report/clinton-compiles-mental-list-of-people-to-destroy`

Burfoot, C., & Baldwin, T. (2009, Aug). Automatic Satire Detection: Are You Having a Laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 161–164). Association for Computational Linguistics.

Cambridge Analytica offices searched over data storage. (2018). BBC. Retrieved from `http://www.bbc.com/news/uk-43522775`

Chafe, W. (1982). Integration and Involvement in Speaking, Writing, and Oral Literature. *Spoken and Written Language: Exploring Orality and Literacy*, 35–54.

Chafe, W. (1984, Feb). How People Use Adverbial Clauses. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 10, pp. 437–449). eLanguage.

Chalker, S. (2003). *The Oxford Dictionary of English Grammar: 1000 Entries.* Oxford, United Kingdom: Oxford University Press.

Chandran, N. (2018). *Facebook Attacked by Critics Over 'Fake News' - but outside the US this Time.* CNBC. Retrieved from `https://www.cnbc.com/2018/01/17/facebook-and-fake-news-controversy-in-philippines-around-rappler.html`

Chen, A. (2017). The Fake-News Fallacy. *The New Yorker*. Retrieved 2017-09-20, from `https://www.newyorker.com/magazine/2017/09/04/the-fake-news-fallacy`

Chen, D., & Manning, C. (2014). A Fast and Accurate Dependency Parser Using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 740–750).

de Becker, G. (1997). *The Gift of Fear: And Other Survival Signals that Protect Us from Violence*. New York, NY: Dell Publishing.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC* (Vol. 6, pp. 449–454).

Dewey, C. (2016). Facebook fake-news writer: 'I think Donald Trump is in the White House because of me'. *Washington Post*. Retrieved from `https://www.washingtonpost.com/news/the-intersect/wp/2016/11/17/facebook-fake-news-writer-i-think-donald-trump-is-in-the-white-house-because-of-me/`

Durando, J. (2017). *U.S. adds Osama bin Laden's Son to Global Terrorist List.* Gannett Satellite Information Network. Retrieved from `https://www.usatoday.com/story/news/world/2017/01/05/terrorism-sanctions-osama-bin-laden-son/96197152/`

Eggins, S. (2004). *Introduction to Systemic Functional Linguistics.* A&C Black, address = New York, NY.

Ekman, P., O'Sullivan, M., & Frank, M. G. (1999). A Few Can Catch a Liar. *Psychological Science*, *10*(3), 263–266.

Enos, F., Benus, S., Cautin, R. L., Graciarena, M., Hirschberg, J., & Shriberg, E. (2006, Sep). Personality Factors in Human Deception Detection: Comparing Human to Machine Performance. In *Ninth International Conference on Spoken Language Processing.* doi: 10.1.1.448.5012

Exposed: Undercover secrets of Trump's data firm. (2018). Channel 4. Retrieved from `https://www.channel4.com/news/exposed-undercover-secrets-of-donald-trump-data-firm-cambridge-analytica`

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., ... Treinen, M. (2001, Mar). What Makes Web Sites Credible?: a Report on a Large Quantitative Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 61–68). ACM. Retrieved from `http://dl.acm.org/citation.cfm?id=365024.365037`

Fowler, R. (2013). *Language in the News: Discourse and Ideology in the Press.* New York, NY: Routledge.

Frank, M. G., Menasco, M. A., & O'Sullivan, M. (2008). *Human Behavior and Deception Detection.* John Wiley & Sons. doi: 10.1002/9780470087923

Gorsuch, R. L. (1983). *Factor Analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gottfried, J., & Shearer, E. (2016). News Use Across Social Media Platforms 2016. *Pew Research Center*, *26*.

Graham, R. (2017). Google and Advertising: Digital Capitalism in the Context of Post-Fordism, the Reification of Language, and the Rise of Fake News. *Palgrave Communications*, *3*(1), 45.

Hagen, S. (2017). *Opensources.* GitHub. Retrieved from `https://github.com/BigMcLargeHuge/opensources` (Github Repository)

Hagen, S., & Lutzenberger, T. (2016). *B.S. Detector.* The Self Agency, LLC. Retrieved from `http://bsdetector.tech` (Browser Add-on)

Hambrick, D. Z. (2018). *Cognitive Ability and Vulnerability to Fake News.* Scientific American. Retrieved from `https://www.scientificamerican.com/article/cognitive-ability-and-vulnerability-to-fake-news/`

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On Lying and Being Lied to: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes*, *45*(1), 1–23.

Ho, S. M., Hancock, J. T., Booth, C., Liu, X., Timmarajus, S. S., & Burmester, M. (2015, May). Liar, Liar, IM on Fire: Deceptive Language-action Cues in Spontaneous Online Communication. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 157–159).

Horne, B. D., & Adali, S. (2017). This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. *CoRR*, *abs/1703.09398*. Retrieved from `https://arxiv.org/abs/1703.09398`

Hu, X. (2016). Mining Misinformation in Social Media. In *Big Data in Complex and Social Networks* (pp. 123–152). CRC Press. Retrieved from `https://books.google.com/books/about/Big_Data_in_Complex_and_Social_Networks.html?id=CA4NDgAAQBAJ`

Hu, Y., Song, R. J., & Chen, M. (2017). Modeling for Information Diffusion in Online Social Networks via Hydrodynamics. *IEEE Access*, *5*, 128–135.

Huntley, A. (2017). *Newfoundland Sunk After Collision with Iceberg.* Retrieved from https://www.thebeaverton.com/2017/04/newfoundland-sunk-collision-iceberg/

Isaac, M. (2016). *Facebook, in Cross Hairs After Election, Is Said to Question Its Influence.* Retrieved from https://www.nytimes.com/2016/11/14/technology/facebook-is-said-to-question-its-influence-in-election.html

Kiely, E., & Robertson, L. (2016). *How to Spot Fake News.* FactCheck.org. Retrieved from https://www.factcheck.org/wp-content/cache/wp-rocket/www.factcheck.org/2016/11/how-to-spot-fake-news//index.html_gzip

Kruzel, J. (2018). *Trump's Fake News Awards Could Land Staffers in Hot Water.* Politifact. Retrieved from http://www.politifact.com/truth-o-meter/statements/2018/jan/11/norm-eisen/donald-trumps-fake-news-awards-could-land-white-ho/

Kumar, S., West, R., & Leskovec, J. (2016, Apr). Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In (pp. 591–602). International World Wide Web Conferences Steering Committee. Retrieved from http://dl.acm.org/citation.cfm?id=2872427.2883085

Lakoff, G. P., & Duran, G. (2018). *Trump Has Turned Words into Weapons. And He's Winning the Linguistic War.* The Guardian News and Media. Retrieved from https://www.theguardian.com/commentisfree/2018/jun/13/how-to-report-trump-media-manipulation-language

Lazer, D., Baum, M., & Mele, N. (2017, Feb). Combating Fake News: An Agenda for Research and Action. Harvard University, Northeastern University, Cambridge, MA: North Eastern University, Harvard University.

Leistyna, P., & Meyer, C. F. (2003). *Corpus Analysis: Language Structure and Language Use* (Vol. 46). New York, NY: Rodopi.

Levine, T. R. (2014). Truth-Default Theory (TDT) A theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 378–392.

Lindsey, B. (2017). *Oh My GOD! Trump Just PROVED He's the People's President! What He Did On Stage Will Bring You To TEARS!* Retrieved from https://libertywriters.com/2017/02/trump-just-proved-he-peoples-president-stage-will-bring-tear/

Londoño, E. (2018). *Brazil Looks to Crack Down on Fake News Ahead of Bitter Election.* The New York Times. Retrieved from https://www.nytimes.com/2018/02/17/world/americas/brazil-election-fake-news.html

Lopez, G. (2014). *Department of Justice to Announce Findings of Investigation into Cleveland Police.* Vox. Retrieved from http://www.vox.com/xpress/2014/12/4/7335025/cleveland-police-doj

Love, J., & Cooke, K. (2016). *Google, Facebook move to restrict ads on fake news sites.* Retrieved from https://www.reuters.com/article/us-alphabet-advertising/google-facebook-move-to-restrict-ads-on-fake-news-sites-idUSKBN1392MM

Lytvynenko, J. (2017). *If You Get 41/55 On This Quiz, Fake News Didn't Fool You This Year.* BuzzFeed. Retrieved from https://www.buzzfeed.com/janelytvynenko/massive-2017-fake-news-quiz?utm_term=.bcdEP5BgWq#.opxjEvx9ZK

MacIntyre, I. (2016). *Brazil Assures World that Country Will be Only be Partly on Fire by Olympics.* Retrieved from https://www.thebeaverton.com/2016/06/brazil-assures-world-that-country-will-be-only-be-partly-on-fire-by-olympics/

Martin, J. (1991). Distilling Knowledge and Scaffolding Text. *Functional and Systemic Linguistics: Approaches and Uses*, *55*, 307.

Marzilli, T. (2018). *Old and Young US Adults Most Susceptible to Fake News.* YouGov. Retrieved from https://today.yougov.com/topics/media/articles-reports/2018/06/13/old-and-young-us-adults-most-susceptible-fake-news

McKinney, K. (2014). *Fashion Designer Stella McCartney Says Strong Women Aren't Attractive.* Vox. Retrieved from http://www.vox.com/xpress/2014/10/1/6875259/Amal-Almuddin-Stella-McCartney-body-image-problem

Mihalcea, R., & Strapparava, C. (2009, Aug). The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 309–312). Retrieved from http://dl.acm.org/citation.cfm?id=1667583.1667679

Nelson, L. (2017). *Watch Rachel Maddow Reveal Part of President Trump's Tax Returns.* Vox. Retrieved from http://www.vox.com/2017/3/14/14930662/trump-tax-returns-rachel-maddow-watch-video

Nini, A. (2015). *Multidimensional Analysis Tagger (Version 1.3).* Retrieved from http://sites.google.com/site/multidimensionaltagger (Software)

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011, Jun). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 309–319). Association for Computational Linguistics.

Parkinson, H. J. (2016). Click and Elect: How Fake News Helped Donald Trump Win a Real Election. *The Guardian*, *14*. Retrieved 2018-01-14, from http://www.theguardian.com/commentisfree/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies

Pelley, S. (2017). *How Fake News Becomes a Popular, Trending Topic.* CBS News. Retrieved from https://www.cbsnews.com/news/how-fake-news-find-your-social-media-feeds/

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001. (Software)

Pérez-Rósas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic Detection of Fake News. *arXiv preprint arXiv:1708.07104*. Retrieved from https://arxiv.org/abs/1708.07104

Pérez-Rósas, V., & Mihalcea, R. (2015). Experiments in Open Domain Deception Detection. , 1120–1125. Retrieved from http://anthology.aclweb.org/D/D15/D15-1133.pdf

Porat, R. (2017). *Alphabet Investor Relations: 2016 Annual Report.* Mountain View, CA: Alphabet Inc.

Radutzky, M. (2017). *What's "fake news"? 60 Minutes producers investigate.* 60 Minutes. Retrieved from https://www.cbsnews.com/news/whats-fake-news-60-minutes-producers-investigate/

Rannard, G. (2017). Harvey Fake News Imam Never Visited Texas. Retrieved from http://www.bbc.com/news/blogs-trending-41147572

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017, Sep). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2931–2937). Retrieved from http://www.aclweb.org/anthology/D17-1316

Read, M. (2016). *Donald Trump Won Because of Facebook.* NYMag. Retrieved from http://nymag.com/selectall/2016/11/donald-trump-won-because-of-facebook.html

Risdal, M. (2016). Getting Real about Fake News. Retrieved from https://www.kaggle.com/mrisdal/fake-news (Dataset)

Roberts, H. (2016). *This is What Fake News Actually Looks Like - We Ranked 11 Election Stories that Went Viral on Facebook.* Business Insider.

Retrieved from http://uk.businessinsider.com/fake-presidential-election-news-viral-facebook-trump-clinton-2016-11

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2017, Jun). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of NAACL-HLT* (pp. 7–17). Retrieved from http://www.aclweb.org/anthology/W16-0802

Schleppegrell, M. J. (2001). Linguistic Features of the Language of Schooling. *Linguistics and Education*, *12*(4), 431–459.

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The Spread of Misinformation by Social Bots. *arXiv:1707.07592 [cs.SI]*. Retrieved from https://arxiv.org/abs/1707.07592

Soll, J. (2016). *The Long and Brutal History of Fake News.* Retrieved 2018-01-15, from https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535

Sorkin, A. D. (2017). *What Gillibrand Got Wrong About Military-Sexual Assault.* The New Yorker. Retrieved from https://www.newyorker.com/news/amy-davidson/what-gillibrand-got-wrong-about-military-sexual-assault

Spinney, L. (2017). The Shared Past That Wasn't. *Nature Publishing Group*, *543*, 168–170.

Subramanian, S. (2017). *Meet the Macedonian Teens Who Mastered Fake News and Corrupted the US Election.* WIRED. Retrieved from https://www.wired.com/2017/02/veles-macedonia-fake-news/

Sydell, L. (2016). *We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned* (Vol. 23). Retrieved from https://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs

Tynan, D. (2016). *How Facebook Powers Money Machines for Obscure Political 'News' Sites.* The Guardian. Retrieved from https://www.theguardian.com/technology/2016/aug/24/facebook-clickbait-political-news-sites-us-election-trump

Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017, Jul). Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 647–653). Retrieved from https://scholar.google.ca/scholar?q=Separating+Facts+from+Fiction%3A+Linguistic+Models+to+Classify+Suspicious+and+Trusted+News+Posts+on+Twitter&btnG=&hl=en&as_sdt=0%2C5

Vosoughi, S., Roy, D., & Aral, S. (2018). The Spread of True and False News Online. *Science*, *359*(6380), 1146–1151.

Wang, A. B. (2017). *ABC News apologizes for 'serious error' in Trump report and suspends Brian Ross for four weeks.* Retrieved from `https://www.washingtonpost.com/news/arts-and-entertainment/wp/2017/12/03/abc-news-apologizes-for-serious-error-in-trump-report-suspends-brian-ross-for-four-weeks/?noredirect=on`

Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. Retrieved from `https://arxiv.org/abs/1705.00648`

Woolf, V. (1938). *Three Guineas.* New York, NY: Harcourt, Brace and Co.

Xiao, R. (2009). Multidimensional Analysis and the Study of World Englishes. *World Englishes*, *28*(4), 421–450.

Yang, F., Mukherjee, A., & Dragut, E. (2017). Satirical News Detection and Analysis using Attention Mechanism and Linguistic Features. *arXiv preprint arXiv:1709.01189*. Retrieved from `http://arxiv.org/abs/1709.01189`

Zheleva, E. (2017). Vox Articles published before March 21, 2017. Retrieved from `https://data.world/elenadata/vox-articles` (Dataset)

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., & Tolmie, P. (2016). Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLoS One*, *11*(3), e0150989. Retrieved from `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150989`

Appendix A

Z-Score Variables

Table A.1: Linguistic variables, the mean z-scores for each corpus, and its associated Biber dimension. Features that were flagged by MAT bear an asterisk.

Feature	Deceptive	Credible	Biber Dimension	Flagged
Amplifiers	-0.5	-0.4	D1	
Independent clause coordination	0.1	0.9	D1	
Average word length	0.6	0.8	D1, D2	
'Be' as a main verb	-1.3	-1.1	D1	
By-passives	0.1	0	D5	
Causative adverbial subordinators	-0.1	0.1	D1	
Concessive adverbial subordinators	0	0.2		
Conditional adverbial subordinators	-0.2	-0.1	D4	
Conjuncts	0.1	0.2	D5	
Contractions	-0.7	-0.3	D1	
Demonstratives	-0.3	-0.3	D1	
Demonstrative pronouns	-0.8	0	D1	
Discourse particles	-0.3	-0.4	D1	
Downtoners	-0.3	0.1		
Emphatics	0	0.6	D1	
Existential 'there'	-0.5	-0.2	D6	
First person pronouns	-0.6	-0.6	D1	
Gerunds	-0.7	-0.6		
Hedges	-0.3	-0.2	D1	
Indefinite pronouns	-0.5	-0.5	D1	
Attributive adjectives	-0.1	0.5	D1, D2	
Necessity modals	-0.5	-0.5	D4	
Total other nouns	3.9	3.1	D1	*
Nominalizations	0.1	0.2	D3	
Other adverbial subordinators	0.7	0.5	D5, D3	
Agentless passives	-0.2	-0.5	D5, D1	
Past participial clauses	0.8	1.3	D2, D5	
Perfect aspect	-0.4	-0.4	D2	

Phrasal coordination	1	1	D3, D6	
Total prepositional phrases	-0.7	-0.6	D1	
Pied-piping relative clauses	-0.3	-0.1	D3	
Pronoun 'it'	-0.3	0.1	D1	
Place adverbials	0	0	D1, D3	
Possibility modals	-0.5	-0.1	D1, D4	
Predicative adjectives	0.3	0.9	D1	
Present participial clauses	1	0.7		
Private verbs	-0.5	-0.6		
Predictive modals	-0.2	-0.1	D4	
Pro-verb 'do'	-0.5	-0.5	D1	
Public verbs	1	0	D2	
Total adverbs	-2.2	-1.8		*
Sentence relatives	2.3	3.4	D1	*
'Seem' and 'appear'	-0.1	0.3		
Split auxiliaries	-0.7	-0.7		
Second person pronouns	-0.4	-0.4	D1	
Stranded prepositions	-0.4	-0.4	D1	
Suasive verbs	0.3	0.2	D4	
Synthetic negation	-0.3	-0.3	D2	
'That' adjective complements	0.2	0.4	D6	
Subordinator 'that' deletion	0.2	-0.1	D1	
'That' verb complements	0.3	0	D6	
Time adverbials	-0.2	-0.3	D3	
Infinitives	0.5	0.4	D4	
'That' relative clause on object	1.2	1	D6	
Third person pronouns	-0.1	-0.4	D2	
'That' relative clauses on subject	1.9	2.8		*
Type-token ratio	-1.6	0.1	D1, D5	
Past tense	0.2	-0.2	D2	
Present tense	-0.6	-0.6	D1, D2	
Wh-clauses	0.2	0.2	D1	
Wh-clauses on object	-0.7	-0.7	D3	
Direct wh-questions	0.1	0.3	D1	
Wh-clauses on subject	-0.2	-0.2	D3	
Past participial WHIZ deletion	-0.4	-0.4	D5, D1, D2	
Present participial WHIZ deletion	1.1	0.6	D1	
Analytic negation	-0.7	-0.3	D1	

Appendix B

Text Type Comparisons

The following charts map the credible corpus and deceptive corpus with other types of text. The corpus is plotted with the x axis being the approximate score, and the y axis being the textual dimension. The other text types marked in the charts are 'intimate interpersonal interaction', 'informational interaction', 'imaginative narrative', 'situated reportage', 'involved persuasion', 'general narrative exposition', 'learned exposition', and 'scientific exposition'. The corpora are marked by a cyan dot in their respective chart. Appendix B.1 shows the deceptive corpus, while appendix B.2 shows the credible corpus. At the top of the chart, the closest text type to the corpus is noted, which has been identified as 'general narrative exposition' for both corpora.

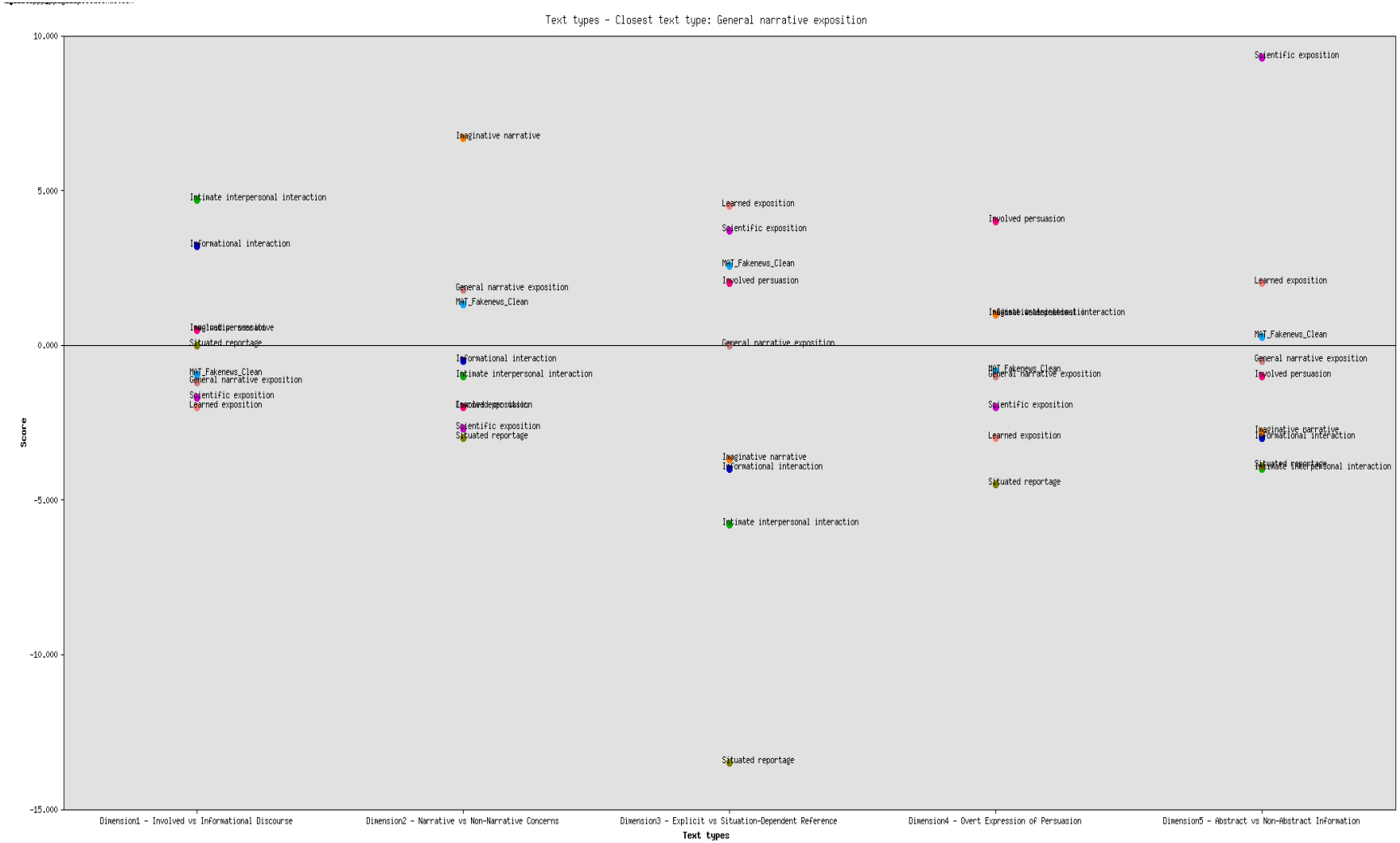

Figure B.1: The results of the fake news corpus compared to Biber's text types. The corpus is signified by a light blue dot with the label 'MAT-Fakenews-Clean',

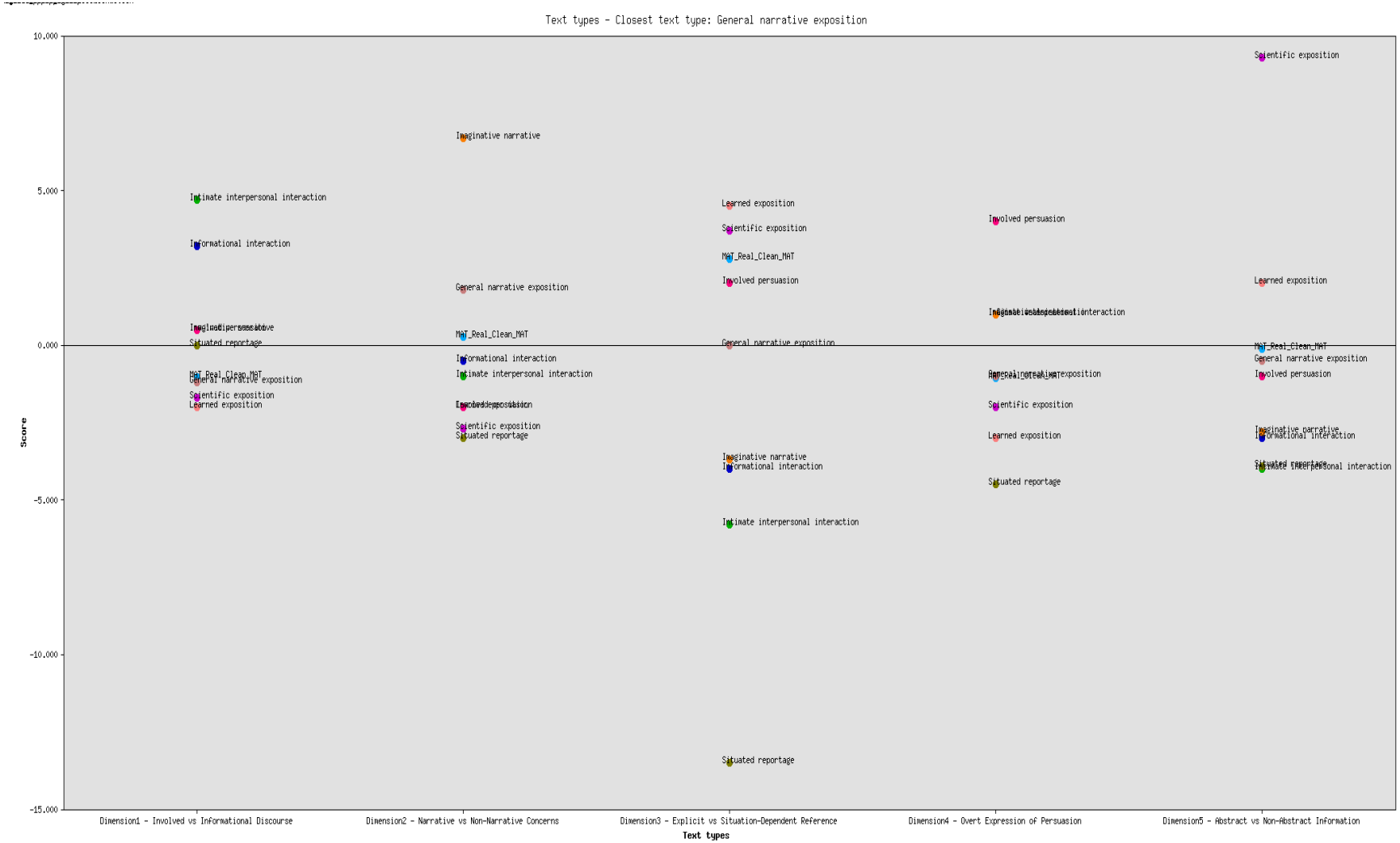

Figure B.2: The results of the reliable news corpus compared to Biber's text types. The corpus is signified by a light blue dot with the label 'MAT-Real-Clean'.