2019

Application of Boolean Logic to Natural Language Complexity in Political Discourse

Austin Taing
University of Kentucky, austin.taing001@gmail.com

Digital Object Identifier: https://doi.org/10.13023/etd.2019.131

Recommended Citation

Taing, Austin, "Application of Boolean Logic to Natural Language Complexity in Political Discourse" (2019). *Theses and Dissertations--Computer Science*. 77.
https://uknowledge.uky.edu/cs_etds/77

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Austin Taing, Student

Dr. Judy Goldsmith, Major Professor

Dr. Miroslaw Truszczynski, Director of Graduate Studies

Application of Boolean Logic to Natural Language Complexity in Political Discourse

THESIS

A thesis submitted in partial
fulfillment of the requirements for
the degree of Master of Science in
the College of Engineering at the
University of Kentucky

By
Austin Taing
Lexington, Kentucky

Director: Dr. Judy Goldsmith, Professor of Computer Science
Lexington, Kentucky 2019

ABSTRACT OF THESIS

Application of Boolean Logic to Natural Language Complexity in Political Discourse

Press releases serve as a major influence on public opinion of a politician, since they are a primary means of communicating with the public and directing discussion. Thus, the public's ability to digest them is an important factor for politicians to consider. This study employs several well-studied measures of linguistic complexity and proposes a new one to examine whether politicians change their language to become more or less difficult to parse in different situations. This study uses 27,500 press releases from the US Senate between 2004–2008 and examines election cycles and natural disasters, namely hurricanes, as situations where politicians' language may change. We calculate the syntactic complexity measures *clauses per sentence*, *T-unit length*, and *complex-T ratio*, as well as the *Automated Readability Index* and *Flesch Reading Ease* of each press release. We also propose a proof-of-concept measure called *logical complexity* to find if classical Boolean logic can be applied as a practical linguistic complexity measure. We find that language becomes more complex in coastal senators' press releases concerning hurricanes, but see no significant change for those in election cycles. Our measure shows similar results to the well-established ones, showing that logical complexity is a useful lens for measuring linguistic complexity.

KEYWORDS: linguistic complexity, readability, natural language processing, logical complexity

Author's signature: Austin Taing

Date: May 3, 2019

Application of Boolean Logic to Natural Language Complexity in Political Discourse

By
Austin Taing

Director of Thesis: Judy Goldsmith

Director of Graduate Studies: Miroslaw Truszczynski

Date: May 3, 2019

ACKNOWLEDGMENTS

This thesis, while an individual work, was supported by academic contributions from a number of people. Dr. Judy Goldsmith, my advisor, provided the original impetus for the work, which took a number of turns before taking the form of this thesis. Dr. Goldsmith also provided comments and guidance towards other resources which I could use to further the work, including thesis committee member Dr. Justin Wedeking, whose previous related work helped contribute to a useful launching point for my own. Finally, the whole of the thesis committee, Dr. Goldsmith, Dr. Wedeking, and Dr. Brent Harrison, provided insightful feedback and asked poignant questions which helped improve the final product.

I would also like to briefly acknowledge the support of a great many other individuals over the course of my work. Friends, family, and others gave me the support needed to fully apply myself and create this thesis. I thank all of them for the aid they provided me.

TABLE OF CONTENTS

Acknowledgments . . . iii

Table of Contents . . . iv

List of Tables . . . v

List of Figures . . . v

Chapter 1 Introduction . . . 1
  1.1 Background and Motivation . . . 1
  1.2 Layout of Thesis . . . 2

Chapter 2 Related Work . . . 3

Chapter 3 Measuring the complexity of political text . . . 4
  3.1 Syntactic complexity . . . 4
  3.2 Readability . . . 5
  3.3 Logical complexity . . . 6

Chapter 4 Data set . . . 7
  4.1 Classifying press releases . . . 7
  4.2 Experimental groups . . . 7

Chapter 5 Experimental Design . . . 8
  5.1 Syntactic and logical complexity . . . 8
  5.2 Readability . . . 8

Chapter 6 Results . . . 9
  6.1 Syntactic complexity and readability results . . . 9
  6.2 Logical complexity results . . . 10
  6.3 Discussion . . . 10

Chapter 7 Conclusions . . . 18
  7.1 Future Work . . . 19

Chapter A Appendices . . . 20
  A.1 Examples from the corpus . . . 20
  A.2 Logical complexity keywords . . . 22

Bibliography . . . 23

Vita . . . 25

LIST OF TABLES

6.1 Syntactic complexity results for reelection senators 11
6.2 Readability results for reelection senators 12
6.3 Syntactic complexity results for coastal senators 13
6.4 Readability results for coastal senators 13
6.5 Logical complexity results for reelection senators 14
6.6 Logical complexity results for coastal senators 14

LIST OF FIGURES

6.1 Syntactic complexity results for election group 15
6.2 Readability results for reelection group 15
6.3 Syntactic complexity results for coastal group 16
6.4 Readability results for coastal group 16
6.5 Logical complexity results for both groups 17

Chapter 1 Introduction

In this thesis, we attempt to determine whether the linguistic complexity of political discourse changes as a result of events such as election cycles; that is, whether politicians' language becomes easier or harder to understand at these times. We employ several well-studied measures of complexity in this context and propose a new measure, *logical complexity*, to attempt to find such a shift. We also compare our method to the established measures to determine its effectiveness in this context.

1.1 Background and Motivation

For most people, political discourse is not something with which we interact directly; instead, we obtain information and form opinions based on the reports of the press. While there is variance in the sources of such reports, statements that come directly from politicians represent the intent of those politicians in their interaction with the populace. Recently, colleagues Cory Siler, Luke Miles, and Judy Goldsmith explored a theoretical approach to modeling politicians' statements in terms of attracting a voter base [20], as described by Dean, Parikh, and Taşdemir [3, 16]. Dean et al. proposed that politicians speak in language equivalent to Boolean formulae, comprised of individual variables representing positions on specific policy issues. Their model examines different ways for politicians to assign utility to statements based on their appeal to a voter base, in which each voter's opinions are represented by their own Boolean formula with individual utility for each variable. Siler et al. showed that calculating such utility is computationally intractable; Goldsmith conjectured that politicians instead employ a simpler model, such as 2SAT or Horn formulae [7].

This inspired a related question: how complex are those statements in the real world? While exploring ways to tackle this question, we encountered several different approaches to examining complexity through computational linguistics, though none answer this question directly.

Outside of platform documents which detail the stance of a politician on a large number of issues, concise declarations that fit the theoretical Boolean variable description are rare. Instead, the media from which we obtain information is in the form of natural language, expressing opinions in a less direct way. Thus, we choose to examine political discourse through the lens of linguistic complexity in order to bridge the gap to the logical formalism described in the work of Siler et al. However, simply applying measures of complexity to a body of political discourse tells us little of use. Therefore, this study seeks to answer the question of whether politicians' language changes in complexity during the election cycle or in response to significant events.

We choose to examine election cycles as an extension of the theoretical work by Siler et al. regarding politicians' interaction with voter bases. We seek to find whether an effort is made during these times by politicians to communicate or interact with the populace in a different way in order to increase their chances at reelection.

We also choose to examine significant events, as those are prime opportunities for politicians to interact with the populace. For this study, we isolate natural disasters as a representative of significant events. In this work, we introduce a new approach to computing language complexity, *logical complexity*, as a first step towards applying the theory of Boolean logic to linguistic complexity. We use classic measures of syntactic complexity and textual readability as baselines for comparison of this approach.

We hypothesize that to reach the wide audience demanded at stressful times like election cycles and especially natural disasters, the language of press releases becomes more simple. This will reflect in the readability of the text in both cases. However, we hypothesize that logical and syntactic complexity will decrease for natural disasters, but will be largely unaffected by election cycles.

1.2 Layout of Thesis

The thesis begins in Section 2 by examining previous work to establish an academic background for the study. We discuss the relevance of the work of other scholars and their applications to the problem at hand. In Section 3, we describe and define each of the measures of complexity we will use to test our earlier hypothesis as well as introduce our own, *logical complexity*, whose veracity is the other subject of our experiment. Examples are included to illustrate the concepts as they pertain to our experiment. We briefly delve into the specifics of the corpus used for this study, highlighting its relevance and potential limitations in Section 4. Section 5 defines the parameters under which the experiment was run, followed by the results of that experiment in Section 6. Qualitative and quantitative examination of the results, including visualization aids, are given. Finally, we conclude in Section 7 by discussing the meaning of the results in the context of the corpus and experimental design, highlighting potential limitations of the experiment and areas for further study.

Example passages from the corpus and their processed forms for several major sentence forms are present in the appendix. The appendix also contains descriptions of corpus-specific issues that are touched on within the main discussion but do not pertain directly to the discussion at hand.

Chapter 2 Related Work

The term *syntactic complexity* can be used to describe a number of different approaches to measuring text complexity. Early measures of syntactic structure include those defined by Hunt in 1965 [10], including the measures used in this study. These measures have been applied to areas such as language mastery of elementary and middle school students [1, 10], and most prominently in English-as-a-second-language learner development [12, 21]. Lu et al. previously built a syntactic analyzer, the L2 parser, built off the Stanford lexical parser, based on this work with second-language English text (indicated by the L2 abbreviation) [12]. It focuses on classic linguistic measures similar to those used in this study for syntactic complexity.

Readability is a well-researched measure of linguistic complexity [5, 11, 18], and has been applied within the domain of political discourse [2]. This study includes the same measure Siegal & Siegal (1953) applied to speeches by the candidates in the 1952 presidential race [19], *Flesch reading ease*. More recently, Gyasi (2017) used readability measures in regard to press releases to evaluate communication with the public by political parties in Ghana [9]. This study applies the same readability index, *Automated Readability Index*, on a new corpus. By using multiple measures, this study seeks to examine similar text using a wider lens.

Previous other work has also approached to analyzing overall patterns through psychological meaning analysis. Pennebaker et al. have done substantial work in this area, including their work with the Linguistic Inquiry and Word Count (LIWC) tool [17, 22]. This method focuses on comparing appearance rates of predetermined keywords to derive categorical measures from the text, without directly considering the structure or ordering of those words. *Cognitive complexity*, one of the composite measures studied by Pennebaker et al., serves as an interesting comparison point to logical complexity. Its measurement focuses specifically on *exclusion* and *integration* words, which are conceptually similar to the *and/or* used to describe classical logic statements. This method was used by Owens et al. (2011) to examine political discourse in the form of US Supreme Court decisions [15]. As we will discuss in greater detail in the next section, our logical complexity formulation builds upon this framework by taking into account the structure of the text in addition to the frequency of keywords.

Chapter 3 Measuring the complexity of political text

Complexity in this context is not so clearly defined; natural language has many facets that can be examined for some notion of complexity. Therefore, to get a grasp of holistic complexity in discourse, we will use three primary metrics that represent different facets of linguistic communication: syntactic complexity, logical complexity, and readability. *Syntactic complexity* is a linguistic measure of how much information is presented in a sentence, and how the component words relate to each other. *Readability* is the difficulty for a reader to understand the text, codified in computational linguistics. *Logical complexity* examines text as a collection of Boolean formulae—e.g. *a AND b, but NOT c*—as described in 1.2; formulae with more elements are more complex. We put forth an approach which aims to approximate the number of such elements relative to the length of the text.

3.1 Syntactic complexity

Syntactic complexity is a well-studied area in linguistics [6, 10, 12, 21], focusing on the structural composition of a sentence. Measures of syntactic complexity vary in the amount of granularity examined and the type of structures examined, but generally focus on the amount of information contained in structures of a given scope. For the purpose of this study, we focus on three primary measures: T-unit length, clauses per sentence, and complex T-unit ratio.

A *T-unit* is defined as the minimally terminable unit into which writing can be split [10]. In English, a T-unit consists of an independent dominant clause and its accompanying subordinate clauses. Using T-units as a base information unit allows for correcting potentially ill-constructed sentences in the original text, as well as representing a consistent frame of reference for information presented to the reader. A *complex* T-unit, then, is a T-unit with one or more dependent clauses in addition to its dominant clause. Taken together with the number of clauses per sentence, the ratio of complex T-units to the total number of T-units (hereafter simply referred to as the *complex-T ratio*) gives a good estimation of how much information is encoded in each readily digestible segment of the writing. The following paragraphs detail the implementation of these measures, based on the work of Lu et al. [12]. We use the Stanford CoreNLP toolset [13] to construct the relevant parse trees. A full list of tags and their meaning can be found at the end of this document.

First, we define each parse tree as a sentence. Thus, each sentence has exactly one ROOT node, which immediately dominates a *clausal node*: S, SINV, SQ, or SBARQ. The clausal node must also contain a verb phrase (VP) as one of its children. Furthermore, to constitute an independent clause, the S-node must either have another S-node as a sibling or be dominated directly by the ROOT node. However, in the case of a sentence fragment, we treat the whole fragment as a single clause. A clausal node represents a dependent clause when it is immediately dominated by an SBAR node, indicating a clause introduced by a subordinating conjunction. In either case, the node must contain a verb phrase (VP) which itself contains a modal or finite verb (MD, VBD, VBP, VBZ) to be considered a clause.

Given these definitions for clauses, we define a T-unit as any independent clause and all of its dependent clauses. That is, each T-unit contains exactly one independent clausal node and zero or more S/SINV/SQ/SBARQ subtrees immediately dominated by SBAR nodes. Among these, we define complex T-units as those which have one or more such dependent clause subtrees.

3.2 Readability

While understandable in an intuitive sense, readability of text can be expressed numerically in a large variety of ways depending on the context of the writing to be analyzed [5, 11, 14, 18, 23]. For the purpose of this study, we will consider the *Automated Readability Index* (ARI) and *Flesch reading ease* of the text.

Automated Readability Index

The ARI test is a fast method for calculating readability, optimized originally to be calculated as a typist input a passage on a typewriter [18]. For this reason, ARI uses character count as a proxy for syllable count in its formulation, given in Equation 1. The formulation used in this study is the revised ARI [11]. The value given by the revised ARI formula corresponds to an approximate US grade level, representing the number of years of formal education needed to understand the text. Thus, higher ARI values indicate more difficult text. For example, the abstract of this thesis has a revised ARI value of 14.8.

$$ARI = .4(\frac{words}{sentences}) + 6(\frac{characters}{words}) - 27.4 \tag{3.1}$$

The revised ARI serves as a baseline measure for readability due to its original use in print. This enables us to use it with little uncertainty compared with other measures that require more linguistically involved information, such as syllable count. The specific values of the formula were generated by Kincaid et al. to fit data from their corpus of students' writing and military documents to expected grade level; they have no standalone meaning.

Flesch Reading Ease

The Flesch Reading Ease test, given in Equation 2, is among the most well-studied and widely used readability formulas [11]. As the name suggests, higher reading ease corresponds to simpler text. The highest possible score is 121.22, indicating that all sentences in a passage contain only one-syllable words; a score of 60 is considered standard English [5]. As a specific example, the abstract of this thesis has a Flesch score of 29.5.

$$FRE = 206.835 - 1.015(\frac{words}{sentences}) - 84.6(\frac{syllables}{words}) \tag{3.2}$$

3.3 Logical complexity

Logical complexity in the context of natural language writing takes on a larger scope than the syntactic complexity measures described above. While a complete logical assertion can be contained within a single sentence, often in longer form writing, a logical assertion takes the form of several sentences taken together. For the purpose of this study, the overall length in the original text of each logical assertion is not considered; The other measures of complexity used in this study should suffice to describe complexity from that perspective. Instead, this study will examine the structure of the logical assertions based on their length in T-units. The introduction of logical complexity is one of the major contributions of this work.

We define logical complexity using similar metrics to syntactic complexity. For this approach, we define a *logical keyword* as an adverb or conjunction which fulfills the role of one of the three classical logical operations: *and*, *or*, and *not*. These are represented in the parse tree by adverb (RB) and coordinating conjunction (CC) nodes, from which we look for specific words that indicate any one of these three operations. We base this list on keyword lists used in previous work assessing the meaning of words in text [15, 22]. A full list of the words used for each operation can be found in Appendix B.1. The list of words is fairly small compared to lists of keywords for cognitive complexity because we limit the list to words that carry more direct logical structure indication.

We use the ratio of appearance of these keywords to T-units as the baseline of logical complexity. Because a T-unit can be understood as a single logical predicate, we examine the occurrence of logical operators within each one. Thus, higher numerical values will indicate higher complexity, as there is a higher concentration of logical indicator words.

As an example, the following passage would have a logical complexity of 1.5, with two *not* keywords and one *and* keyword across two T-units:

> “It might not seem that a constructed example would be useful, which is reasonable. Even a contrived example needs some context and meaning, not just the right words.”

Chapter 4 Data set

The data set used for this study is the US Senate Press Release corpus, containing transcriptions of public statements and reports made by US senators between 2004 and 2008. Rather than newspaper articles or other journalistic sources, the corpus contains press releases from the Senate offices, so there is no change in wording by third-party journalists [8].

4.1 Classifying press releases

The years spanned by the corpus entries include three election years for the US Senate: 2004, 2006, and 2008. However, entries from 2004 and 2008 are limited, and do not form a sufficient basis for study. Therefore, we focus on the 2006 election and its affected senators. However, the election cycle in the US is not a clearly defined time frame. The Federal Election Commission defines election cycles as lasting the full two years between Senate elections, but this pertains primarily to financing of campaigns. Seat race rankings start at the beginning of the election year in January, and we thus use the ten months from January through October as our definition of the election cycle as it pertains to press releases.

In the early stages of the study, we found that the variance among releases for different types of natural disasters was very significant and posed issues in identifying relevant articles. Therefore, we focus on coastal state senators and their press releases surrounding hurricanes as a representative of natural disasters. We make the assumption that these are sufficient to explore crisis response discourse, but we discuss this assumption at the end of this report.

4.2 Experimental groups

Given these two criteria, we define two experimental groups, which we will hereafter refer to as the *reelection group* and the *coastal group*.

The reelection group consists of the 33 senators who were up for reelection in 2006: Akaka, Allen, Bingaman, Burns, Byrd, Cantwell, Carper, Chafee, Clinton, Conrad, Dayton, DeWine, Ensign, Feinstein, Frist, Hatch, Hutchison, Jeffords, Kennedy, Kohl, Kyl, Lieberman, Lott, Lugar, Menendez, Ben Nelson, Bill Nelson, Santorum, Sarbanes, Snowe, Stabenow, Talent, and Thomas.

The coastal group consists of all senators in the corpus from the southeastern coastal states where hurricanes pose a significant threat: Burr, Chambliss, Cochran, Cornyn, DeMint, Graham, Hutchison, Isakson, Landrieu, Lott, Martinez, Bill Nelson, Sessions, Shelby, and Vitter.

Chapter 5 Experimental Design

For each of the complexity measures, we compare results from three categories of press releases from each senator, as described in the previous section: natural disaster-related, election cycle, and control (consisting of all other press releases). We examine the press releases of each senator separately to isolate the effects of these events on the complexity of the text. We hypothesize that press releases decrease in complexity during election cycles and in response to natural disasters. The following sections further detail how we test this hypothesis using the three complexity measures.

5.1 Syntactic and logical complexity

For each press release, we extract the text from a plain text file. This raw text is then cleaned of extraneous text, such as the publication location, that remains with it; this process is consistent among releases from the same Senator's office within the sample. These sanitized releases are then sent through the CoreNLP parser for parse tree construction, which handles some cases of missing punctuation and other issues as described in the appendix. We discard press releases which fail to process through the CoreNLP parser.

Once the parse tree has been constructed, the necessary information for syntactic complexity measures can be isolated as described in Section 3. We then isolate the key words for logical complexity from the same parse trees. These are counted and the ratio of their appearance to the number of T-units is recorded.

For all three of the measures examined in this study, higher values indicate more complex text. Therefore, we expect to see a significant decrease in these values in the *target text*. Similarly, we expect to see lower logical complexity in releases related to natural disasters, as with syntactic complexity, indicated by lower ratios of keywords to clauses.

5.2 Readability

We use the textacy library tools to extract the basic syntactic information needed to calculate readability scores [4].

As previously described, lower ARI and higher FRE values indicate simpler text. Thus, if our hypothesis is correct, we expect to see a decrease in ARI and increase in FRE in the coastal group's hurricane-related press releases.

Chapter 6 Results

For each senator in the election cycle and coastal states groups, we ran the experiment and obtained the mean and standard deviation of each of the six target measures: clauses per sentence (C/S), T-unit length (T-L), complex-T ratio (CT/T), revised automated readability index (ARI), Flesch reading ease (FRE), and logical complexity as defined in Chapter 3 (LC). Tables 6.1 through 6.6 contain those values, sorted by senator and by measure. We observe immediately a small number of outliers in each group, caused by press releases whose punctuation is missing. These outliers are separated and listed at the bottom of each table; they are excluded from further analysis, as their results are effectively noise due to their improper formatting.

To observe possible shifts in complexity, we chart each senator's mean value for baseline and experimental groups on a scatter plot. The graphs in figures 6.1-6.5 show this data within square bounds for clearer visualization. Each graph includes both the line $y = x$ (which we will call the *reference line*), which indicates no change between groups, and the line of best fit to provide context for the data. Greater distance from the line $y = x$ indicates a greater shift in complexity between the baseline and the experimental group (election cycle or hurricane-related press releases). Similarly, if there is a significant effect on complexity, the points will tend to fall on one side of the line. The line of best fit provides a second guide in this regard; if the line intersects $y = x$ within the data range, the data are distributed roughly evenly above and below the reference line. Furthermore, the line gives a good visualization of how far the data are from $y = x$ as a whole.

6.1 Syntactic complexity and readability results

To test our hypothesis that complexity decreases during election cycles and in response to natural disasters, we first look at the five established measures of linguistic complexity. Qualitative observation of the reelection group results suggests that there is no change in complexity during the months preceding an election, as the data are distributed roughly evenly above and below the reference line. We see further evidence of this when examining the line of best fit for each data set; all five are close in slope to the reference line and four of five intersect the reference line within the range of the data.

However, we see a significant difference when examining the results for the coastal senators' press releases. While there is not a significant shift in clauses per sentence, the four other measures show significant skew towards one side of the reference line. This effect is most pronounced in the two readability measures, for which all points lie on one side of the reference line. We see that for these senators, ARI values are higher and FRE values are lower for hurricane-related press releases, indicating higher complexity. This is mirrored in the syntactic complexity measures to a lesser degree, with notable increases in T-unit length across the board. Complex-T ratio is also increased for most of the coastal senator group. These corroborate the readability evidence that there is an increase in complexity for press releases concerning hurricanes.

6.2 Logical complexity results

We now examine the results of our logical complexity measure, as represented in Tables 6.5-6.6 and Figure 6.5. We see that in both the reelection group and coastal group, the graph shows similar results to the other five measures. There is little correlation and fairly even distribution with respect to the reference line for the reelection group, which indicates that election cycles have little effect overall on complexity. This effect is mirrored in the line of best fit, which also intersects the reference line within the data range.

For the coastal group, we see again that all of the points on the graph lie above the reference line, indicating an increase in complexity across the board. This is similar to the results of the other measures, indicating that language becomes more complex in press releases concerning hurricanes.

6.3 Discussion

As can be seen from Tables 6.1, 6.2, and 6.5, Senator Kennedy was excluded from the final results entirely, as too many of the press releases from the Kennedy office were unable to be parsed and we felt that discarding such a large portion of the data could significantly impact the results. We also feel that it is likely that Senator Kennedy would have been among the outlier cases, as Kennedy was among the small group of senators whose press releases do not have punctuation in the corpus. While Kennedy was the only senator whose data was excluded entirely, in the general case, as a result of discarding releases which could not be processed through the CoreNLP parser, there may be some skewing of the results if there is a particular pattern of complexity among those press releases.

Table 6.1: Syntactic complexity results for reelection senators

Name	**Clauses/sentence**		**T-Length**		**Complex-T ratio**	
	Base	Election	Base	Election	Base	Election
Akaka	1.677	1.748	29.866	28.507	0.459	0.481
Allen	1.416	1.651	33.129	30.96	0.458	0.516
Bingaman	1.353	1.348	35.319	39.523	0.468	0.465
Burns	1.518	1.424	27.571	26.935	0.426	0.431
Byrd	1.567	1.614	24.827	25.424	0.404	0.407
Cantwell	1.462	1.411	28.702	29.253	0.44	0.426
Carper	1.302	1.156	33.629	30.799	0.464	0.437
Chafee	1.376	1.417	32.758	30.123	0.448	0.418
Dayton	1.343	1.291	30.19	30.861	0.46	0.447
DeWine	1.291	1.329	26.73	29.604	0.335	0.38
Feinstein	1.408	1.466	30.141	29.393	0.444	0.452
Frist	1.352	1.392	29.336	31.227	0.4	0.426
Hatch	1.546	1.551	27.002	25.146	0.473	0.427
Hutchison	1.433	1.4	28.824	28.296	0.405	0.399
Jeffords	1.469	1.495	28.806	28.086	0.466	0.488
Kyl	1.569	1.546	31.819	28.962	0.493	0.526
Lott	1.368	1.53	27.077	26.438	0.369	0.402
Lugar	1.353	1.263	29.406	30.94	0.369	0.364
Santorum	1.321	1.534	36.548	31.892	0.443	0.452
Sarbanes	1.4	1.448	33.598	33.326	0.421	0.444
Snowe	1.454	1.458	33.357	31.721	0.486	0.493
Stabenow	1.349	1.395	33.186	29.603	0.436	0.457
Talent	1.384	1.389	27.667	28.455	0.385	0.405
Thomas	1.353	1.372	26.337	26.578	0.383	0.405
Clinton	3.738	3.954	76.573	79.458	0.755	0.764
Ensign	4.356	4.308	73.629	70.81	0.767	0.723
Kohl	2.771	2.837	73.205	69.996	0.757	0.764
Lieberman	3.894	3.345	74.614	72.252	0.759	0.739
Menendez	4.431	4.501	74.704	77.171	0.776	0.779
Nelson, Ben	2.782	3.032	61.012	62.605	0.685	0.694

Table 6.2: Readability results for reelection senators

Name	**ARI**		**Flesch**	
	Base	Election	Base	Election
Akaka	13.696	13.66	41.801	40.532
Allen	12.866	12.906	46.876	46.386
Bingaman	11.575	11.348	49.844	51.421
Burns	11.162	11.372	53.403	50.964
Byrd	10.239	10.587	56.452	54.213
Cantwell	12.555	13.188	47.83	45.608
Carper	12.035	10.642	48.44	52.345
Chafee	13.814	12.898	42.433	45.431
Dayton	12.099	12.134	48.002	47.088
DeWine	12.706	13.095	50.604	46.633
Feinstein	11.898	11.772	47.334	48.464
Frist	12.187	12.671	47.062	44.902
Hatch	10.933	10.384	53.329	56.077
Hutchison	12.112	12.369	48.35	46.471
Jeffords	11.574	11.353	50.318	50.473
Kyl	12.428	11.637	47.644	49.676
Lott	11.425	12.252	49.06	47.504
Lugar	13.845	13.638	41.029	42.161
Santorum	13.92	14.177	40.171	38.698
Sarbanes	13.699	13.501	44.294	45.031
Snowe	13.244	12.616	45.205	47.388
Stabenow	12.823	11.68	48.499	52.865
Talent	11.431	11.606	50.867	50.111
Thomas	10.012	10.346	57.008	55.031
Clinton	25.75	26.781	12.429	9.682
Ensign	24.441	23.868	15.782	17.222
Kohl	19.344	18.926	29.145	29.556
Lieberman	24.921	22.06	13.639	20.228
Menendez	26.415	26.143	12.482	12.29
Nelson, Ben	18.77	19.276	29.08	27.674

Table 6.3: Syntactic complexity results for coastal senators

Name	**Clauses/Sentence**		**T-unit length**		**Complex-T ratio**	
	Base	Hurricane	Base	Hurricane	Base	Hurricane
Burr	1.148	0.961	31.304	36.035	0.359	0.306
Chambliss	1.508	1.43	31.164	36.324	0.458	0.48
Cochran	1.256	1.397	29.828	29.908	0.358	0.375
Cornyn	1.471	1.467	27.751	27.588	0.413	0.427
DeMint	1.502	1.402	26.98	31.833	0.457	0.529
Hutchison	1.425	1.452	28.789	28.006	0.402	0.423
Isakson	1.472	1.478	29.615	30.968	0.444	0.425
Lott	1.347	1.5	26.479	27.642	0.359	0.406
Nelson, Bill	1.655	1.83	30.302	30.48	0.473	0.503
Vitter	1.673	1.637	27.51	27.834	0.449	0.439
Graham	2.152	2.349	63.399	60.259	0.71	0.632
Landrieu	4.286	4.582	72.749	76.401	0.789	0.811
Martinez	3.042	2.805	67.34	67.11	0.734	0.687
Sessions	3.301	2.857	55.532	49.588	0.636	0.614
Shelby	2.503	3.071	59.463	65.077	0.74	0.749

Table 6.4: Readability results for coastal senators

Name	**ARI**		**Flesch**	
	Base	Hurricane	Base	Hurricane
Burr	11.614	12.628	50.419	47.452
Chambliss	12.701	13.567	45.738	43.431
Cochran	12.367	13.593	45.107	41.941
Cornyn	12.185	12.569	48.005	44.995
DeMint	10.524	12.239	54.911	46.229
Hutchison	12.126	12.552	48.267	44.929
Isakson	11.694	12.122	49.193	45.312
Lott	10.961	12.673	51.611	44.034
Nelson, Bill	13.115	13.816	46.739	43.312
Vitter	12.731	13.352	46.934	43.265
Graham	15.953	18.118	38.587	34.109
Landrieu	25.614	27.643	14.127	6.522
Martinez	20.379	21.644	25.957	20.37
Sessions	19.822	15.955	28.373	34.733
Shelby	20.169	22.204	26.092	18.666

Table 6.5: Logical complexity results for reelection senators

Name	**Base**	**Election**
Akaka	1.195	1.089
Allen	1.236	1.266
Bingaman	1.169	1.208
Burns	1.049	1.03
Byrd	0.915	1.022
Cantwell	1.081	1.158
Carper	1.275	1.163
Chafee	1.233	1.042
Dayton	1.019	1.111
DeWine	1.104	1.21
Feinstein	1.07	1.048
Frist	1.138	1.065
Hatch	0.982	0.865
Hutchison	1.057	1.066
Jeffords	1.088	0.997
Kyl	1.179	1.086
Lott	0.974	1.111
Lugar	1.221	1.334
Santorum	1.256	1.191
Sarbanes	1.42	1.467
Snowe	1.277	1.239
Stabenow	1.279	1.109
Talent	0.919	0.992
Thomas	0.877	0.856
Clinton	3.631	3.828
Ensign	3.169	2.993
Kohl	2.934	2.912
Lieberman	3.662	3.153
Menendez	3.325	3.574
Nelson, Ben	2.551	2.441

Table 6.6: Logical complexity results for coastal senators

Name	**Base**	**Hurricane**
Burr	1.064	1.267
Chambliss	1.19	1.356
Cochran	1.035	1.231
Cornyn	1.176	1.197
DeMint	0.971	1.267
Hutchison	1.053	1.123
Isakson	1.086	1.104
Lott	0.956	1.087
Nelson, Bill	0.888	0.952
Vitter	1.022	1.08
Graham	2.715	2.643
Landrieu	3.395	3.597
Martinez	2.797	3.049
Sessions	2.18	2.613
Shelby	2.203	2.806

Figure 6.1: Syntactic complexity results for election group

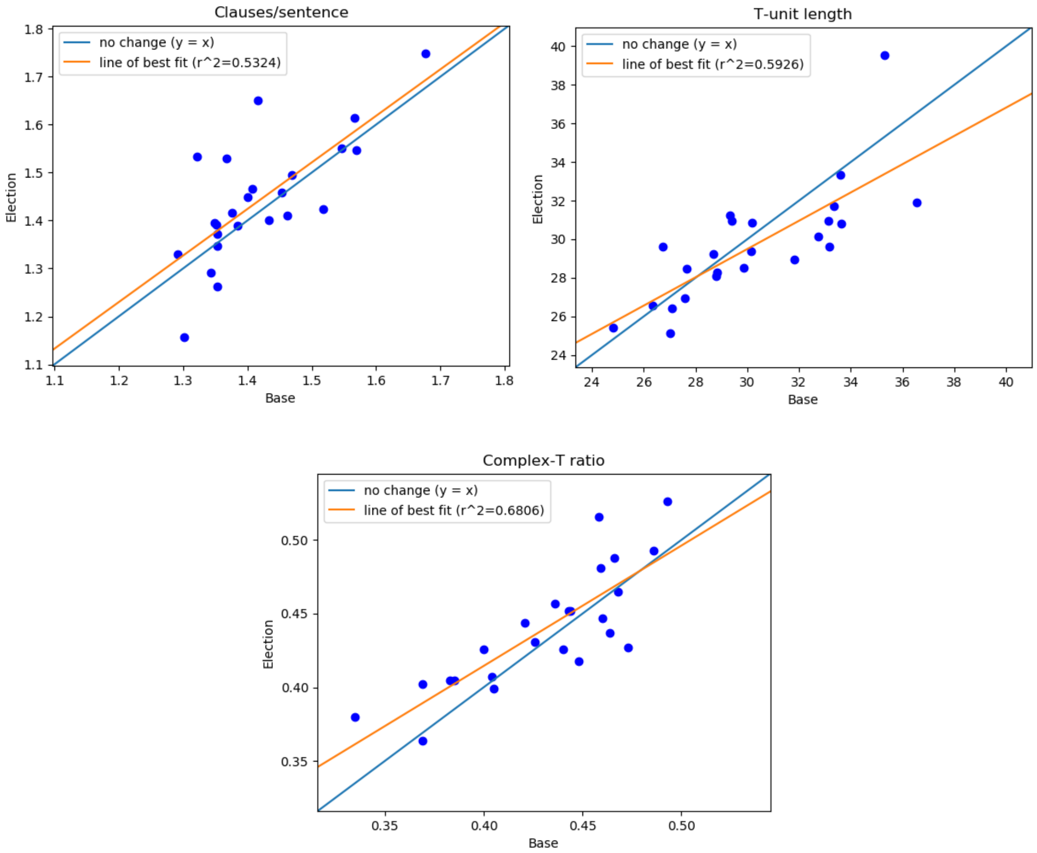


Figure 6.2: Readability results for reelection group

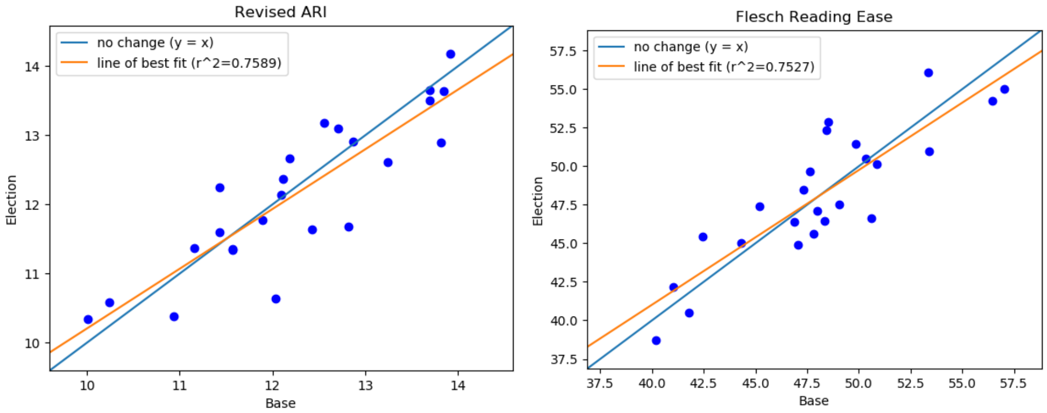


Figure 6.3: Syntactic complexity results for coastal group

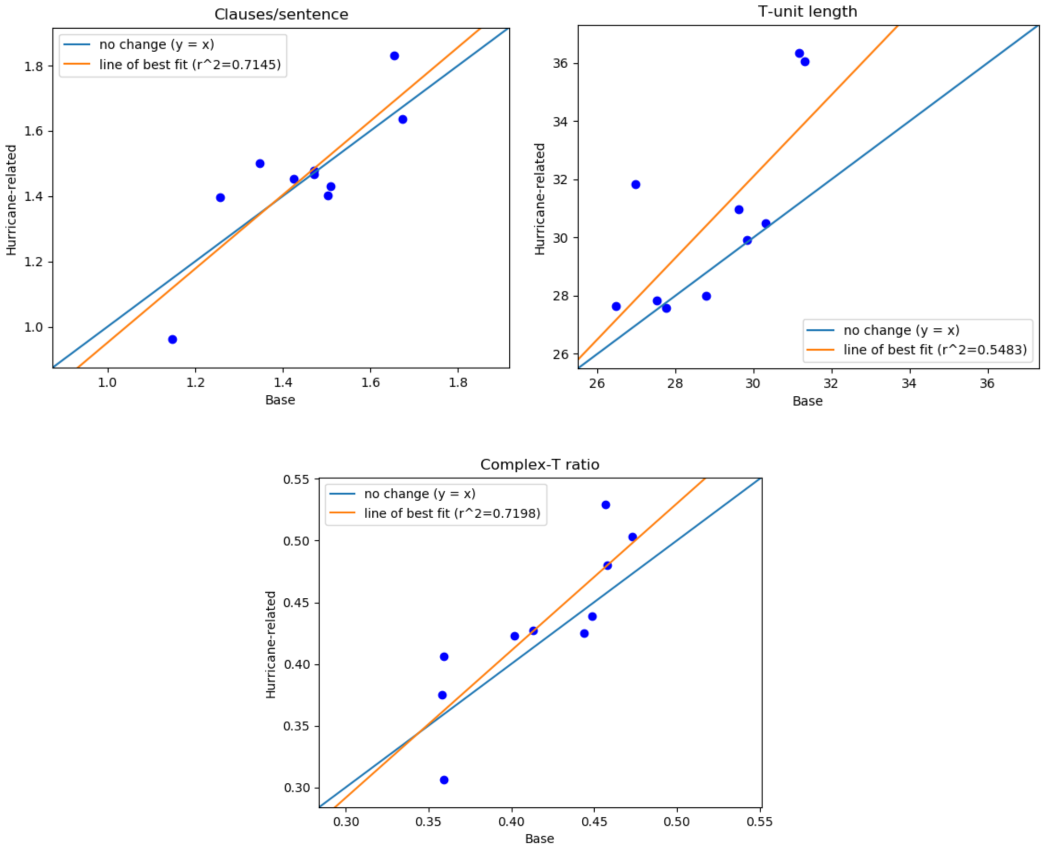


Figure 6.4: Readability results for coastal group

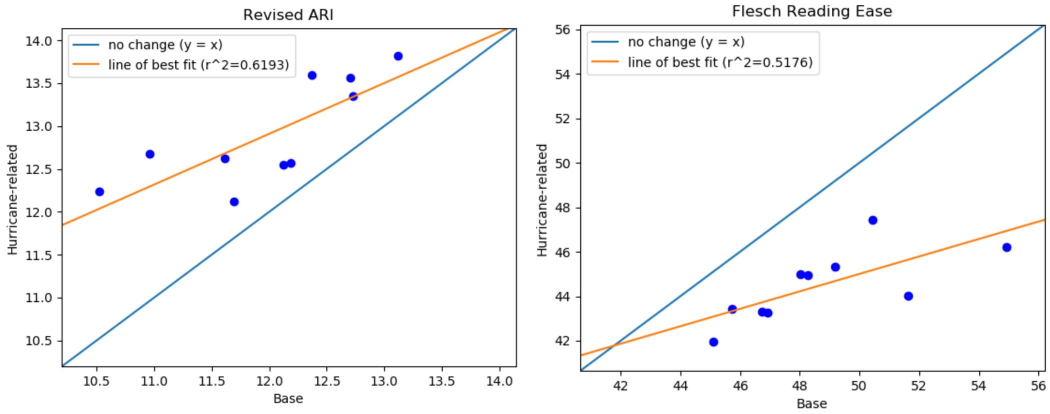
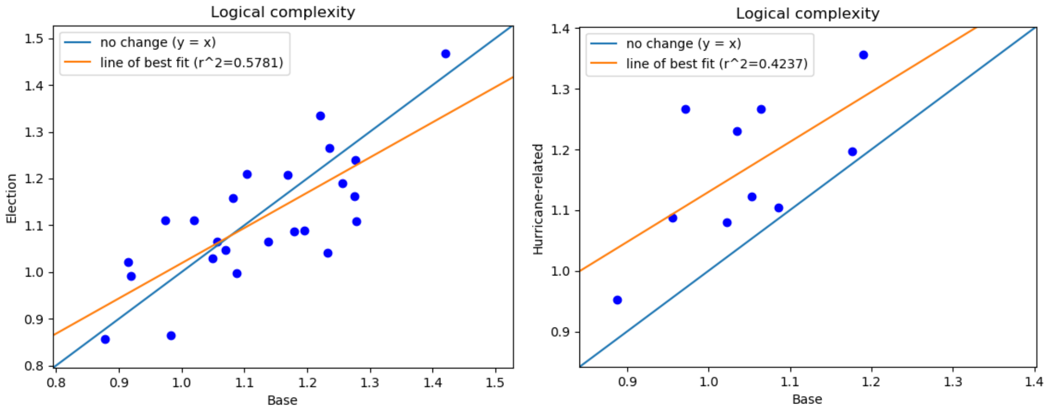


Figure 6.5: Logical complexity results for both groups



Chapter 7 Conclusions

In this study, we proposed a new measure of linguistic complexity, *logical complexity* to serve as an initial foray into applying classical Boolean logic to natural language. Our experiment tested it in comparison to several well-known measures of linguistic complexity which examine text at different angles. We also applied all of the measures to test the hypothesis that politicians' language becomes more simple when discussing natural disasters, and to a lesser extent during election cycles.

Our results suggest that our initial hypothesis was incorrect regarding natural disasters; in these cases, we found that politicians' language becomes more difficult to read and slightly more complex. This effect is shown clearly in the readability results, for which we observe an increase in ARI and decrease in Flesch Reading Ease for all senators. The cause of this is unclear; however, the lack of significant change in clauses per sentence and complex-T ratio suggest that the main contributing factor is raw increase in word count for these press releases. We hypothesize that this may be caused by sentences whose length is extended by lists, such as names of locations or people affected by a disaster. Future work can be done to examine the frequency and effect of long lists on complexity by using similar tools to this study. Lists appear as noun phrases with several nouns, and can thus be isolated for further review.

A potential limitation of this study in this area, however, is the choice to use hurricanes in East coast states as a representative for natural disaster-related press releases. We made this assumption on the expectation that response to such disasters can be generalized to natural disasters as a whole. However, should this not hold, choosing to study a different type of natural disaster, or disasters as a whole, may yield a different result.

For election cycles, our results found no significant change in any of our complexity measures. While we found that there is significant change in individual senators' language, we did not find a consistent trend within the sample. Because this corpus contains press releases but no direct campaign speeches or advertising material, we come to the conclusion that any changes in language would be isolated to those media, rather than affecting the senators' communication with the public at large.

Of some note is that the complexity of this corpus as a whole is itself notably higher than that of standard English. While there are few standard values for syntactic complexity (and obviously none for logical complexity) for standard English, we can compare readability scores to standard values. We recall from 3.2 that ARI values represent the approximate US grade level education needed to understand a text, so even the lowest average ARI in the study of 10.012 (belonging to Thomas's baseline press releases) is significantly higher than the standard English level of 7-8. The same segment of the corpus scores 57.008 in Flesch Reading Ease, which is still more difficult than the standard English 60-70. The range (excluding outliers) extends to Senator Santorum, whose baseline text scores are 13.92 ARI and 40.171 FRE, much more difficult than standard English. It is possible that the magnitude of change identified by this study in hurricane-related press releases is sufficiently small to not significantly impact a typical voter's ability to read the text. However, examining this aspect of the corpus is outside the scope of this study.

Our experiment showed that our logical complexity measure gave similar results to the well-established complexity measures. In particular, the results were similar to those of the readability measures, which is fairly surprising; our definition of logical complexity uses structures of syntactic complexity, T-units, rather than the more general structures employed by readability measures. These results imply that our measure does reasonably well in determining the complexity of text. We assert then that while this study serves primarily as a proof of concept, the results suggest that Boolean logic can serve as a useful basis for practical examination of language complexity.

In summary, we examined the linguistic complexity of Senate press releases between 2004 and 2008 to find if the language used changes in complexity during election cycles or in response to natural disasters, for which we use hurricanes as a representative. We found no change in press releases during election cycles, but saw complexity rise consistently in hurricane-related press releases. The major contribution of this work is our proof-of-concept proposal of logical complexity, which we found to have similar results to well-established measures of linguistic complexity. While this study does not directly apply the model described by Parikh et al., our results indicate that a more detailed application of this Boolean model is likely to be useful in measuring complexity of political discourse.

7.1 Future Work

This work serves as a proof of concept, and thus naturally further work can be done to refine the definition of logical complexity for computational linguistics. Because this study focused on an approach similar to that of cognitive complexity, some of the structural information that could inform more precise Boolean predicates was lost in processing. Furthermore, the list of keywords used in this study did not include key phrases such as *in addition* that may also represent logical operators. A more detailed approach that more carefully examines the placement of key logical operators would serve as a more complete application of these concepts.

Chapter A Appendices

A.1 Examples from the corpus

Sample parse trees

We must meet the problems of today, and prepare for the problems of tomorrow.

```
(ROOT
  (S
    (NP (PRP We))
    (VP
      (MD must)
      (VP
        (VP
          (VB meet)
          (NP
            (NP (DT the) (NNS problems))
            (PP (IN of) (NP (NN today)))))
        (, ,)
        (CC and)
        (VP
          (VB prepare)
          (PP
            (IN for)
            (NP
              (NP (DT the) (NNS problems))
              (PP (IN of) (NP (NN tomorrow))))))))
    (. .)))
```

One independent clause, zero dependent clauses; one non-complex T-unit
ARI = 4.771; FRE = 71.768

This government is here to serve the people, so we need to start putting things in plain-language around here.

```
(ROOT
  (S
    (S
      (NP (DT This) (NN government))
      (VP
        (VBZ is)
        (ADVP (RB here))
        (S (VP (TO to) (VP (VB serve) (NP (DT the) (NNS people)))))))
```

```
    (, ,)
    (CC so)
    (S
      (NP (PRP we))
      (VP
        (VBP need)
        (S
          (VP
            (TO to)
            (VP
              (VB start)
              (S
                (VP
                  (VBG putting)
                  (NP (NNS things))
                  (PP
                    (IN in)
                    (NP
                      (NP (JJ plain-language))
                      (PP (IN around) (NP (RB here))))))))))))
  (. .)))
```

Two independent clauses; two non-complex T-units

ARI = 7.3; FRE = 85.015

Irregular/problem cases

Punctuation absent

> Dakota communities are establishing venture capital funds to invest in new business creation following the recommendation of a Marketplace for Entrepreneurs business panel In its report last year our Marketplace entrepreneurial task force concluded that North Dakota could grow more businesses and create more jobs if there were new sources of venture capital the seeds of new business development And today we are announcing that three separate venture capital funds are setting up in North Dakota Senator Conrad said This is a remarkable response to our Marketplace recommendation Year after year Marketplace shows its value by helping grow North Dakota s economy Business and community leaders in Bismarck Grand Forks and in Fargo are establishing venture capital funds for the purpose of investing in local businesses The Bismarck fund called Northern Plains Investment LLP is to begin fund raising this week organizers hope to raise at least 750 000 All are part of

```
a wider venture capital fund family based in Minnesota
Last year Senator Conrad s Summit on Entrepreneurship
included a task force of business leaders who completed
a report for Senator Conrad with a menu of
recommendations to encourage the growth of small and mid
sized business in North Dakota Among the top
recommendations was the creation of North Dakota based
venture capital funds that would bring much needed
private investment to new North Dakota business ventures
Bismarck business leader Niles Hushka of Kadrmas Lee
Jackson joined Senator Conrad at his announcement today
Bruce Gjovig Director of UND s Center for Innovation who
is coordinating the creation of the funds in Grand Forks
and Fargo also spoke at today s press conference
Providing venture capital to new businesses can make a
real difference in our communities For entrepreneurs on
the East and West coasts finding capital is much less of
a challenge than it is here in North Dakota Senator
Conrad said I m pleased to see our North Dakota
communities following the recommendation of our
Entrepreneurial Task Force and create their own venture
capital funds
```

Irregular article headers

Rather than an example passage from the corpus, we include this entry in the appendix to highlight that many senators' press releases did not have regular markers for the title, date, and publishing location of the press release. In most cases, these press releases contained only the title of each article at the start of the text. However, because we cannot reliably determine where those titles end automatically, we were unable to completely clean the text for these press releases.

A.2 Logical complexity keywords

And	additionally, also, and, but, furthermore, including, plus, with
Or	either, else, or, other, otherwise
Not	excluding, n't[1], not, without

---

[1] For the purpose of tokenization, contractions whose expanded form contains the word *not* have the token *n't* isolated in the parse tree.

Bibliography

[1] Scott F Beers and William E Nagy. Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing*, 22(2):185–200, 2009.

[2] Ryan C Black, Ryan J Owens, Justin Wedeking, and Patrick C Wohlfarth. *US Supreme Court opinions and their audiences*. Cambridge University Press, 2016.

[3] Walter Dean and Rohit Parikh. The logic of campaigning. In *Indian Conference on Logic and its Applications*, pages 38–49. Springer, 2011.

[4] Burton DeWilde. Textacy: NLP, before and after spaCy. `https://github.com/chartbeat-labs/textacy`, 2019.

[5] Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221, 1948.

[6] Stephen J Gaies. T-Unit Analysis in Second Language Research: Applications, Problems and Limitations. *TESOL Quarterly*, pages 53–60, 1980.

[7] Judy Goldsmith, 2018. Personal communication.

[8] Justin Grimmer. Replication data for: Representational Style: The Central Role of Communication in Representation, 2010.

[9] William Kodom Gyasi. Readability and political discourse: An analysis of press releases of Ghanaian political parties. *Journal of Media and Communication Studies*, 9(6):42–50, 2017.

[10] Kellogg W Hunt. Grammatical Structures Written at Three Grade Levels. NCTE Research Report No. 3. 1965.

[11] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) For Navy Enlisted Personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

[12] Xiaofei Lu. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496, 2010.

[13] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[14] G Harry Mc Laughlin. SMOG Grading-a New Readability Formula. *Journal of Reading*, 12(8):639–646, 1969.

[15] Ryan J Owens and Justin P Wedeking. Justices and Legal Clarity: Analyzing the Complexity of U.S. Supreme Court Opinions. *Law & Society Review*, 45(4):1027–1061, December 2011.

[16] Rohit Parikh and Çağıl Taşdemir. The strategy of campaigning. In *Fields of Logic and Computation II*, pages 253–260. Springer, 2015.

[17] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[18] RJ Senter and Edgar A Smith. Automated Readability Index. Technical report, Cincinnati University, OH, 1967.

[19] Arthur I Siegal and Estelle Siegal. Flesch readability analysis of the major pre-election speeches of Eisenhower and Stevenson. *Journal of Applied Psychology*, 37(2):105, 1953.

[20] Cory Siler, Luke Harold Miles, and Judy Goldsmith. The Complexity of Campaigning. In *International Conference on Algorithmic Decision Theory*, pages 153–165. Springer, 2017.

[21] Susana M Sotillo. Discourse Functions and Syntactic Complexity in Synchronous and Asynchronous Communication. *Language Learning & Technology*, 4(1):77–110, 2000.

[22] Yla R Tausczik and James W Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[23] Wilson L Taylor. “Cloze Procedure”: A New Tool for Measuring Readability. *Journalism Bulletin*, 30(4):415–433, 1953.

Vita

Austin Taing

Education

- Bachelor of Science in Computer Science, University of Kentucky, 2016

Academic Employment

- Undergraduate Teaching Assistant, University of Kentucky, January 2015 - May 2016
- Graduate Teaching Assistant, University of Kentucky, January 2017 - May 2017