Measuring and Explaining Political Sophistication Through Textual Complexity*

Kenneth Benoit† Kevin Munger‡ Arthur Spirling§

August 26, 2017

Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application to political text suffers from a number of theoretical and practical issues. We develop a new benchmark of textual complexity which is better suited to the task of determining political sophistication. We use the crowd to perform tens of thousands of pairwise comparisons of snippets of State of the Union Addresses, scale these results into an underlying measure of reading ease, and "learn" which features of the texts are most associated with higher levels of sophistication, including linguistic markers, parts of speech, and a baseline of word frequency relative to 210 years of the Google book corpus *n*-gram dataset. Our refitting of the readability model not only shows which features are appropriate to the political domain and how, but also provides a measure easily applied and rescaled to political texts in a way that facilitates comparison with reference to a meaningful baseline. We apply the model to various political corpora to demonstrate how substantive conclusions differ when using our improved approach.

`sophistication` software available, on request: `github.com/kbenoit/sophistication`

*Prepared for presentation at the 2017 Annual Meeting of the American Political Science Association, San Francisco, 31 August - 3 September 2017.

†Professor of Quantitative Social Research Methods, London School of Economics (kbenoit@lse.ac.uk)

‡PhD Candidate, Department of Politics, New York University (km2713@nyu.edu)

§Associate Professor of Politics and Data Science, New York University (arthur.spirling@nyu.edu

1 Introduction

A key question in the field political communication concerns how the nature of political communication has changed. At the same time that the challenges of governing have grown in complexity, the sophistication of political speech, by many measures, appears to have declined. Outside of academic study, typically as part of a broader discussion concerning the "dumbing down" (Gatto, 2002) of political communication, observers have applied measures of textual complexity from educational fields to find that the sophistication of political language has steadily declined over the past 200 years. For example, in 2013, *The Guardian* newspaper[1] used the Flesch-Kincaid grade-level estimates to document a decline in the textual complexity of US Presidential State of Union Addresses.[2]

Social scientists have used measures of textual complexity to link linguistic sophistication to outcomes, often identifying concrete benefits to clarity. Jansen (2011), for instance, studies the reading level of communications by four central banks, equating lower reading levels of bank communication with greater clarity, which they link to positive effects on the volatility of returns of financial markets. Likewise, Owens and Wedeking (2011) and Spriggs (1996) examine the complexity of Supreme Court decisions, pointing to the importance of clarity in court opinions, although their approaches do not rely on measures of reading level. In the context of the British parliament, Spirling (2016) applied readability measures to document the democratizing effects of franchise reform on elite speeches. Finally, as a meta-analysis to defend against charges of elitism and jargon (e.g. Diamond, 2002; Kristof, 2014), Cann, Goelzhauser and Johnson (2014) show that while the reading ease of articles in the top political science journals has declined since 1910, the typical political science article requires less reading ability than the typical article in *Time Magazine* or *Reader's Digest*.

These applications share one trait: They equate important substantive characteristics of political, economic, or legal communication such as clarity or sophistication with indexes such as the

[1] http://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level

[2] Although see Benoit, Munger and Spirling (Forthcoming) for a data-driven critique of this claim.

Flesch Reading Ease score (Flesch, 1948) . Yet such measures were developed decades earlier in entirely different contexts, namely educational research and applied psychology. As a consequence, we are uncertain as to the true direction of change for specifically political communication. More importantly perhaps, we are also unclear about what any such change actually represents in terms of underlying dynamics of language. For example, if communication is indeed becoming simpler, it might be a 'good' thing insofar as politics is becoming clearer; or it might be a 'bad' thing insofar as it represents dumbing down; or it might be both, or neither.

To remedy this, here we critically review the properties of current measures of textual complexity, and develop a new measure of the sophistication of political language. Our approach uses experimental data based on human pairwise comparisons of short extracts of political speech, which we then use to scale linguistic sophistication using a simple but well-defined statistical model. In particular, we employ a Bradley and Terry (1952) approach in which clarity of a text is treated as "ability." By moving measurement to a model-based approach, with all the statistical mechanics that brings, we allow for sensible statements about uncertainty and probabilistic inference: thus, one can make claims about the probability that a given text is easier or harder than another. This is impossible with all extant techniques of which we are aware. For convenience, and to be consistent with previous efforts, we also provide a continuous version of our measure on the (0,100) interval. Our preferred model is more general than others in the sense that it considers the association of a large collection of features on the difficulty of political texts (rather than just one or two somewhat arbitrarily chosen variables). This includes a comprehensive measure of rarity, extracted from the Google books corpus. Precisely because it is trained on a relevant domain, this technique yields a measure of textual complexity that is by construction more appropriate for political text than classical measures and with a model fit that is at least marginally better than more traditional alternatives.[3] Furthermore, we can be precise about each feature's relative contribution to political complexity. More generally, our methodological contribution is to provide a 'work-flow' for scholars interested in measuring textual complexity for any substantive area.

[3] Although for reasons we explain, this is a tricky comparison to make.

To demonstrate how this new measure allows us to gain new insight on old problems, we compare it to the Flesch Reading Ease in two related but different applications of elite discourse. In the first—the State of the Union addresses since the founding of the Republic—we show that our measure has considerably more variation that FRE and, if anything, the mean ease of understanding of language has become higher over time than suggested with the more traditional approach. However, once we introduce uncertainty bounds via a text-based bootstrap, general claims about 'dumbing down' are much more dubious. Second, we apply our approach in its continuous form to three million speeches from the UK's *Hansard* House of Commons records for the period 1935–2013. We provide (admittedly speculative) evidence that technology changes in broadcasting the speeches to the public may have induced politicians to alter the nature of their utterances. Importantly, however, the substantive conclusions one draws from our approach differ in interesting ways from those drawn from current techniques.

2 The Challenges of Measuring Linguistic Sophistication

Measuring linguistic complexity is not a new endeavor (see Klare, 1963, for an overview), with early work dating at least to the 19th Century (e.g. Sherman, 1893). The context is typically education, in the sense that the task is matching learning materials to students, based on their age and cognitive ability. Thus, the interest is in the fast estimation of the "readability" of a document. While there are a large number of indices for this task—indeed, Michalke (2015) references and implements no less than 27 of them—this variety conceals two facts. First, the measures are actually very similar to one another in principle and in practice. And second, a few of the methods completely dominate applied work in terms of use and citation.[4] To see this latter point, in Table 1 we list eight commonly seen metrics—some of which have been adjusted over the years and republished in very similar forms—and their Google Scholar citations at the time of our writing. Inevitably, the number of citations understates the actual use of the methods in practi-

[4]We ignore metrics for languages other than English here, though there certainly exists a literature dealing with them (e.g. Fucks, 1955; Yuka, Yoshihiko and Hisao, 1988)

Table 1: Overview of commonly used 'reading ease' measures in order of citation via Google scholar at the time of writing.

Author	Name of Method	Year	Citations
Flesch	Flesch Reading Ease	1948/49	3793
McLaughlin	SMOG	1969	1402
Dale and Chall	Dale-Chall	1948	1389
Gunning	Gunning Fog Index	1952	1232
Kincaid et al	Flesch-Kincaid Grade Level	1975	1093
Fry	Fry Graph	1968	1007
Spache	Spache Formula	1953	355
Coleman and Liau	Coleman-Liau	1975	261

cal scenarios, but readers can nonetheless see that the various Flesch-based measures (including the Flesch-Kincaid measure) garner the lion's share of attention, with SMOG and the Dale-Chall easure somewhat behind. Readers will also note that while scholars have continued to be interested in the problem of studying readability after 1975 (e.g. Anderson, 1983), these measures were generally not designed or validated in the modern period.

In terms of technical details, for a given document, the available measures take into account some combinations of:

- (average) sentence length; (e.g. Flesch, 1948, 1949; Gunning, 1952; Fry, 1968; Kincaid et al., 1975)
- the (average) number of syllables per word (e.g. Flesch, 1948, 1949; Gunning, 1952; Wheeler and Smith, 1954; Fry, 1968; Kincaid et al., 1975);
- the parts of speech represented in the document (e.g. Coleman and Liau, 1975); and
- the familiarity of the terms used (e.g. Dale and Chall, 1948; Spache, 1953).

To get a sense of what it means to 'take into account' these characteristics, consider the original work of Flesch (1948) (later updated by Kincaid et al. 1975). Flesch studied the reading comprehension of school children. In particular, he was interested in the average grade level of students who could correctly answer at least 75% of some multiple choice questions regarding some texts.

Running a regression with a constant and two predictors (average sentence length and average number of syllables per word), ultimately yielded the following formula for scoring texts:

$$206.835 - 1.015\left(\frac{\text{total number of words}}{\text{total number of sentences}}\right) - 84.6\left(\frac{\text{total number of syllables}}{\text{total number of words}}\right). \qquad (1)$$

In the original application, this index of "Flesch Reading Ease" typically took values between zero and 100.[5] Subsequently, Kincaid et al. (1975) introduced a mechanical conversion of the formula that yields values roughly equivalent to the US grade school level required to understand a text.

Clearly, other than via syllable information, the Flesch formula does not explicitly take into account the actual familiarity of the words used in a text. An example of an approach that does is Dale-Chall (Dale and Chall, 1948), the formula for which has been adjusted over time but for exposition may be rendered as

$$0.1579\,(\text{percentage of difficult words}) + 0.0496\left(\frac{\text{total number of words}}{\text{total number of sentences}}\right). \qquad (2)$$

This yields an (average) grade level at which a reader could be expected to comprehend the document in question. Here, the 'percentage of difficult words' refers to any terms not a pre-ordained list of 763 (subsequently around 3000) 'familiar words' in English. We will discuss this idea in more detail below, but for now note that these terms are ones that 80% of fourth grade children (in 1948) would know.

While social scientists have not ignored the measurement of readability *per se* (e.g. Cann, Goelzhauser and Johnson, 2014), there has not been especially great interest in using such methods to produce independent or dependent variables for analysis. Nonetheless, there is some work in this area: Lim (2008), for example, considers the changing nature of presidential rhetoric in the United States since the founding of the Republic. Jansen (2011) studies central bank communications

[5] In practice, the statistic is bounded at an upper "ease" limit of 121.22 for texts consisting of one-syllable, one-word sentences, and bounded from below only by an offset of the average word length.

and Owens and Wedeking (2011) look at the complexity of Supreme Court decisions. Finally, Spirling (2016) studies the democratizing effects of franchise reform on elite speeches in the British parliament.

Outside of academic study, typically as part of a broader discussion concerning 'dumbing down' (Gatto, 2002), observers of politics have made use of these metrics. For example, in 2013, *The Guardian* newspaper[6] used Flesch-Kincaid grade level estimates to assess the complexity of Presidential State of Union Addresses, while in April 2016 *Politico*[7] claimed that "Donald Trump Talks Like a Third-Grader" using similar methods.

2.1 The 'Out-of-Domain' Prediction Problem

Regardless of the specific mechanical details behind current techniques, they were not designed, optimized or tested on the types of social science data to which they are being applied. Indeed, when political scientists score documents using these methods, they are essentially calculating 'out-of-domain' (and obviously 'out-of-sample') *predictions*.[8] While we are not the first to note this issue (see, e.g., Loughran and McDonald, 2014, on issues specific to financial reporting), it is helpful to explicate the series of problems it induces.

First, the approaches were designed to match texts to the formal education level of potential readers. They were never intended for the more general task of measuring the 'sophistication' of texts in a given domain such as politics, where abstract conceptual appeals to 'democracy' or 'liberty' might make documents significantly more difficult to follow over and above their sentence structure or average number of syllables. Second, and closely related, the indices were originally for assessing children, rather than adult citizens. Yet this second group will differ not simply in their education level from younger people, but also in their knowledge and understanding of the

[6] http://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level

[7] http://www.politico.com/magazine/story/2015/08/donald-trump-talks-like-a-third-grader-121340

[8] Technically, the term 'out-of-sample' could also be used alone here, but we opt for the stronger 'out-of-domain' to draw attention to the fact that the concern is not simply that the estimates are applied to children in the 1940s or 1950s who happened not to be in the original study via random sampling: they are applied to completely different subjects in completely different contexts.

political process, since presumably they will be exposed to affairs of state on a more regular basis. Third, as the citation dates make clear, these indices were mostly created in the 1940s and 1950s, subsequent to which we can well imagine that human understanding has evolved as language and linguistic style develops.[9]

Fourth, while the measures are certainly simple—typically consisting of two or three easily calculable text features multiplied by constants—the objective functions they embody are poorly defined when applied to new data. To see this, consider the FRE. This is derived from a linear regression where, as usual for such approaches, the minimization problem (ordinary least squares) is well-defined. In the original context it would yield an R^2 variance explained statistic. However, when taken to State of the Union speeches, it is difficult to know whether the measure—i.e. the model—is performing well or not. That is, the scores of the documents represent out-of-sample predictions, yet there is no readily available metric for assessing the quality of those predictions. An immediate consequence of this issue is that, fifth, it is hard to compare measures (models for the data, essentially) and contend that one is systematically better than another in a given context. Put very simply, if measure A has document i as more difficult than document j, yet measure B implies the opposite, it is not clear which should be preferred, nor on what criteria the predictions ought to be judged. Crudely, once out-of-domain, there is no 'ground truth' with which to compare.

Precisely because all scores are out-of-domain, a sixth problem emerges: there is no 'natural' way to interpret fine-grained differences in document scores. Consider, for example documents i and j which score as 70 and 75 respectively on the FRE. In principle, one could claim that were the original sample of children given the speeches, a particular proportion would understand questions relating to the texts in a way that gives rise to the scores. This is a strange counterfactual since, of course, all the texts may have been written after the original study took place. But in any case, the interpretation is extremely awkward. The researcher would like to know the *probability* of understanding one speech over another, or their relative appeal were they in a head-to-head contest for a reader of a given comprehension level. But such information is not forthcoming.

[9] As a trivial example to fix ideas, '`computer`' may have been difficult to understand in 1956, but much less so in 2015.

The scores are hard to interpret for a related, seventh reason: there are typically no uncertainty estimates around these out-of-sample, out-of-domain, predictions. That is, if document *i* is scored similarly to document *j* in terms of point estimates, we would surely be more confident in such a measurement for *i* if it was 3000 words long relative to *j* at 30 words.

2.2 Other Problems

As was made clear above, many of the commonly used approaches, such as Dale-Chall and FRE, as essentially *composite indices*. That is, they combine information from a text's characteristics to produce document level estimates. It would be helpful if the researcher could interpret the coefficients on the subparts of the index directly, but that is difficult under current conditions. Thus, although we may agree that the FRE for State of the Union addresses has fallen over time, what we would really like to know is *why*: as a function of sentence length, or word length, or both? FRE or its close allies cannot answer this directly.

As we discussed, there are obvious reasons to take into account the familiarity of the language used when calculating a document score. For a modern reader, "Indeed, the shoemaker was frightened" would presumably be easier to understand than "Forsooth, the cordwainer was afeared", yet both would be scored identically by FRE. When such matters are taken into account by current approaches however, it is in *ad hoc* way: thus the Dale-Chall method provides a list of 3000 familiar words, with any word outside this set having a constant weight, regardless of its actual commonality. Also, as is obvious, such a list is not updated as language changes: thus, within the Dale-Chall words we find '`telephone`' which would have been meaningful in 1948 but not 1848. Conversely, '`locomotive`' makes the list, but would probably not be a term familiar to a reader in the United States now. By the same token '`television`' does not occur, though of course it would be common for contemporary students.

2.3 Qualities of a Better Approach

Some of the problems we discuss are straightforward to solve. For example, a better approach will study adults in obviously political settings for the contemporary period. This will immediately rectify the central 'out-of-domain' issue. Other matters are more subtle. Ultimately, as in the educational literature, humans are the 'gold standard' for coding complexity of language. With that broad understanding in mind, an ideal way forward is to either use small numbers of experts or, better yet, large numbers of non-experts who can code texts in a fast, reproducible mode via 'crowd-sourcing' (Benoit et al., 2016).

Because our interest is more general than education, we want the coders to score the documents directly. At least since the work of Thurstone (1927), we know that having humans perform (large numbers of) pairwise comparisons between texts is likely preferable to other hand-coding systems. In the pairwise case, political scientists (e.g. Loewen, Rubenson and Spirling, 2012; Lowe and Benoit, 2013) have used the Bradley-Terry model (Bradley and Terry, 1952) as a fast and well-grounded way of converting the pairwise binary decisions over items (here, documents) and placing them into continuous score space. This simple approach has a natural interpretation, insofar as its fundamental building block is the probability that 'i beats j'—here, the probability that i is easier to understand than j—when the two documents are compared one to another. This probability is well-defined, and has the usual desirable properties, such as being strictly between zero and one. Furthermore, since it is a maximum likelihood approach, the Bradley-Terry method allows for uncertainty estimates around the scores themselves and the usual confidence and prediction interval machinery. Finally, because the latent characteristic of the item can be modeled via a linear predictor—that is, $\mathbf{X}\beta$—one can talk meaningfully about the 'effects' of certain characteristics, such as document length, syllable number, the familiarity of tokens etc on the linguistic complexity of a document. Notice that such estimates will be *sample specific* and once some domain coding has been undertaken, one is not required to simply apply a rote formula again and again however dubious a given application might be.

3 Methods: Crowdsourcing Complexity

With the above considerations in mind, we aim to "discover" the textual features that constitute complexity. We employ human coders in the crowd to perform hundreds of pairwise comparisons to evaluate which of a pair of snippets is more complex. We drew these snippets from the 70 SOTU Addresses delivered since 1950. This corpus is appropriate because the purpose of the SOTU has remained relatively constant in the postwar period. Traditional readability scores like FRE have also been applied to SOTUs in the popular press[10], and changes in this measure have substantive implications for the nature of accountability and democracy in the United States.

[KEN – Which is it? I'm pretty sure we only used the modern era.]

3.1 Preparing the Snippets: Gold Standards and Crowdflower

We begin with the raw texts of the 70 SOTU Addresses delivered since 1950. Until about 1913, written copies of these Addresses were delivered to the Congress, while afterwards they were presented directly to the public via radio and later television. Each form of address brings some organizational non-sentence pieces of text, which we remove. We broke the Addresses up into one- and two-sentence snippets of text.[11]

These snippets vary dramatically in the number words they contain. This is clearly an important component of textual complexity—so important, in fact, that we take it as given and measure the variation in complexity among snippets of similar lengths. We match snippets of approximately equal length, to avoid comparisons where deciding on the "easier" snippet appears easy because one is noticeably shorter than the other. Along the same lines, we take the FRE scores as the baseline against which our measure should be compared. Within each group of snippets of similar lengths, we sort the snippets once by FRE scores in ascending order and again in descending order,

[10] Guardian, Feb 12, 2013., http://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level.

[11] We disqualified some snippets from consideration: any outside of the 0-121 range of the Flesch reading ease index; any containing more than two years as tokens; any with large numbers (usually referencing specific financial outlays); and any beginning with the title of a section in the document.

and combine these two lists to create a list of comparisons that vary from extremely dissimilar to extremely similar FRE scores.[12]

Similar to Benoit et al. (2016), we then use the Crowdflower platform to recruit coders to evaluate which of each snippet is more complex. We labeled the task as "Identify Which Of Two Text Segments Contains Easier Language." Upon accepting the task, we provide the workers with a number of example comparisons, with one option correctly labeled as more complex. The specific instructions provided to each worker were:

> Your task is to read two short passages of text, and to judge which you think would be easier for a native English speaker to read and understand. An easier text is one that takes a reader less time to comprehend fully, requires less re-reading, and can be more easily understood by someone with a lower level of education and language ability.

A crucial aspect of crowdsourcing any coding task is ensuring that workers provide high quality responses. To that end, we employ "gold standard" tasks: interspersed among the comparisons of interest are comparisons in which one snippet is unambiguously more complex than the other, at a rate of one test question in ten. To create the gold standard questions, we select the 15% of the snippet pairs with the largest disparity in FRE scores, verified through inspection. If the worker incorrectly classified more than 30% number of these gold standard tasks, she was removed from the pool of workers and her answers removed from the dataset.[13] Prior to being accepted for the task, a crowd worker also had to pass a qualification test consistently entirely of test questions, answering at least 7 of 10 correctly.

To create the snippets, we formed two-sentence segments from the State of the Union corpus, with three levels of ranges of the total number of characters: between 345–360, 360–375, and 375–390 in length, from which we randomly selected 2000 pairs of snippets for direct comparison, in a way that guaranteed the connectivity of pairs for comparison to enable Bradley-Terry scaling.[14]

[12]To be precise, we matched three sets of two-sentence snippet pairs: those with lengths between 345-360, 360-375, and 375-390 respectively. We also created an additional 210 randomly selected bridging pairs, to form a fully linked network of pairs to enable pairwise scaling.

[13]Following Berinsky, Margolis and Sances (2014), we also included some "screener" questions, which appear to be the same as normal comparisons but include at some point the phrase "Disregard the content and code this sentence as EASIER." Approximately 10% of the test questions were screeners, and approximately 10% of the total comparison tasks were test questions.

[14]To increase the range of data and to use results from an earlier coding of snippets, we also combined the post-1950

Finally, we added another 15% of gold questions plus 5% of special gold "screener" questions. After removing duplicates, our dataset of snippets to be compared consisted of 7,236 total pairings for comparison, including 836 "gold" questions, of which 310 were screeners. We crowd-sourced the comparisons using a minimum of three coders per pair, yielding 19,810 total comparisons, of which 13,430 did not involve screeners or test questions. To aid the automation of this process and to provide both reproducibility and transparency, we implemented all of the functions to sample snippets, create pairs and test questions, prepare the data for Crowdflower, and to process the crowd-coded data in an R package sophistication, which also includes the cleaned version of the SOTU corpus. [KEN–above two paragraphs are contradictory, implying either 1% or 5% gold screeners.]

3.2 Incorporating Familiarity: Google n-grams and parts of speech

Corpus linguistics has progressed significantly since the early measures of reading ease were developed, giving us access to a huge amount of detail about word rarity and how it evolves over time. Our test data spans political speech dating to the 1790s, and a major contribution of our measure is that it incorporates a benchmark of how unusual (and hence how difficult to understand) each word from that time span is in contemporary usage. To this end, we downloaded the unigram frequency datasets from the Google Book corpus,[15] which yields token counts on a yearly basis from 1505 until 2008. Processing this enormous data and discarding any years prior to 1790 resulted in a total set of 615,362,456,717 token counts from 85,623 word types, after filtering out tokens that occurred fewer than five times or that did not match a dictionary of 133,000 English words and word forms.[16] To smooth out individual differences in the yearly samples, we combined the frequency counts by decade.

texts with some with one- and two-sentence snippets from an earlier set of crowd coding. This earlier set used a range of 180–300 characters and 180–400 characters respectively, but our dataset included just nine unique snippets, used in 99 different comparisons with post-1950 snippets, and in all of the 36 pairwise comparisons against one another.

[15] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

[16] One reason for the massive drop in the number of word types is that many appear to be artifacts of errors introduced in optical character recognition. (You should have paid more attention to typing those CAPTCHAs correctly.)

To assess the benchmark of a how unusual was a text, we computed the frequency of each term in a decade relative to the frequency of the word "the", the most common word in the English language and also one whose relative frequency has remained relatively unchanged in several hundred years. This allowed us to compare the relative frequencies of terms without being affected by changes in overall word quantities or transcription accuracies (which vary significantly over the time sampled). For instance the word "husbandry" (the cultivation and breeding of crops and animals) was used 8.5 times more frequently in the 1790s than it was in the 2000s. [ARTHUR–Do we actually use the Google corpus for any year except 2000?] Because we think this would make it harder for a contemporary audience (such as our crowd coders), we computed the score of each text based on the frequencies from the decade of the 2000s. Our computation looked up every token in a text from the Google unigram frequency corpus in 2000, and computed a mean relative word frequency as well as the relative frequency of the *least* frequent word.

For example, considering the following two snippets:

> Numerous are the providential blessings which demand our grateful acknowledgments... too important to escape recollection. (George Washington, 1791)
>
> Now, we have to build a fence. And it's got to be a beauty. (Donald Trump, 2015)

These are 15 and 14 tokens in length, but the mean frequency relative to *the* in the 2000s for the first was 0.11, and 0.14 for the second, indicating that the mean word in Washington's speech was relatively much less frequently used than in Trump's. The word that is used least commonly (relative to *the*) in the two snippets induces a large difference in the measurements of the texts: for Washington, it is *providential* which has a ratio of 0.00002085 relative to *the* (implying *the* is used about $48,000$ times as often). For Trump, the relevant word is *fence*, for which the ratio is an order of magnitude higher, at 0.00025 (meaning *the* is used about 4000 times as often). (We note also that the Flesch Reading Ease for the Washington text is 5.5, compared to 105.1 for the Trump snippet.)

We also computed the relative frequency of parts of speech in each text, tagging the snippets using the Google Universal tagset[17] using the spacyr package built on the spaCy NLP library for

[17] See https://github.com/slavpetrov/universal-pos-tags.

Python.[18] This follows some readability indexes, such as Tränkle and Bailer (1984), that consider conjunctions and prepositions, and Coleman's "C3" and "C4" indexes (Coleman and Liau, 1975) that take into account the frequency of pronouns and propostions. Converting these to relative frequencies for each snippet gave us information on the proportions of nouns, adjectives, verbs, prepositions, and so on. This allowed us to include these quantities in the feature set for fitting models below to predict reading ease. [ARTHUR–did we also include syntactic dependency measures?]

3.3 Bradley-Terry Regression Analysis

Exposition of the Bradley-Terry model (Bradley and Terry, 1952) can be found in numerous textbooks (e.g. McCullagh and Nelder, 1989), but we follow the presentation found in Turner and Firth (2012) for our work here. The input data is the result of our human coders having declared winners in the large number of "contests" between snippets. For a given contest, crowd workers must decide which of two snippets i and j for is 'easier' to comprehend (no ties are allowed). If the 'easiness' of i is α_i, and the 'easiness' of j is α_j, then the odds that snippet i is deemed easier than j may be written as $\frac{\alpha_i}{\alpha_j}$.

Defining $\lambda_i = \log \alpha_i$, the regression model can be rewritten in logit form:

$$\text{logit}[\Pr(i \text{ easier than } j)] = \lambda_i - \lambda_j. \tag{3}$$

Subject to specifying a particular snippet as a "reference snippet" (whose easiness is set to zero), this set-up allows for maximum likelihood estimation of each snippet's easiness. For current purposes though, we wish to make the easiness of the snippets a product of covariates—that is, the average length of words they contain, the number of syllables the words have etc. This is achieved

[18] See http://spacy.io.

by modeling the easiness of a given snippet as

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir}. \tag{4}$$

This is known as the *structured* Bradley-Terry model: the set of β coefficients then tells us the marginal effect of each x-variable on the perceived (relative) easiness of the snippets. Notice further that, on estimating the β parameters, the covariates pertaining to a given document may be used to obtain the (predicted) easiness of that text (even if it did not appear in sample, or not in that given form).

This is a simple model, and it is worth emphasizing what is being assumed about the data generating process when we interpret its relevant output. First, we assume that the outcomes of the contests are (statistically) independent of one another: that what happens in the kth contest does not affect what happens in the $k+1$th contest. Second, we are making no allowance for variability between snippets which have otherwise identical covariate values. That is, we are not using any kind of random effects for the snippets themselves. This means, equivalently, that the contest results for a given snippet are not modeled as correlated. Third, we make no attempt to include so-called "contest-specific predictors" either in their indirect form—such as effects for (the proclivities of) given human coders—or directly—such as allowing for consequences of the order in which the snippets were presented to the subjects who judged them.

To be clear, the model is sufficiently flexible to allow all these concerns to be addressed—we have simply chosen not to do so. Our primary interest is in estimating the complexity of documents by predicting (that is, scaling up) from the snippet results. This means we care mostly about the point estimates of the coefficients, not their attendant uncertainty. Since the former is unaffected by our simplification here, that is how we proceed.

3.3.1 Variable Selection via Machine Learning

As noted above, it is not *a priori* obvious which variables should be included in a given model of readability: in this case, the interest is in which features should appear in the linear predictor. To attack this problem, we first fit an *unstructured* Bradley-Terry model, which returns an estimate of an "ability" λ_i (in this case, relative easiness) for each snippet, but makes no use of covariates.[19] We then use all our various text characteristics as features to predict these (unstructured) abilities. Specifically, we use a random forests approach (Breiman, 2001), and then we inspect the (relative) 'variable importance' estimates for each covariate. Once those characteristics that matter most are identified, they can be used in the structured model of Equation (4) to obtain the relevant coefficient estimates. We return to the results of that process momentarily.

3.4 A "Bag-of-Snippets" Approach

Finally, as is perhaps obvious by now, ours is a "bag-of-snippets" approach, and is analogous to the "bag-of-words" commonly used in this literature. We assume that the order in which the snippets appear in a document has no effect on how difficult or easy they are to comprehend. Clearly, in some cases this abandoning of context will be an oversimplification. As an example, consider the following two sentences said by one character to another in Tokien's *The Hobbit*:

> There is more in you of good than you know, child of the kindly West. Some courage and some wisdom, blended in measure.

It is not hard to imagine that the second sentence is easier to understand once one has read the first, because the subject of the sentence is implied rather than explicitly stated. In contrast, consider the following two sentence snippet from Joyce's *Finnegans Wake*:

> Begin to forget it. It will remember itself from every sides, with all gestures in each our word.

[19] In practice, it is occasionally the case in our sample that a snippet never wins or never loses. The usual consequence of this kind of data separation would be infinite ability estimates. In one run of the model, we simply deleted those missing values, and in another we used the bias-reduction technique of Firth (1993) to ameliorate this problem. The results, in terms of the variable importance order are essentially identical, either way.

Here, the first sentence is arguably harder to understand once the second sentence is presented: it is then unclear what *it* might refer to. How often such context issues arise routinely in political texts is an open question, but we ignore them here (or rather, we assume that, on average, their effects are zero).

4 Results

We have two main sets of results: in the first part, we compare the standard measures as applied to our political texts. Secondly, and more importantly, we provide a new measure of complexity based on our crowdsourced data and the inferences we draw from our machine learning approach.

4.1 Comparing the Standard Measures

Our setup allows us to compare the performance of the various standard measures in a systematic way. In Table 2 we consider two natural ways to do this. For each of the traditional measures, we fit a Bradley-Terry model which has one predictor: the score for the snippets on a given measure. Thus, the first row refers to a model in which the only covariate is the (difference in the) snippets' Flesch scores (a model we return to below), the second row refers to a model in which the only covariate is the (difference in the) snippets' Dale-Chall scores, and so on. We report the Akaike information criterion for each of these models, along with the proportion of contests correctly predicted by the model. This latter statistic is calculated by generating the relevant λ_is from the linear predictor, using the $\hat{\beta}$ from the model, multiplied by xs for a given snippet. We then calculate the probability that the snippet which actually won a contest would be expected to do so given the estimated parameters—in the sense of Equation 3. If this probability is greater than 0.5, then we declare that a success for the model.

One observations is immediate: the models all perform very similarly, with very little to separate them in terms of either AIC or accuracy. The best performer on our data was the Spache measure, but the FRE is almost exactly as useful and might be preferred on familiarity grounds.

Table 2: Model performance of the standard measures. The overall fit of the Bradley-Terry model using the scores for a given measure is reported in two ways: the Akaike information criterion (AIC) and the Proportion of contest results correctly predicted (where a correctly predicted contest is one in which there is > 0.5 probability that the actual winner would win).

	AIC	Proportion Correct
FRE	26269.2	0.568
Dale-Chall	26227.9	0.573
FOG	26084.8	0.573
SMOG	26188.2	0.526
Spache	26025.6	0.577
Coleman-Liau	26574.4	0.550

We use it in our running comparison for what follows.

4.2 Augmented Bradley Terry Approach

In Appendix A we report the variable importance plots for the random forest models that we ran on the unstructured abilities. In particular, the model favors the rarity measure based on the recording the least commonly occurring term in the snippet (relative to 'the' in the Google corpus)—denoted as `google_min_2000`. And it also suggests average sentence length measured in characters (`meanSentenceChars`) is about as important. Given our discussion above, the fact that these variables are useful is unsurprising. In principle, of course, we could stop there (especially given the relatively large distance of the 'top two' from the other variables). In experiments, however, we found that the third most important variable, `pr_noun`—the proportion of words from the text that are nouns—helped model fit. We thus include that one too to form a 'basic' machine learning model.

How does this simple model perform? To assess that, we construct a 'baseline' model which uses the Flesch reading ease (FRE) as its (only) covariate content. We do this in two ways. First, we include the FRE of the snippet using the weights from Flesch's (1948) original formula. Second, we include the variables Flesch (1948) includes, but allow the model to calculate the optimal weights for our political data. In Table 3 we report the findings from those models, in the leftmost two columns. For the 'FRE baseline' model (original weights) we see that the Akaike information

criterion (AIC) is 26269, while the proportion (of contests in the data) correctly predicted (PCP) is 0.568. When we allow the weights on the relevant variables to adjust to local conditions (column 2) we see a commensurately better model fit: the AIC falls to 25912.69, and the proportion correctly predicted rises to 0.583. This is in line with our thinking above: in particular, that models work best when fit to relevant data. Column 3 represents our 'basic' three variable model as discussed above. Clearly, it does better than the Flesch model with the original weights, but—perhaps surprisingly—not as well as the re-weighted version (AIC is higher, PCP is lower).

Studying the model, we note that it doesn't include a measure of word length—yet the success of FRE tells us that almost certainly matters. Looking down the variable importance plots, the first measure of word length to be recommended (i.e. the one 'highest up' in importance terms) is the average number of characters per word (`MeanWordChars`). As an experiment, we added this variable our machine learning model and re-ran the analysis. The results of that process are in the fourth column of Table 3 titled 'Best Model'. Clearly, it now outperforms every other version, with the lowest AIC (25739.93) and the highest PCP (0.587). In an effort to ascertain the robustness of this model, we dropped the parts-of-speech variable (`pr_noun`) and added the next highest rated one (`pr_verb`), but in both cases the fit got worse. This is our preferred model for the analysis that follows. Note, in passing, that all the variable 'effects' are expected (and are statistically significant at conventional levels): in particular, *ceteris paribus* texts that contain words which have low (minimum) rarities are easier to understand ('Minimum Google books rarity' is positive), texts that contain longer sentences ('mean Sentence Chars') are harder, and texts with longer words ('mean Word Chars') are also more difficult to comprehend. More nouns ('noun proportion'), on average, also adds to simplicity and this is, in fact, in keeping with earlier work by Flesch (1948) who proposed a 'human interest' index in which a text with more (pro)nouns was generally found to be more compelling than one with fewer.

On what types of data, exactly, does our model do better? Unsurprisingly, given they share core terms, it's when two documents are similar other than the proportion of nouns they contain, or the rarity of their words. And, on inspection—i.e. looking at the contests for which our model

outperforms the Flesch version to the greatest extent—it's the frequency term that matters. To get a sense of this, compare these two snippets. The first is from Obama's 2009 address, and has an FRE of around 50:

> I speak to you not just as a President, but as a father, when I say that responsibility for our children's education must begin at home.

The second is from Cleveland's 1889 effort, which has an FRE of approximately 67:

> The first cession was made by the State of New York, and the largest, which in area exceeded all the others, by the State of Virginia.

Thus the FRE model predicts this to be a relatively straightforward win for Cleveland's speech. Our model, of course, penalizes the estimate of its simplicity due to the presence of the relatively rare term *cession* (along with there being slightly fewer nouns in the second document). Indeed, the frequency of the least common term in Obama's speech is over three orders of magnitude larger than that of Cleveland's speech. Put crudely, if researchers think the commonality of terms matters for measuring complexity, our approach is preferred.

It is helpful to be candid about several issues pertaining to our results. First, clearly, while we are outperforming the most widely-used measure of readability, our gains are not huge in an absolute sense. The largest gains in predictive accuracy come from refitting the Flesch model appropriately to the data rather than using its usual "off-the-shelf" mode. Second, these gains are large in a *relative* sense. Our task was intentionally designed to be difficult. The baseline Flesch predictive accuracy was 56.8%–a mere 6.8% better than chance. Our final model is 8.7% better than chance, a relative increase of 28%. Third, whether or not one uses our *specification*, the general *approach*—of training on relevant data and providing model-based estimates—is surely correct, precisely for the reasons we gave above. Indeed, even if one wanted simply to use the Flesch set up (in terms of its component variables) based on Table 3 we would recommend 'local' data for that purpose.

Table 3: Model comparison, post feature-selection. Note that the last column represents our 'optimal' model. 'PCP' is proportion (of contests) correctly predicted by the model.

	FRE Baseline	FRE re-weight	Basic RF model	Best Model
FRE	0.02*			
	(0.00)			
mean Sentence Length		−0.06*		
		(0.00)		
mean Word Syllables		−1.78*		
		(0.07)		
Minimum Google books rarity			1310.41*	1332.49*
			(153.27)	(155.85)
mean Sentence Chars			−0.01*	−0.01*
			(0.00)	(0.00)
noun proportion			0.61*	0.63*
			(0.19)	(0.19)
mean Word Chars				−0.31*
				(0.02)
N	19430	19430	19430	19430
AIC	26269.20	25912.69	25917.49	25739.93
PCP	0.568	0.583	0.580	0.587

Standard errors in parentheses

* indicates significance at $p < 0.05$

5 Applications: snippets, *State of the Union* and *Hansard*

We can apply the results of our model in various ways. We outline three obvious approaches before demonstrating how they might be used in practice. First, given Equation (3) and Equation (4), we can obtain a (point) estimate of the probability that any given text i is more difficult than any other text j by calculating

$$\Pr(i \text{ easier than } j) = \frac{\exp(\lambda_i)}{\exp(\lambda_i) + \exp(\lambda_j)}. \tag{5}$$

To see how this works, consider two snippets, one from Eisenhower,

> Here in the District of Columbia, serious attention should be given to the proposal to develop and authorize, through legislation, a system to provide an effective voice in local self-government. While consideration of this proceeds, I recommend an immediate increase of two in the number of District Commissioners to broaden representation

Table 4: Examples of covariates from two snippets in the data.

snippet	Min Google rarity	Mean Sent Chars	noun proportion	mean Word Chars
Eisenhower	3.501e-07	158.5	0.23	5.37
Bush	1.40e-08	153.5	0.31	4.72

of all elements of our local population.

and one from George W. Bush

> And the victory of freedom in Iraq will strengthen a new ally in the war on terror, inspire democratic reformers from Damascus to Tehran, bring more hope and progress to a troubled region, and thereby lift a terrible threat from the lives of our children and grandchildren. We will succeed because the Iraqi people value their own liberty - as they showed the world last Sunday.

For each of these snippets, Table 4 gives the relevant covariate values for our 'best' model above.

Using the coefficients from Table 3, it is a simple matter of matrix multiplication to form

$$\lambda_{\text{Eisenhower}} = (1332.49 \times 3.501e-07) + (-0.01 \times 158.5) + (0.63 \times 0.23) + (-0.31 \times 5.37) = -3.10$$

and

$$\lambda_{\text{Bush}} = (1332.49 \times 1.40e-08) + (-0.01 \times 153.5) + (0.63 \times 0.31) + (-0.31 \times 4.72) = -2.80.$$

Following the algebra above, we have

$$\Pr(\text{Eisenhower snippet easier than Bush snippet}) = \frac{\exp(-3.10)}{\exp(-3.10) + \exp(-2.80)} = 0.425.$$

Of course, these comparisons can be made between *any* two documents—so long as we have covariate values for them—including fifth grade texts, as in Flesch's (1948) original work. In our case, we obtained a set of fifth grade texts from a university education department[20], and estimated the relevant λ as above to be -2.175897. Thus, the probability that the Eisenhower text is easier

[20] https://projects.ncsu.edu/project/lancet/fifth.htm

than a fifth grade text is estimated to be 0.284; and the probability that the Bush text is easier to follow than the fifth grade works is 0.324. We can place confidence intervals around the point prediction by resampling the sentences in the texts (in the sense of Lowe and Benoit, 2013).

Along with model-based estimates, researchers may also want a quantity analogous to the continuous 0–100 scores from the Flesch (1948) (regression) formula. There are at least two ways to obtain this. First, using Equation (5) denote the $\Pr(i \text{ easier than } j)$ term as p. Then, supposing that we have an appropriate example of a (set of) fifth grade text(s), we can substitute $\exp(\lambda_i)$ for 100 (or, indeed, any number preferred) and then rescale $\exp(\lambda_j)$ as $100 \times (\frac{1}{p} - 1)$. Though this preserves the model-based interpretation of the quantity of interest, in practice it tends to return quite low numbers once one is even slightly removed from a fifth grade text. For example, a spotcheck on a document with an FRE of around 84 implies a rescaled score of 35, which 'seems' very low. Again, this is not 'wrong'—it is simply rescaling in a way that preserves the probability structure inherent in the model. But it may well be confusing for end-users, who expect a number approximately commensurate with the grade-level interpretation given by Flesch.

With this problem in mind, an alternative is to simply rescale all the λs (that is, the $\mathbf{X}\beta$s, without applying the exponential function) themselves to be on the relevant interval. For a given data set, a sensible way to proceed is to include a text(s) at the fifth grade level, and one at the post-college level (or whatever minimum and maximum is preferred), and to then scale all resulting λs from 100-0, based on those two end points.[21]

5.1 The Original Snippet Dataset

Experimenting with the continuous measure on the corpus we have performs well in the sense that it returns point estimates on a 0–100 scale that are commensurate (but not identical) to the FRE equivalents. This 'works' because it replaces a logit-style calculation that is not linear in the $\mathbf{X}s$ with a linear sum (i.e $\sum_{r=1}^{p} \beta_r x_{ir}$), exactly like the regression-based formula for FRE. In Figure 1

[21] We used the collection of fifth grade texts we mentioned above for the 'easy' end of the scale, and the most difficult snippet (which had an FRE of around 3) for the 'hard' end.

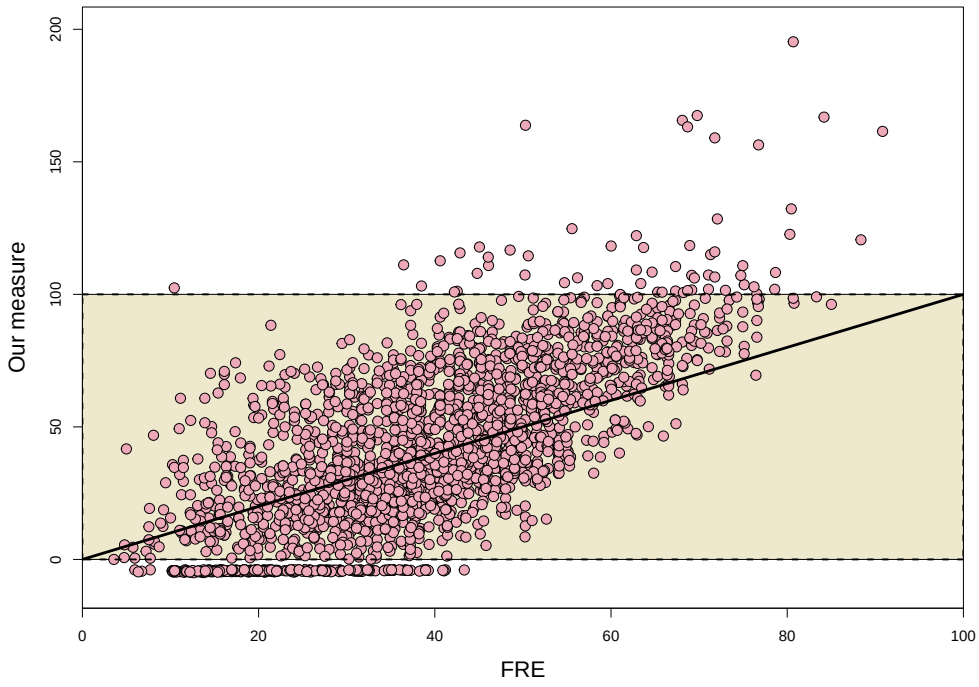

Figure 1: Comparing the "linear" version of our measure to FRE of the snippets. Correlation is generally high, especially for the theoretical range of the FRE (inner box).

we provide a scatterplot of our measure for the snippets (y-axis) relative to the FRE for the same data (x-axis). Clearly the correlation (~ 0.7) is reasonably large and positive. The internal box allows for a more direct comparison of our measure to the (theoretical) minimum and maximum of the FRE: in general, our measure performs similarly. This implies that for the great majority of documents for which FRE is used, our measure—preferred on theoretical grounds—is a good choice that will behave 'as expected'. Outside the box, particularly to the 'top right' of the plot, our measure tends to score the points differently. Indeed, we assign a considerably higher ('easier') rating for the simplest texts.

5.2 Reanalyzing the *State of the Union* addresses

Recall that our snippets came from the *State of the Union* time-series, a dataset of considerable interest to academics and journalists. Using our model-based probability measure—here, with a fifth grade text as a baseline for comparison—Figure 2 reports the relevant point estimates and 95% (simulated) confidence intervals (y-axis) plotted against the date of the relevant text. The proba-

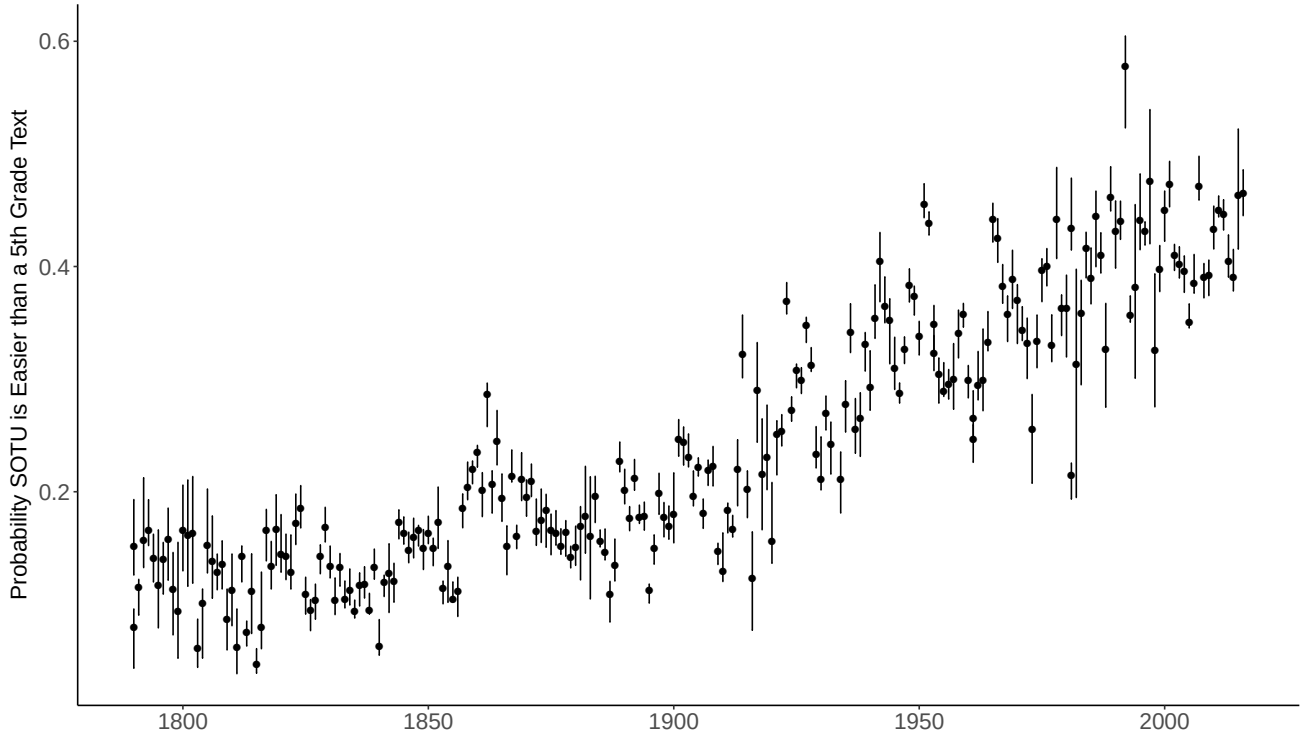

Figure 2: The probability that a State of the Union address is easier to understand than a fifth grade text baseline.

bility estimates are drifting upwards over time, but generally stay below 0.50. But because we are using a well-defined statistical model, we can say more about the data. In particular, the confidence intervals allow us to make comments about sampling uncertainty. Note that there is considerable overlap between the intervals for the post-war period (for example, some of the speeches in the early 2000s are not so different to those in the early 1950s). This implies that statements about 'dumbing down'—or simplification—may be correct in some aggregate sense if we consider the entire period since the founding of the Republic, but less clear for modern times specifically.

Of course, since other measures in the literature are not based directly on a statistical model, it is hard to compare our output here with more traditional approaches. Fortunately, the continuous version of our measure does allow a direct comparison, and in 3 (where we label it 'MBE' for [m]odel [b]ased [e]stimate(s)) we show it plotted against the FRE (which has been smoothed and given a 95% confidence band calculate by sentence-level bootstrap). Clearly, the conclusions from the measures agree in terms of general direction: addresses become easier over time. But

conclusions differ in terms of magnitude. In particular, our measure has the speeches prior to around 1910 being considerably more difficult to understand than FRE claims they were. And then, post 1910, our measure tends to have the estimated ease of understanding the passages as higher than FRE. To the extent that one believes that new technology, such as the radio and the television, lead to speeches that are easier to follow after the first decade of the 20th Century, this makes sense. And, to reiterate, our model is actually trained on appropriate, political data. Why do we estimate the earlier speeches as being so much more difficult than FRE has them? Mostly, this is because of our rarity variable. Recall that it uses the relative commonality of a word in 2000 as a baseline. Of course, as one moves back into history words that are rare and archaic today become more common. Thus, our measure allows us to more accurately judge how difficult texts are from the *perspective of a modern reader*. Notice that if this is undesirable, e.g. for those wanting difficulty estimated for contemporaneous audiences for 1800, 1810, 1820 etc, our framework allows one to do that. It would simply require using the relevant Google books corpus for the decade in which the text originated: that is, this rarity would become a dynamic variable in the modeling set-up, rather than fixed to its levels in 2000.

5.3 *Hansard*, 1935–2013

As our final application, and to demonstrate the different types of conclusions one might reach using our measure versus FRE, we analyzed 78 years of House of Commons debates. This *Hansard* corpus includes essentially all speeches (some 3 million in number) by all Members of Parliament (MPs) for the period under study. The data is described in Rheault et al. (2016). To keep our analysis simple, we focus solely on Labour and Conservative MPs, who represent around 90% of all MPs in the corpus, and use our continuous measure as described above. The data is compiled in 'sessions' of parliamentary time, which last around a year a piece. For each of these sessions, we estimated the mean of the FRE and our continuous measure, for all MPs. The results of those calculations can be seen in Figure 4. Clearly, the lines differ: though they start in approximately the same place, our measure (denoted MBE) implies the speeches quickly become easier, before

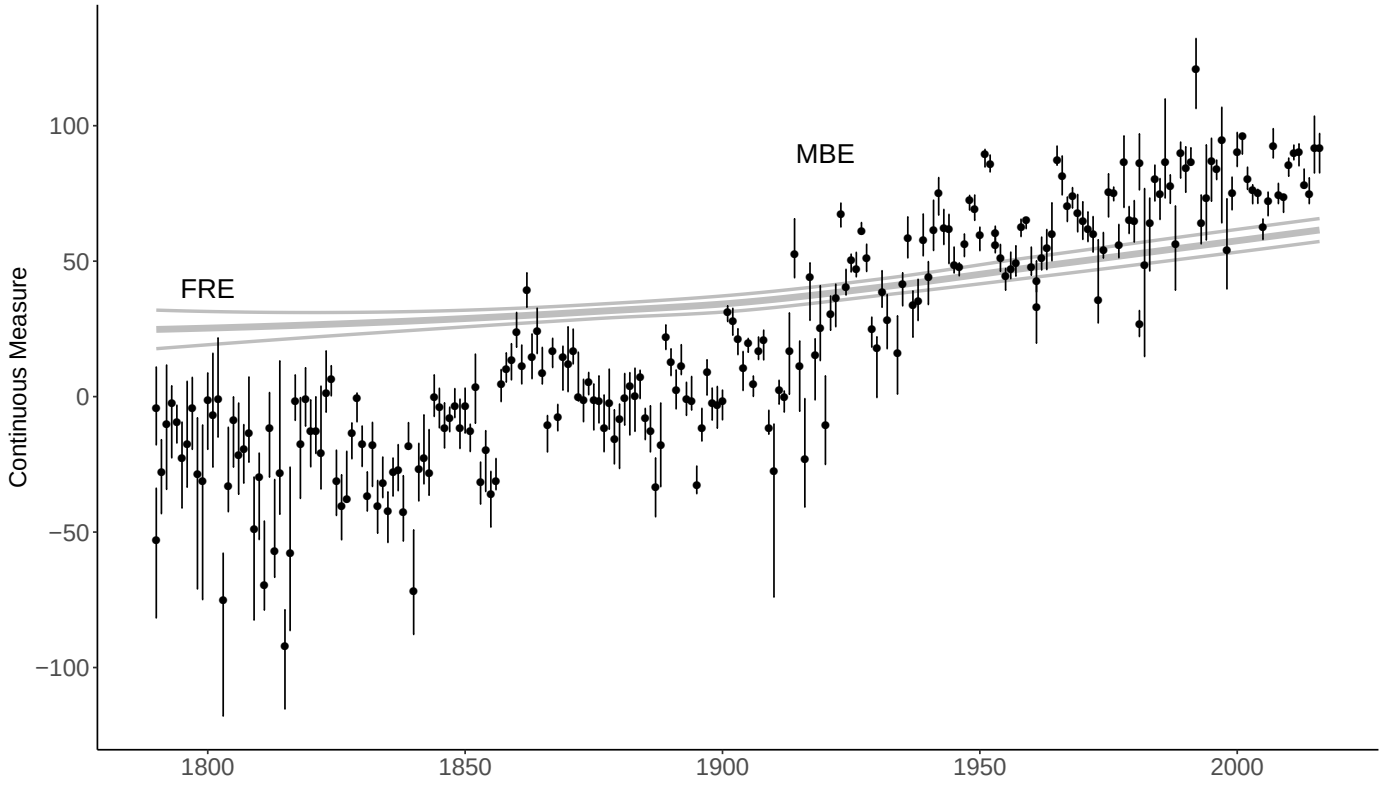

Figure 3: Comparing the linear, continuous version of our model based estimates (points plus 95% confidence intervals, denoted MBE) to FRE (smooth lines, with outer edges representing 95% confidence intervals) of the State of Union addresses. Note that the time series are similar in direction, but our estimates start lower and end higher, implying a larger effect of new broadcasting technologies at the beginning of the 20th Century. Confidence intervals estimated by sentence-level bootstrap.

declining fairly precipitously by the end of the time series. For FRE, there is also a rise and decline, but the variation is more muted: indeed, it is almost constant at around 55 on the 0–100 scale. But it is not simply the magnitudes of the measures that are distinct: their time series properties are also different. To see this, consider the simple linear regression of the measure on the session number (1 through 79). Clearly, the effect of 'time' on (mean) difficulty is not linear: for both time series, it increases, and then decreases. But the point at which the trend reverses is not the same for each measure. Specifically, we conducted a generalized version of the Chow test: that is, for each session in the data, we segmented the time series into two parts (before and after the session in question). We then looked for evidence of structural instability in the two segments. We did this

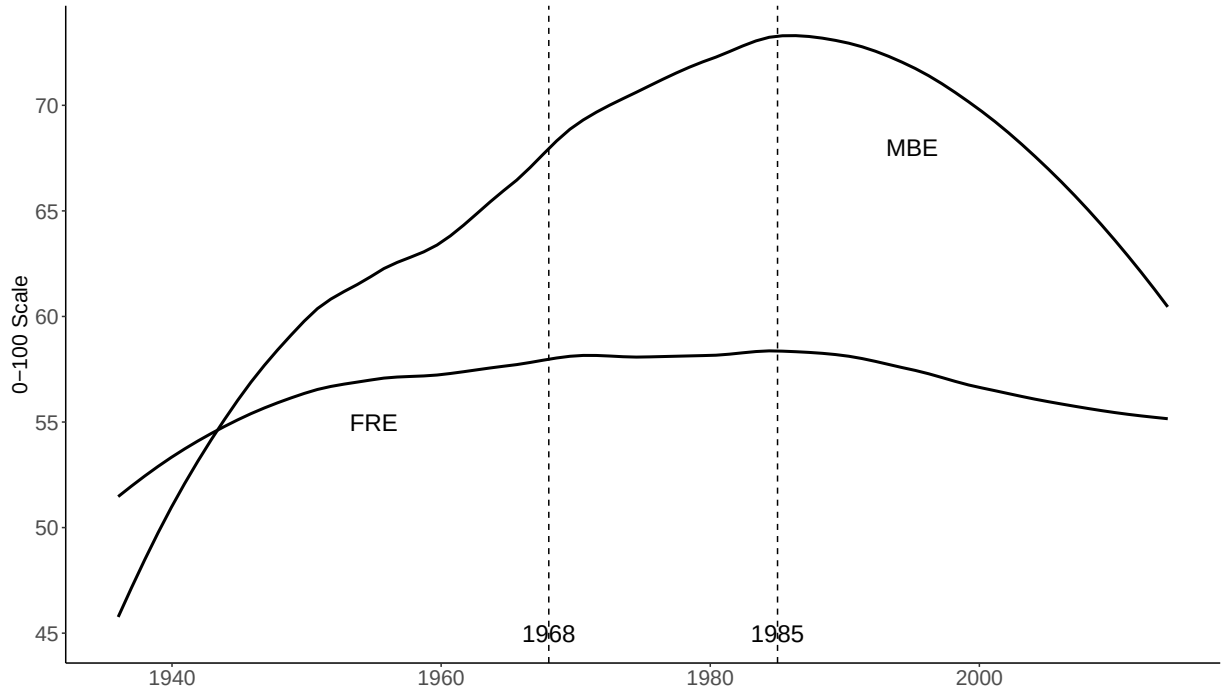

Figure 4: Comparing our mean model based estimate (MBE) with FRE estimates for 3 million speeches delivered by Members of Parliament. Note the break point for our measure is 1985, while for the FRE it is 1968.

using standard defaults as described by Zeileis et al. (2002). For the FRE series, the optimal break is in session 33, or around 1968. For our preferred approach, the optimal break is in session 50, or around 1985. Interestingly, both of these change points correspond approximately to technological shifts in terms of recording and broadcasting House of Commons proceedings.[22] In particular, in the spring of 1968, the House of Commons experimented with sound broadcasting. Ultimately, parliament would install permanent means of recording in 1978. By contrast, it wasn't until the late 1980s that television recording was approved—and it began in November 1989.

Obviously, it is very difficult to make causal claims from such aggregated, observational data. Still, the effects seem to be similar: with new technology, and new visibility, speeches become (on average) more complex. Why might this be? One argument made in the press [23] is that television,

[22] See House of Commons briefing on 'Broadcasting Proceedings of the House': https://www.parliament.uk/documents/commons-information-office/g05.pdf

[23] See e.g. 'Have TV cameras in Parliament made political debate coarser?' http://www.telegraph.co.uk/news/politics/11244147/Have-TV-cameras-in-Parliament-made-political-debate-coarser.html

in particular, encourages members to make longer opening speeches in debates. The idea here is that they do this to ensure their presence is noted by cameras, and that they can be quoted—possibly at length—on news programs. In general, making longer, more structured reports as speeches will tend to depress readability indices, especially if they substitute for shorter, 'punchier' statements. In the Canadian context,[24] there is some belief that television broadcasting encourages MPs to read their speeches, rather than speaking off-the-cuff. If so, this formalism will tend to drive the average statement to be more complex as measured by any approach.

The central lesson of this section is not the substantive finding of the breaks: we acknowledge this is somewhat speculative in nature. Rather, our purpose is to show another example demonstrating that FRE and our measure differ in important ways. Note that while, of course, the data here is similar—it is political, from elites, and in English—the precise setting is very different. Again, to the extent that our approach is trained in the right way, and the model is well thought out, we encourage users to prefer it over more traditional methods.

6 Discussion

The nature of the messages that political actors send one another are of obvious interest to Political Science, whether it be in American politics, international relations or from a comparative perspective. Yet a curious gulf has emerged in our studies. On the one hand, we have plenty of theory and empirical evidence that such communication matters: whether it be 'dog whistle' in nature, rhetorical, vague, or more explicitly designed to appeal to certain types of agents. On the other hand, *pace* some work we mentioned above, the discipline has been slow to adopt textual complexity measures in any context, preferring instead to code documents via pre-existing dictionaries, or in a more unsupervised manner altogether. This is despite the fact that the various readability measures are easy to use and scale in a straightforward way—which is important, given the sheer amount of text data now available to scholars. Presumably, part of this reticence is lack of familiarity with

[24] See 'Television and the House of Commons', https://lop.parl.ca/content/lop/ResearchPublications/bp242-e.htm

such approaches. But part of it is likely a very reasonable skepticism about the merits of these educational measures—a concern echoed in various non-Political Science arenas (e.g. Sirico, 2007; Loughran and McDonald, 2014) and indeed, increasingly in education itself Ardoin et al. (2005).

Rather than attempt to rehabilitate the indices, here we focused on producing something better. In particular, we used human coders (via the crowd) to provide relative assessments of short texts, and from there we built a well-defined statistical model. That model uses variables that differ from standard approaches, including word-rarity and parts-of-speech information. The final version performs better in fit terms too, although precisely because the approach is on much firmer probabilistic grounds it is hard to compare directly to previous approaches. Fundamentally then, we believe we have improved practice here: the approach is transparent, sensible and model-based and trained on relevant domain data. It is also flexible, in the sense that the work-flow and software we have designed allows end-users to calibrate the method to their specific problems. Quite how well our default model does on a given problem outside of the type reviewed here is, of course, to be determined. Nonetheless, we suggest that it will give more plausible answers than FRE or a related technique.

While our contribution is helpful for those interested in communication in politics, it is hardly the last word on the matter. We have provided a statistical machinery, and variables, for thinking more carefully about the measurement of sophistication or clarity in texts. What we have not done is produced a straightforward way to distinguish between more subtle understandings of such concepts. For example, one can imagine a politician—a president of the United States even—who uses relatively common terms in simple sentence constructions, but is not especially clear. By contrast, great academic writers might be able to describe extremely complicated ideas in straightforward ways. Our approach would generally be better than previous ones, but is unlikely to place these two extremes correctly on the same scale. This is, of course, because a 'sophisticated' idea (like democracy, or inclusively or conservatism) need not be complicated in expression, and vice versa. Going forwards, more attempts should be made—not least at the coding/crowdsourcing level—to iron out these differences, possibly by introducing different dimensions of complexity at the point

of testing or modeling. We leave such efforts for future work.

A Random Forest Variable Importance Plots

As noted in text, we ran our random forest model (1000 trees, otherwise standard defaults in the sense of Liaw and Wiener (2002)) for both sets of unstructured estimates—that is, with and without bias-reduction. The results of that process, in terms of the variable importance plots, are given in Figure 5. As usual, variables (on the *y*-axis) with points further right are deemed 'more important' for predicting the outcome (here, the snippet's ability). Notice that the ordering of the variables is similar, regardless of which approach we take (i.e. with or without bias reduction).

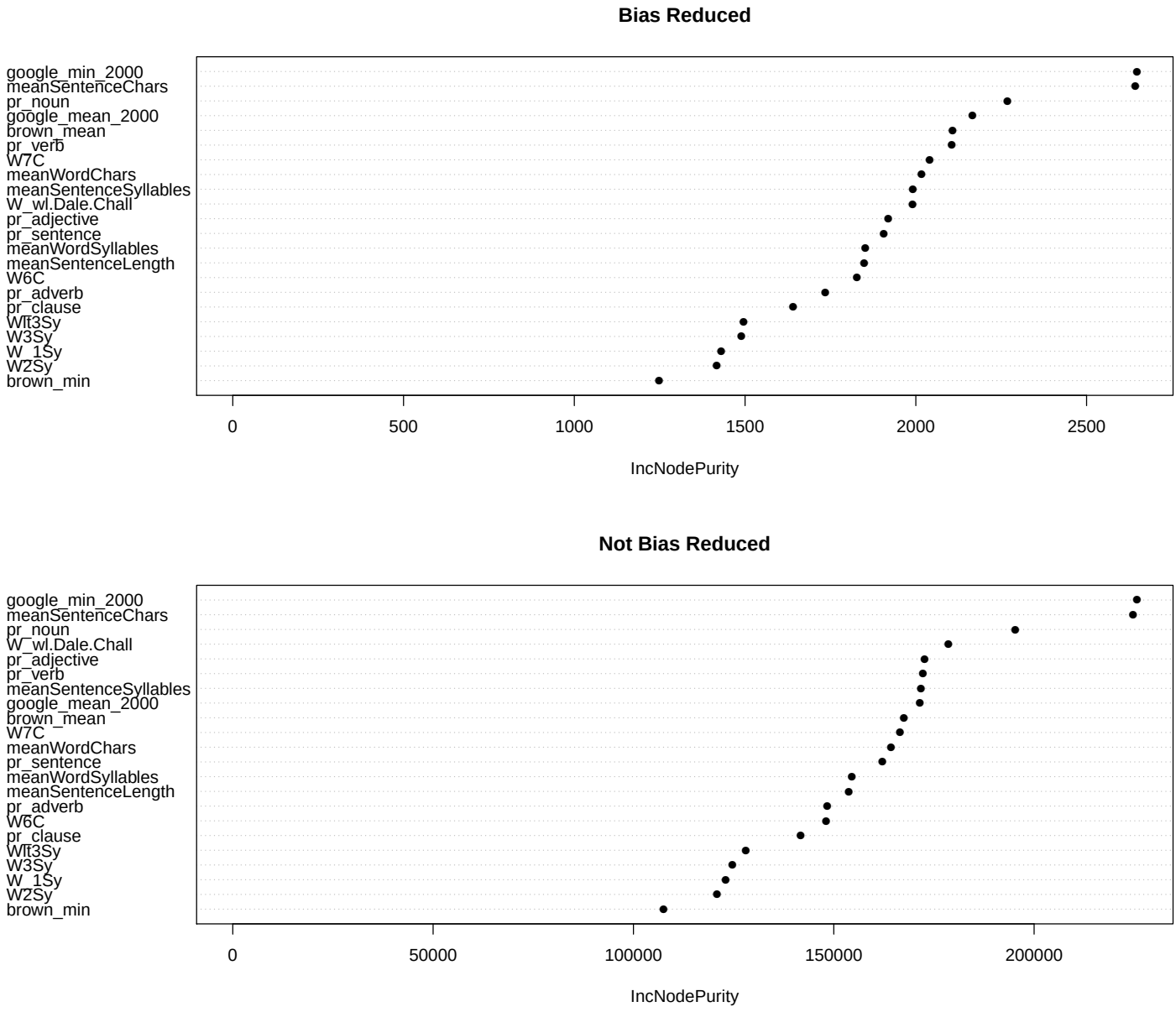

Figure 5: Variable Importance Plots for (unstructured) readability estimates. Note that points further to the right imply more 'important' variables. Top panel is for bias-reduced estimates; bottom panel is for non-bias reduced estimates.

References

Anderson, Jonathan. 1983. "Lix and Rix: Variations on a Little-known Readability Index." *Journal of Reading* 26(6):490–496.

Ardoin, Scott P, Shannon M Suldo, Joseph Witt, Seth Aldrich and Erin McDonald. 2005. "Accuracy of Readability Estimates' Predictions of CBM Performance." *School Psychology Quarterly* 20(1):1.

Benoit, Kenneth, Drew Conway, Benjamin Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(2).

Benoit, Kenneth, Kevin Munger and Arthur Spirling. Forthcoming. *Anxieties of Democracy*. Cambridge: Cambridge University Press chapter Dumbing Down? Trends in the Complexity of Political Communication.

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science* 58(3):739–753.

Bradley, Ralph and Milton Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39(3/4):324–345.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Cann, Damon, Greg Goelzhauser and Kaylee Johnson. 2014. "Analyzing Text Complexity in Political Science Research." *PS: Political Science & Politics* 47:663–666.

Coleman, M and T Liau. 1975. "A computer readability formula designed for machine scoring." *Journal of Applied Psychology* 60(2):283–284.

Dale, Edgar and Jeanne Chall. 1948. "A Formula for Predicting Readability." *Educational Research Bulletin* 27(1):11–20.

Diamond, Larry. 2002. "What Political Science Owes the World." *PS: Political Science & Politics Online Forum* pp. 113–27.

Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80(1):27–38.

Flesch, Rudolf. 1949. *The Art of Readable Writing*. New York: Harper.

Flesch, Rudolph. 1948. "A new readability yardstick." *Journal of Applied Psychology* 32(3):221–233.

Fry, Edward. 1968. "A Readability Formula That Saves Time." *Journal of Reading* 11(7):513–578.

Fucks, Wilhelm. 1955. *Unterschied des Prosastils von Dichtern und anderen Schriftstellern: ein Beispiel mathematischer Stilanalyse*. Bouvier.

Gatto, John Taylor. 2002. *Dumbing us down: The hidden curriculum of compulsory schooling*. Vancouver: New Society Publishers.

Gunning, Robert. 1952. *The Technique of Clear Writing*. New York: McGraw-Hill.

Jansen, David-Jan. 2011. "Does the Clarity of Central Bank Communication Affect Volatility in Financial Markets? Evidence from Humphrey-Hawkins Testimonies." *Contemporary Economic Policy* 29(4).

Kincaid, J Peter, Robert Fishburne, Richard Rogers and Brad Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease formula) for Navy Enlisted Personnel*. Vol. Research Branch Report 8-75 Naval Air Station Memphis: Chief of Naval Technical Training.

Klare, George. 1963. *The measurement of readability*. Ames, Iowa: University of Iowa Press.

Kristof, Nicholas. 2014. "Professors, We Need You!".

Liaw, Andy and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2(3):18–22.
URL: *http://CRAN.R-project.org/doc/Rnews/*

Lim, Elvin. 2008. *The Anti-Intellectual Presidency*. New York: Oxford University Press.

Loewen, Peter, Daniel Rubenson and Arthur Spirling. 2012. "Testing the power of arguments in referendums: A Bradley–Terry approach." *Electoral Studies* 31(1).

Loughran, Tim and Bill McDonald. 2014. "Measuring Readability in Financial Disclosures." *The Journal of Finance* 69(4):1643–1671.

Lowe, Will and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21(3):298–313.

McCullagh, Peter and John Nelder. 1989. *Generalized linear models*. New York: CRC press.

Michalke, Meik. 2015. *koRpus: An R Package for Text Analysis*. (Version 0.05-6).

Owens, Ryan and Justin Wedeking. 2011. "Justices and Legal Clarity: Analyzing the Complexity of Supreme Court Opinions." *Law & Society Review* 45(4):1027–1061.

Rheault, L, Beelen K, Cochrane C and Hirst G. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis." *PLOS ONE* 11(12).

Sherman, Lucius. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston: Ginn.

Sirico, Louis J. 2007. "Readability Studies: how technocentrism can compromise research and legal determinations." *QLR* 26:147.

Spache, George. 1953. "A new readability formula for primary-grade reading materials." *The Elementary School Journal* 53(7):410–413.

Spirling, Arthur. 2016. "Democratization of Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915." *Journal of Politics* 78(1):120–136.

Spriggs, James F. II. 1996. "The Supreme Court and Federal Administrative Agencies: A Resource-Based Theory and Analysis of Judicial Impact." *American Journal of Political Science* 40:1122–1151.

Thurstone, L. L. 1927. "A law of comparative judgment." *Psychological Review* 34(4):273–286.

Tränkle, U. and H. Bailer. 1984. "Kreuzvalidierung und Neuberechnung von Lesbarkeitsformeln für die deutsche Sprache." *Zeitschrift f ̈ur Entwicklungspsychologie und Pädagogische Psychologie* 16(3):231–244.

Turner, Heather and David Firth. 2012. "Bradley-Terry Models in R: The BradleyTerry2 Package." *Journal of Statistical Software* 48(1):1–21.

Wheeler, Lester and Edwin Smith. 1954. "A practical readability formula for the classroom teacher in the primary grades." *Elementary English* 31:397–399.

Yuka, Tateisi, Ono Yoshihiko and Yamada Hisao. 1988. A Computer Readability Formula of Japanese Texts for Machine Scoring. In *Proceedings of the 12th Conference on Computational Linguistics - Volume 2*. COLING '88 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 649–654.

Zeileis, Achim, Friedrich Leisch, Kurt Hornik and Christian Kleiber. 2002. "strucchange: An R Package for Testing for Structural Change in Linear Regression Models." *Journal of Statistical Software* 7(2):1–38.