Measuring and Explaining Political Sophistication Through Textual Complexity*

Kenneth Benoit† Kevin Munger‡ Arthur Spirling§

October 30, 2017

Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a measure of word rarity derived from term frequencies in the Google books dataset. Our technique 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 reassess patterns in US and UK political corpora to demonstrate how substantive conclusions differ when using our improved approach.

`Sophistication` software available: `http://github.com/kbenoit/sophistication`.

Word Count: 9,494 (excluding Supporting Information)

*This research was partly supported by the European Research Council grant ERC- 2011-StG283794-QUANTESS.

†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 concern in the study of politics is 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. Thus, within academic studies, typically as part of a broader discussion concerning "dumbing down" (Gatto, 2002), observers have applied measures of textual complexity from educational fields to find that the sophistication of political language has steadily decreased over the past 200 years (e.g. Lim, 2008). Such concerns are echoed in popular presentations too: 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]

By contrast, and with more optimistic conclusions, other social science studies have used measures of textual complexity to link linguistic sophistication to outcomes, with a focus on the 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. In the context of the British parliament, Spirling (2016) applies readability measures to document the democratizing effects of franchise reform on elite speeches. Studying post-war Austrian and German elections, Bischof and Senninger (Forthcoming) find that simpler manifestos make for better informed voters. 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 average article in *Time Magazine* or *Reader's Digest*.

These applications share one trait: They equate important substantive characteristics of po-

[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.

litical, economic, or legal communication such as clarity or sophistication with indexes such as the Flesch Reading Ease (FRE) score (Flesch, 1948) (or something similar to it). These measures, however, were developed decades earlier in entirely different contexts, namely educational research and applied psychology. And it is not clear that they are still relevant for our applications—or indeed if they ever were. 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, a trend toward greater verbal simplicity could be a positive development if it improves the clarity of communication, but also might be negative if it represents "dumbing down" in the form of reduced sophistication.

To address such unresolved questions, here we systematically review the properties and statistical performance of current measures of textual difficulty, and develop a new measure of for political language. In what follows, we use the terms "difficulty," "sophistication," and "complexity" interchangeably. Our approach uses experimental data based on human pairwise comparisons of short extracts of political speech (e.g. Lowe and Benoit, 2013; Montgomery and Carlson, Forthcoming), which we then use to scale linguistic sophistication using a simple but well-defined statistical model. In particular, we employ a scaling approach developed by Bradley and Terry (1952) in which clarity of a text is treated as "ability." By moving measurement to a model-based approach, with the statistical mechanics that brings, we allow for sensible statements about uncertainty and inference: thus, one can make claims about the *probability* that a given text is easier or harder than another. This allows us to make meaningful *ratio-level* claims: that, for example, one text is twice as easier (on average) than another (relative to a baseline). 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 (designed to be) 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 better than more traditional alternatives.[3] Furthermore, we can be precise about each feature's relative contribution to complexity—via the inspection of a $\hat{\beta}$ in a standard generalized linear model arrangement. More generally, our methodological contribution is to provide a work-flow for scholars interested in measuring textual complexity for any substantive area.

To demonstrate how this new measure allows us to gain new insights on old problems, we compare it to the FRE 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 than the FRE and, if anything, texts in the modern period are much easier to follow than traditional approaches would suggest. That said, 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 show that by our measure, speeches since 1985 have increased in sophistication, mainly because of a rise in the usage rate of unusual terms, which classical measures developed from other domains fail to capture. We relate this to technological changes in how speeches are recorded and broadcast. Furthermore, we show that Labour governments look increasingly like Conservative ones, in terms of the language they use—especially after the 1980s. By setting out clear principles for measuring linguistic sophistication in the political domain, furthermore, we demonstrate the methodological superiority of our approach, and outline a general method for fitting appropriate measures to any context.

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

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

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, with the emphasis being on the easy measurement of the "readability" of a document. While there are a large number of indices for this task—indeed, Michalke (2015) references and implements no fewer 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 practical 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 measure 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.

[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)

In terms of technical details, for a given document, the available measures take into account some combination 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 of students who could correctly answer at least 75% of some multiple choice questions regarding a few select texts. This dependent variable was subsequently transformed to a zero to 100 scale. Fitting a linear regression with a constant and two predictors (average sentence length and average number of syllables per word), ultimately yielded the following formula for scoring documents:

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

As designed for the original application, this "Flesch Reading Ease" measure had the intended range "for almost all samples taken from ordinary prose" (225 Flesch, 1948).[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.

Other than indirectly through syllable counts, 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

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

[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.

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, deemed to be those known by 80% of fourth grade children (in 1948).

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. None of the studies to which we refer above developed their own measures fit to the domain, but rather adopted some variant of the existing indices, giving rise to an "out-of-domain prediction problem."

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. When political scientists score documents using these methods, they are essentially calculating out-of-domain (and obviously out-of-sample) *predictions*.[6] The problems caused by jumping contexts has frequently been noted in dictionary applications of text analysis (see e.g. Loughran and McDonald, 2014, on using generic sentiment dictionaries on financial documents), but produces more specific problems when designed to measure the sophistication of language.

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

[6]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.

simply in their education level from younger people, but also in their knowledge and understanding of the 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 language and linguistic style has developed considerably.[7]

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" for comparison.

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

[7] We return to this idea in some detail below, but as a trivial example to fix ideas, the term *computer* may have been difficult to understand in 1956, but much less so in 2016.

contest for a reader of a given comprehension level. But such information is not forthcoming. 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

Existing measures of readability are *composite indices* whose inputs are weighted. Since those weights are static (i.e. from one point in time), applying them to dynamic data such as time series causes particular inferential problems. To see this, suppose we hypothesize that the State of the Union addresses have gradually adopted less sophisticated language over time. If we use FRE or its close allies to assess this claim, we assume that the only relevant information for the hypothesis test comes from the features of the documents—that is, the Xs. But "dumbing down" could occur (or not) as a consequence of changing *weights* (the $\hat{\beta}$s) too. Traditional approaches cannot speak to such claims directly.

For the reasons we have advanced above, there are compelling 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 a fairly arbitrary manner. For instance, the Dale-Chall method provides a list of 3,000 familiar words, with any word outside this set having a constant weight, regardless of its actual commonality. Such lists are not updated as language changes. Within the Dale-Chall words, we find *locomotive*, a term relatively unknown in this century outside of children's shows; and *telephone*, a term signifying technological advances in 1948, but unknown in 1848 and archaic in 2008. By contrast, *television* is absent from the list.

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 manner, recruited through a crowd-sourcing platform (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 (see Montgomery and Carlson, Forthcoming, for discussion). 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 is strictly between zero and one. 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, the researcher is not required to simply apply a rote formula again and again however dubious a given application.

3 Method: Crowdsourcing Complexity

With the above considerations in mind, we aim to discover the textual features that constitute complexity, in the context of specifically political language. At an intuitive level, our procedure is quite simple, and it begins by producing a series of short texts of one or two sentences each—fragments we refer to as "snippets"—which are given to human coders to compare, pairwise.[8] The coders tell us which of the two texts is easier to understand, and they do this multiple times for various pairwise combinations of snippets. We go from these pairwise decisions to a continuous scale of reading ease via the application of an unstructured Bradly-Terry model. Then, given those scores on the scale, we learn features of the snippets best predict their relative difficultly, as rated by the humans.

The human coding is performed on a crowdsourcing platform in batches of ten short comparison tasks, following the general procedures described by Benoit et al. (2016). The precise questions asked of coders, the way in which we ensured consistent quality in their responses, and the exact nature of the comparisons required from them is discussed in some detail in Supporting Information A. In our particular case, the snippets were drawn from the 70 State of the Union Addresses (SOTUs) delivered after 1950.[9] We used these texts because the purpose of the SOTU addresses has remained relatively unchanged in the postwar period, and because of the attention these speeches have received in previous examinations of readability. This gives us a benchmark of interpretation to which to compare our findings below, although our approach may easily be adapted to measure linguistic sophistication in other contexts.

Some preprocessing of the addresses prior to creating snippets was required: in particular, we removed some organizational non-sentence pieces of text (mostly referring to the medium by which the address was delivered). Once cut down for comparison, we disqualified some snippets from consideration: those which were outside the 0–121 range of the FRE; any containing more

[8] We take a candidly "bag of snippets" approach as a document model: we assume all relevant information is within the snippet rather than where or how it occurs in the document.

[9] As we explain below, we subsequently supplement these with some a small amount of earlier pilot comparison data we had.

than two numeric years; any with large numbers; and any beginning with the title of a document section.

We constrained the snippets drawn for comparison from our texts to three bands of approximately equal lengths, to avoid comparisons where deciding on the "easier" snippet appears easy because one is noticeably shorter than the other. Within each group of snippets of similar lengths, we sorted the snippets once by their FRE scores in ascending order and again in descending order, and combined the two lists to create a set of comparisons that vary from (very) dissimilar to (very) similar FRE scores.[10]

3.1 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 dataset,[11] which yields token counts on a yearly basis from 1505 until 2008.

To assess the how unusual might be a text for a modern audience, we computed the frequency of each term it contained relative to the frequency of the word *the* today.[12] 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, *husbandry* (the cultivation and breeding of crops and animals) was used much more often in the 1790s than in current times. Its inclusion in a speech would therefore make that document harder

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

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

[12] We used *the* since it is the most common word in the English language and because its relative frequency has remained relatively unchanged in several hundred years.

for a contemporary audience (such as our crowd coders). (To smooth out individual differences in the yearly samples, we combined the frequency counts for all years from 2000 through 2008.) We give more details on this process in Supporting Information B.

We also computed the relative frequency of parts of speech in each text, to obtain proportions of nouns, adjectives, verbs, prepositions, and so on. We did the same for some syntactic complexity markers such as the number (subsequently, proportion) of clauses in sentences. This allowed us to include these quantities in the feature set for fitting models below to predict reading ease. Our approach to obtaining these quantities is explained in Supporting Information C.

3.2 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 "easiness contests" between snippets. For a given contest, crowd workers must decide which of two snippets i and j 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 α_i/α_j.

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

$$\text{logit}[\Pr(i \text{ easier than } j)] = \lambda_i - \lambda_j. \tag{1}$$

Subject to specifying a particular snippet as a "reference snippet" (whose easiness is set to zero), this setup 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 average word's number of syllables, etc. This is achieved

by modeling the easiness of a given snippet as

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir}. \tag{2}$$

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.

The model is sufficiently flexible to be adapted to address these concerns directly, although here we have kept our formulation deliberately simple. Our primary interest is in estimating the complexity of documents by predicting (that is, scaling up) from the snippet results, for which we need estimates of their relative weights in predicting the human ratings of easiness, not a fully specificed model of coder and sentence effects.

3.2.1 Variable Selection via Machine Learning

For any specific application, it is not obvious which variables should be included in a given model of readability, but with our measures from the *unstructured* Bradley-Terry scaling, we can attempt to predict the variation in this ability scale and use the results to choose the relevant covariates for fitting our own, domain-specific measure. Our scaling returns an estimate of an "ability" λ_i (in this case, relative easiness) for each snippet, but makes no use of covariates.[13] We then use all our various text characteristics as features to predict these (unstructured) abilities using a random forests approach (Breiman, 2001), and then 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 2 to obtain the relevant coefficient estimates.

4 Results

We have two main sets of results. First, we can compare the standard measures as applied to specifically political text: the first such attempt that we know of. In Supporting Information D we give more details but one observation is worth noting immediately: the models all perform very similarly, with little to separate them in terms of either model fit (Akaike information criterion) or accuracy (proportion correctly predicted). The best performer on our data was the Spache measure, but the FRE is almost exactly as useful and will be preferred on familiarity grounds. We use it in our running comparison for what follows.

Second, and much more importantly, we provide a new measure of complexity based on our crowdsourced data and the inferences we draw from our machine learning approach.

[13] 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.

4.1 Augmented Bradley-Terry Approach

In Supporting Information E we report details of the random forest models that we ran on the unstructured abilities, along with variable importance plots for the same. We find that the model favors the rarity measure based on the recording the least commonly occurring term in the snippet (relative to the frequency of *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)'s 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 2 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).

Our model does not include a measure of word length, despite this feature being one of the two core components of the Flesch index. 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

Table 2: 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

All coefficients are statistically significant at the $p \leq .05$ level.

number of characters per word (`MeanWordChars`). As an experiment, we added this variable to our machine learning model and re-ran the analysis. The results of that process are in the fourth column of Table 2 titled "Best Model," which 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 as 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. 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 performs best when two documents are similar other than the proportion of nouns they contain, or the rarity of their words. In the contests for which our model outperforms the Flesch version to the greatest extent, it is the word rarity input that matters most. 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,[14] 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, however, 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

[14]This snippet appears per discussion in Supporting Information A about including some older texts from an earlier pilot study.

Table 3: 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

general *approach*—of training on relevant data and providing model-based estimates—is preferable for the reasons we gave above. Even if one wanted simply to use the Flesch set up (in terms of its component variables) based on Table 2 we would recommend local data for that purpose.

5 Applications to political text

We can apply the results of our model in various ways. We outline two obvious approaches before demonstrating how they might be used in practice. First, given Equations 1 and 2, we can obtain a (point) estimate of the probability that any given text i is easier (or conversely, more difficult) than any other text j by calculating

$$\Pr(i \text{ easier than } j) = \frac{\exp(\lambda_i)}{\exp(\lambda_i) + \exp(\lambda_j)}. \tag{3}$$

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 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 3 gives the relevant covariate values for our best model above.

Using the coefficients from Table 2, 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,[15] and estimated the relevant λ to be -2.175897. Thus, the probability that the Eisenhower text is easier 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). Note that the differences between texts mean something extremely well-defined here: we can make concrete statements about *how much* easier one document is relative to another, and the quantity refers back to a sensible model. This is quite unlike FRE, where as we noted, a difference of 5 points on the scale has no natural, cardinal interpretation.

Along with model-based estimates, researchers may also want a quantity analogous to the continuous 0–100 scores from the Flesch (1948) (regression) formula. Our proposal 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.[16] For a given data set, a sensible way to proceed is to include a text(s) at the

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

[16] See Supporting Information F for an alternative approach.

fifth grade level (designated a score of 100), and one at the post-college level (designated a score of 0)—or whatever minimum and maximum is preferred)—and to then (linearly) scale all resulting λs based on those two end points.[17]

Experimenting with the continuous measure on the SOTU snippet corpus performs well in the sense that it returns point estimates on a 0–100 scale commensurate (but not identical) to the FRE equivalents. This works because it replaces a logit-style calculation that is not linear in the predictors 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 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 over the full range of points (~ 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 bottom left of the plot, our measure tends to score the points differently. Indeed, we assign a considerably lower ("harder") rating for the hardest texts.

5.1 Reanalyzing the *State of the Union* addresses

Recall that our snippets came from the SOTU 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 plots the relevant point estimates and 95% (simulated) confidence intervals (y-axis) plotted against the date of the relevant text. The probability 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

[17] 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.

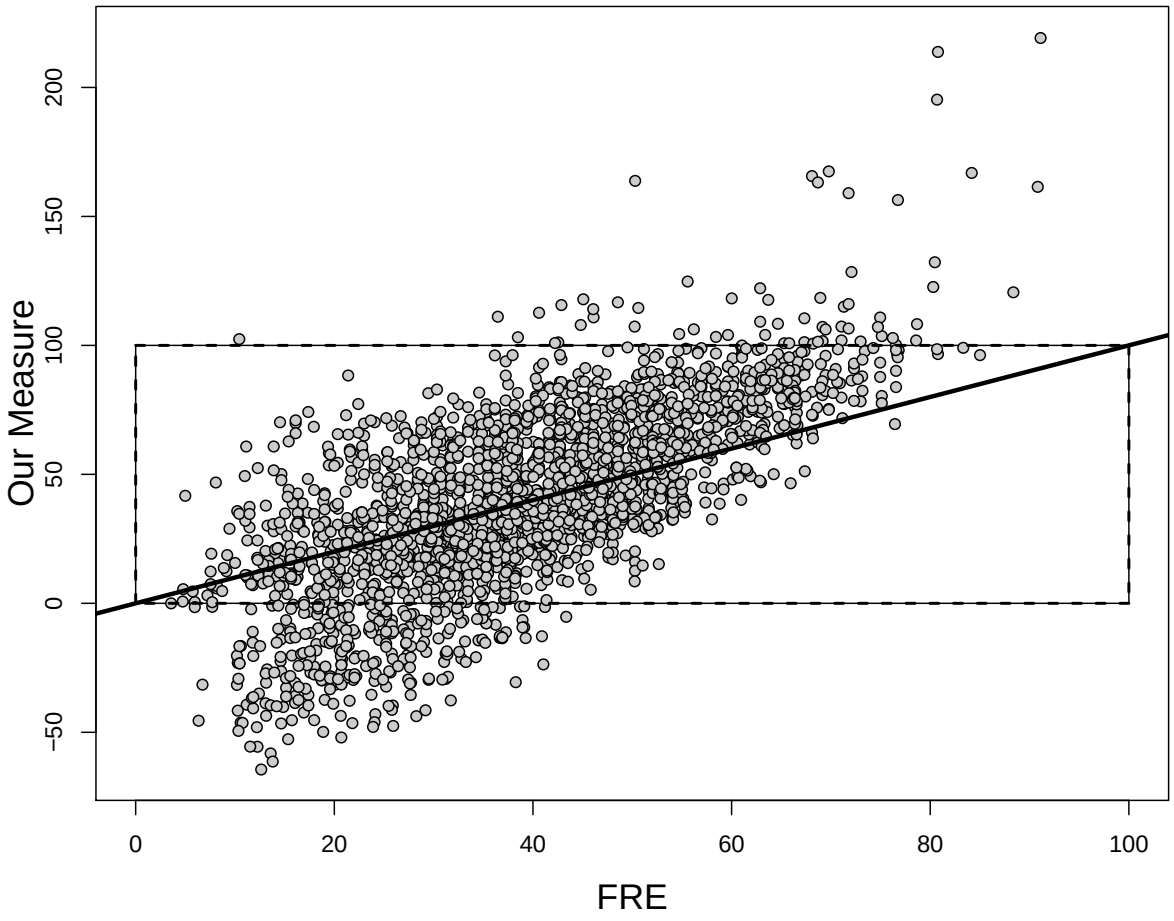

Figure 1: Comparing the "linear" version of our measure to FRE of the snippets. Correlation is generally high, especially for the theoretical range of the FRE (inner box).

2000s are not so different to those in the early 1950s). This implies that statements about the simplification of language 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 Figure 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

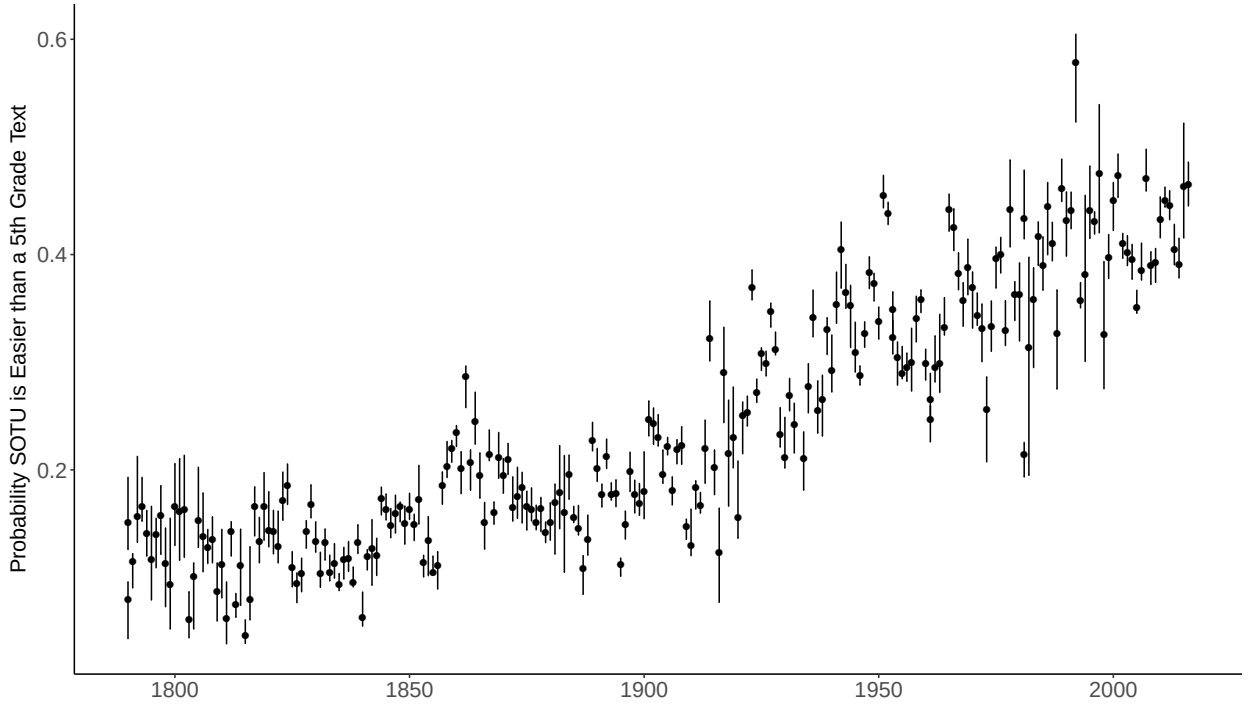

Figure 2: The probability that a State of the Union address is easier to understand than a fifth grade text baseline.

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. one may want difficulty estimated for contemporaneous audiences in 1800, 1810, 1820 etc, our framework allows one to

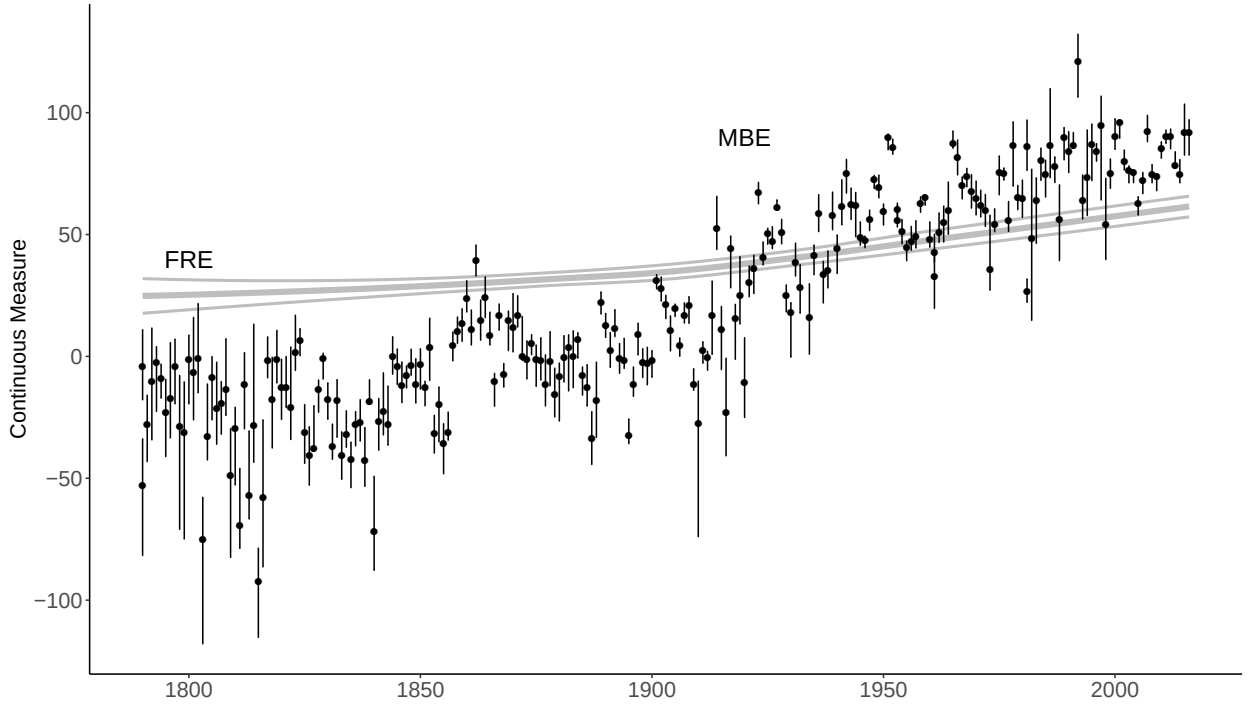

Figure 3: Comparing the linear, continuous version of our model based estimates (points plus 95% confidence intervals, denoted MBE) to FRE (smooth lines, with outer edges representing 95% confidence intervals) of the State of Union addresses. Confidence intervals estimated by sentence-level bootstrap.

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.[18]

5.2 *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 (See Rheault et al., 2016, for description). To keep our analysis simple, we focus solely on Labour and Conservative legislators, who represent around 90% of all

[18]Of course, we do not have coders from any other period, so one would need to make simplifying assumptions about the relevant coefficients.

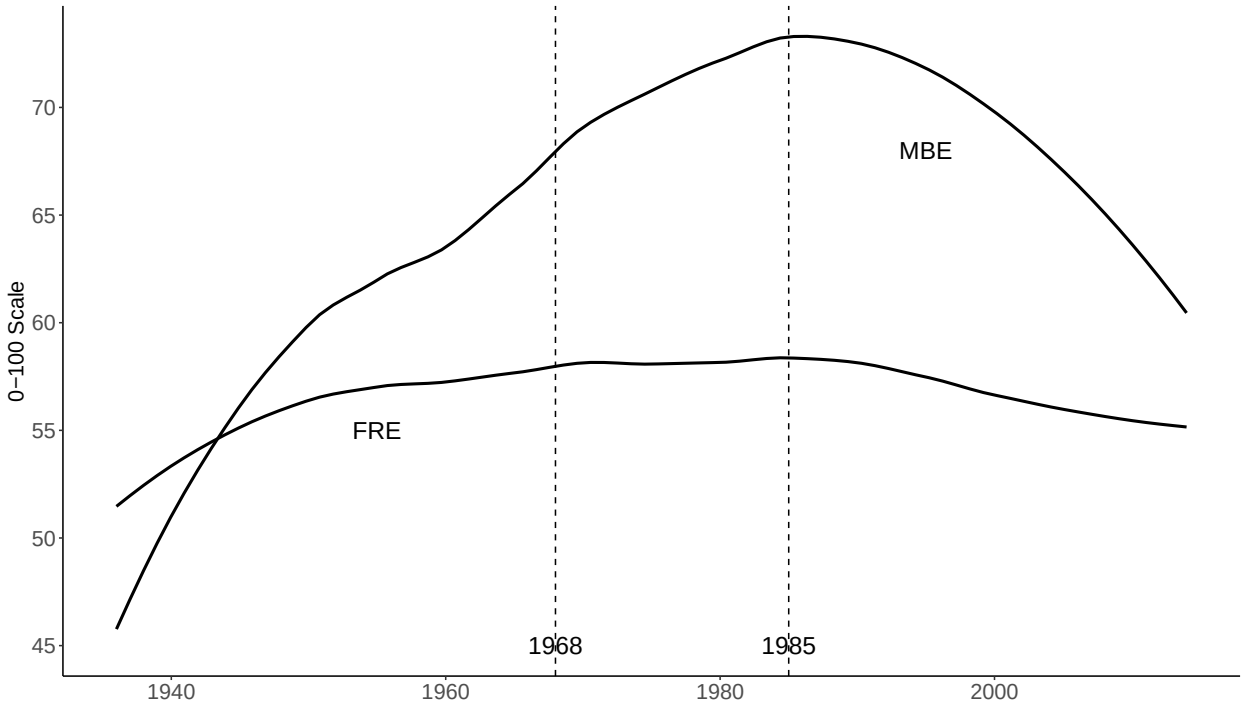

Figure 4: Comparing our mean model based estimate (MBE) with FRE estimates for 3 million speeches delivered by Members of Parliament. Note the break point for our measure is 1985, while for the FRE it is 1968.

MPs in the corpus. The data is compiled in "sessions" of parliamentary time, which last around a year a piece. We begin with by looking at the behavior of our continuous measure (relative to FRE) and then study the model based probabilities.

5.2.1 Speeches and Technology Changes

To begin, for each of the parliamentary 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.

Although the lines start in approximately the same place region of complexity (the rescaled measure on the *y*-axis), the speeches quickly become easier according to the MBE measure, before the trend reverses in 1985. The FRE, by contrast, is almost constant at around 55 on the 0–100 scale, after 1968. To identify the different inflection points, we conducted a generalized version of the Chow (1960) test. 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

between the two segments, 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.[19] 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[20] is that television, 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,[21] 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. To get a sense of the plausibility of this argument, in Figure 5 we disaggregate our measure into its four component parts, and study their (mean) behavior over time. We add a lowess curve in each case, and vertical lines for estimated breakpoints (see Bai and Perron, 2003) in the data (as implemented by Zeileis et al., 2002). The patterns are clear: the proportion of nouns per speech is rising over time (top left); the average length of words is rising (bottom left); speeches contain words that are rarer (bottom right); sentence lengths got shorter and then longer again (top right). This latter point is

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

[20] 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

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

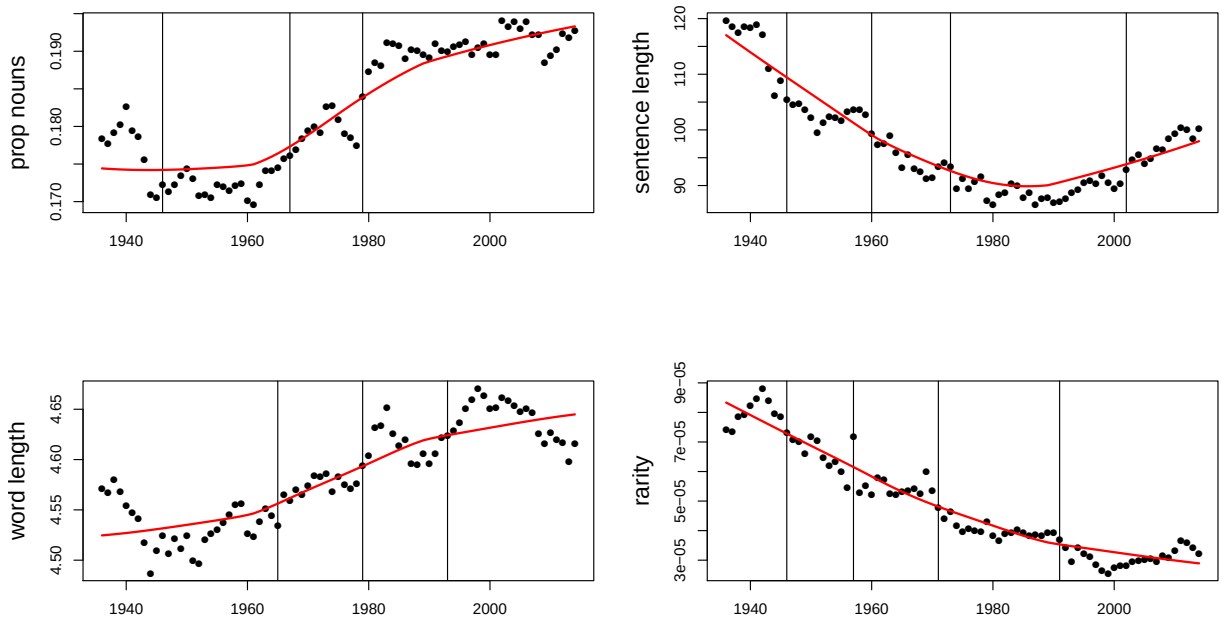

Figure 5: *Hansard* time series disaggregated by covariate in our measure, each point representing the average value for that session (with lowess line smoother added). Horizontal axis is the date of the session. Vertical lines represent estimated change points.

the key for our inference here: that is, only sentence length shows a pattern consistent with Figure 4. In particular, it seems that the most contemporary speeches involve longer sentences, which corroborates our earlier claims about the effects of television: somewhere between 1980 and 2000, something—we would argue the introduction of television—altered the data generating process.

5.2.2 Sociological Change in the House of Commons

The idea that descriptive representation might be an important characteristic of elected officials is not new (Pitkin, 1967). In recent times, however, scholars of British politics have specifically addressed its effects in the context of social class in the House of Commons (e.g. Heath, 2016). Empirically, they note that fewer and fewer Labour MPs in the post-war period come from (objectively) working class backgrounds, with an especially steep decline during and after the 1980s (Heath, 2015). Other scholars note that this is also true of subjective measures, wherein MPs are

asked to self-identify in class terms (Norris and Lovenduski, 1995). While the typical focus is on voter perceptions of politics, our measure allows to investigate how such changes affect discourse in parliament. Recall from Equation 3 that it is trivial for us to produce a probability that one text is easier than another. For the entirety of the *Hansard* data, we do just that for the mean value of λ (the "easiness" of a speech) for all Conservative and Labour MPs. That is, we calculate, for every session, the probability that the mean Conservative speech is easier to comprehend than the mean Labour speech. Note the contrasting strength of our approach with the weakness of traditional efforts. In particular, such a ratio is not directly interpretable in the Flesch context: e.g. the fact that text *A* has an FRE of 100, and text *B* has one of 50, does not mean *A* is "twice as easy" as *B*. For us though, the probabilities can be interpreted directly in these terms.

The results of this calculation are shown in Figure 6 as the plotted points. Those points are (blue) circles when the Conservatives are in government, and (red) squares when it is Labour. The means are equal at the 0.5 point on the *y*-axis, as noted by the broken line. We see immediately that when parties are in government, their (average) speech is easier to follow: this must be the case, because all the Conservative sessions in power are above the $\Pr(\text{Conservative easier}) = 0.5$ line, while all the Labour sessions are below it. But more interestingly, the trend of the data is towards a probability of 0.5, and we see this from the solid lowess line we imposed on the plot.[22] Put otherwise, Labour and Conservative speeches increasingly resemble one another in terms of difficulty. Why might this be? One possibility is that, with the general decline of working-class Labour MPs, both Conservative and Labour members are more similar in education, class and background than before. If we think social background matters for communication styles, then the convergence may be simply a consequence of sociological change in the House of Commons. In Supporting Information G we show that one possible mechanism is via changing word rarity: in particular, especially from around the 1980s onwards, Labour speeches (on average) use words that are rarer than in the past, and indeed rarer than those used by contemporaneous Conservatives. That is, one possibility here is that in line with their changing social and educational position,

[22] A simple linear regression with a dependent variable equal to the absolute deviation of the relevant probability from 0.5, with session number and party of government as regressors, corroborates the trend claim.

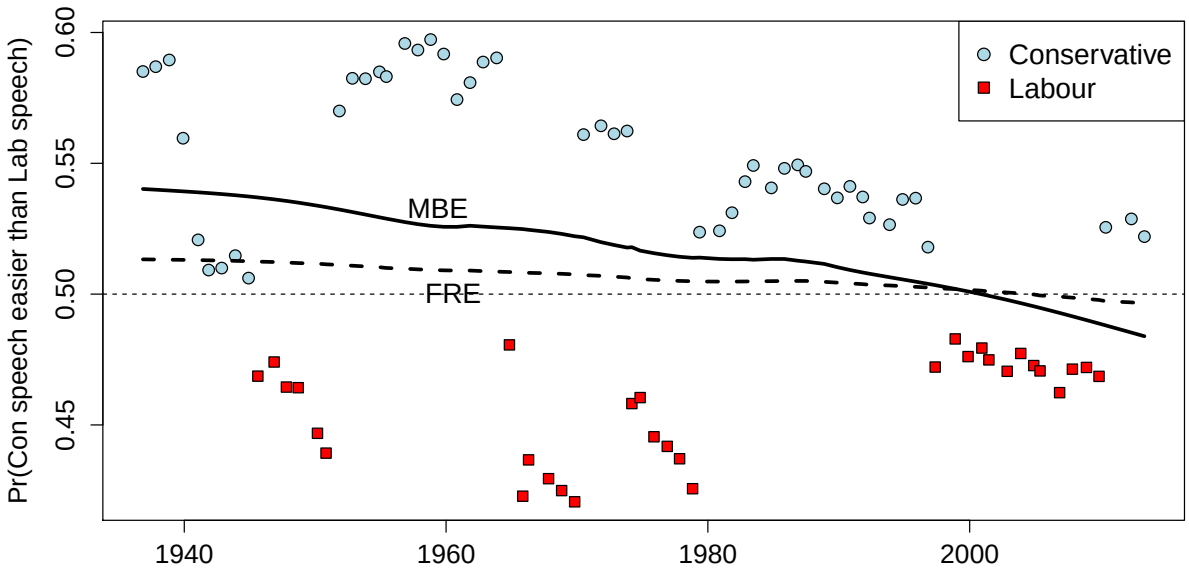

Figure 6: Estimated probability that the mean Conservative speech is easier than the mean Labour speech, over time. Point colors and shapes represent which party was in government at the time: (red) squares are Labour, (blue) circles are Conservative. Solid black line is lowess of probability over time ("MBE"). Broken line is lowess of the FRE *ratio* measure.

Labour members are departing from more basic vocabularies in favor of (relatively) more abstruse terms.

To show how our measure here improves over standard approaches, we include the lowess for an FRE *ratio*: the mean FRE for Conservative speech in a given year divided by the sum of the means for the Conservatives and Labour. While this is not a well-defined probability, its interpretation is more directly comparable to our model-based probabilistic estimate. One observations is immediate: the FRE ratio is considerably less variable than our measure, showing more stability over time. The shallower angle indicates that it fails to capture the full effect of the changing pattern in the political sophistication of language shown by our technique. (In Supporting Information G, we provide regression-based details, based on detrending the time series, that brings this difference in magnitude changes into starker relief.) In sum, our model is more sensitive to changing patterns in linguistic complexity in the Hansard example, because it was fit to the specific context

required.

6 Summary and Discussion

The nature of the messages that political actors send one another are of key 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 (Albertson, 2015), rhetorical (Riker, 1996), vague (Lo, Proksch and Slapin, 2016), or more explicitly designed to appeal to certain types of agents. On the other hand, the discipline has been slow to adopt textual complexity measures in any context, preferring instead to code documents using pre-existing dictionaries. 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 textual data now available to scholars. Presumably, part of this reticence is lack of familiarity with such approaches. But part of it is likely a very reasonable skepticism about the merits of these educational measures—a concern echoed in other fields of social science (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: Table 4 summarizes our contribution with respect to the problems we raised in Section 2.

Table 4: Summary of our approach as a solution to a series of problems with traditional approaches.

Problem with traditional approach	**Solution via our approach**
1. Designed for education	1. Designed for *politics*
2. Tested/validated on children	2. Tested/validated on *adults*
3. Designed for readers in 1940/50s, not easily updated	3. Designed for *contemporary* readers, easy to update (via crowdsourcing).
4. Cannot assess quality/fit of predictions for documents	4. Straightforward to assess *absolute model fit* (in training set) via usual metrics like percent correctly predicted
5. Cannot compare models of different forms	5. Straightforward to assess *relative model fit* (in training set) via usual metrics like AIC, BIC.
6. Cannot interpret fine-grained differences in document scores	6. Natural *model-based interpretation* of document estimates (via Bradley-Terry model).
7. No uncertainty around estimates.	7. Uncertainty *estimates available* both for variables in model, and on document scores (via bootstrap).
8. Composite indices/aggregate form hides changes in variables "under the hood."	8. Straightforward to examine all *changes to component parts*.
9. Rarity of terms accounted for in *ad hoc* inflexible way, if at all.	9. Rarity of terms *systematically derived* from large corpus, and available for any period of interest in past 200 years.

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 workflow and software we have designed allows end-users to calibrate the method to their specific problems.

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 con-

cepts. 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 for popular audiences. Our approach would generally be better than previous ones, but is still unlikely to place these two extremes correctly on the same scale. This is, of course, because a sophisticated idea (like democracy, or inclusivity or conservatism) need not be complicated in expression, and vice versa. 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.

References

Albertson, Bethany. 2015. "Dog-Whistle Politics: Multivocal Communication and Religious Appeals." *Political Behavior* 37(1):3–26.

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.

Bai, Jushan and Pierre Perron. 2003. "Computation and analysis of multiple structural change models." *Journal of Applied Econometrics* 18(1):1–22.

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. "Dumbing Down? Trends in the Complexity of Political Communication." `http://kmunger.github.io/pdfs/BenoitMungerSpirling_SSRCchapter.pdf`. Prepared for 'Anxieties of Democracy' volume (editors Frances Lee and Nolan McCarty).

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.

Bischof, Daniel and Roman Senninger. Forthcoming. "Simple politics for the people? Complexity in campaign messages and political knowledge." *European Journal of Political Research* .

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.

Chow, Gregory C. 1960. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica* 28(3):591–605.

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.

Heath, Oliver. 2015. "Policy Representation, Social Representation and Class Voting in Britain." *British Journal of Political Science* 45(1):173193.

Heath, Oliver. 2016. A growing class divide: MPs and voters. In *Sexier Lies and the Ballot Box*, ed. Philip Cowley and Robert Ford. Biteback.

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!" https://www.nytimes.com/2014/02/16/opinion/sunday/kristof-professors-we-need-you.html. New York Times, Online, February 15, 2014.

Liaw, Andy and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2(3):18–22.

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

Lo, James, Sven-Oliver Proksch and Jonathan B Slapin. 2016. "Ideological clarity in multiparty competition: A new measure and test using election manifestos." *British Journal of Political Science* 46(3):591–610.

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, V0.05-6*. https://cran.r-project.org/web/packages/koRpus/vignettes/koRpus_vignette.pdf.

Montgomery, Jacob and David Carlson. Forthcoming. "Human computation scaling for measuring meaningful latent traits in political texts." *American Political Science Review* . Accessed October 30, 2017: http://pages.wustl.edu/montgomery/sentimentit.

Norris, Pippa and Joni Lovenduski. 1995. *Political Recruitment*. Cambridge: Cambridge University Press.

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

Pitkin, Hanna. 1967. *The Concept of Representation*. Berkeley, CA: University of California Press.

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

Riker, William H. 1996. *The strategy of rhetoric: Campaigning for the American Constitution*. New Haven: Yale University Press.

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˙ür 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.

Supporting Information

A Details on crowd-sourcing, "gold questions" and snippet construction

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 operation is ensuring that workers provide high quality responses. To that end, we employ "gold standard" tasks: tests in which one snippet is unambiguously easier than the other, interspersed with normal rating tasks at a rate of one in ten. To create the gold standard test questions, we select the snippet pairs with the largest disparity in FRE scores, verified through inspection. Prior to being accepted for the task, a crowd worker had to pass a qualification test consistently entirely of test questions, answering at least 7 of 10 correctly. Following successful qualification, a coder performed job lots of ten pairwise comparisons, where one of these was a test question. Workers who did not maintain an overall accuracy rate of 30% correct on the test questions were removed from the pool of workers and their answers dropped from the dataset.[23]

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.[24]

[23] 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." Of the test questions, approximately 10% were screeners.

[24] To increase the range of data and to use results from an pilot study of coding, 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.

B Details on using the Google-books corpus

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, we ended up a table of frequencies for 82,558 unique word types from the total corpus.[25]

To see how this works, consider 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 2000s) 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*

texts with some with one- and two-sentence snippets from an earlier set of crowd work. 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.

[25] This was a fairly massive reduction from the over 615 billion term counts in the original term-year dataset. 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.

(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.)

C Details on obtaining part-of-speech information

We began by tagging the snippets using the Google Universal tagset[26] using the spacyr package built on the spaCy NLP library for Python.[27] 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 propositions. Converting these to relative frequencies for each snippet gave us the information required.

D Comparing the standard measures

In Table 5 we consider two natural ways to compare the fit of the standard approaches in the literature. 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 inputs 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 1. If

[26] See `https://github.com/slavpetrov/universal-pos-tags`.
[27] See `http://spacy.io`.

this probability is greater than 0.5, then we declare that a success for the model.

Table 5: 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

E 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 7. 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).

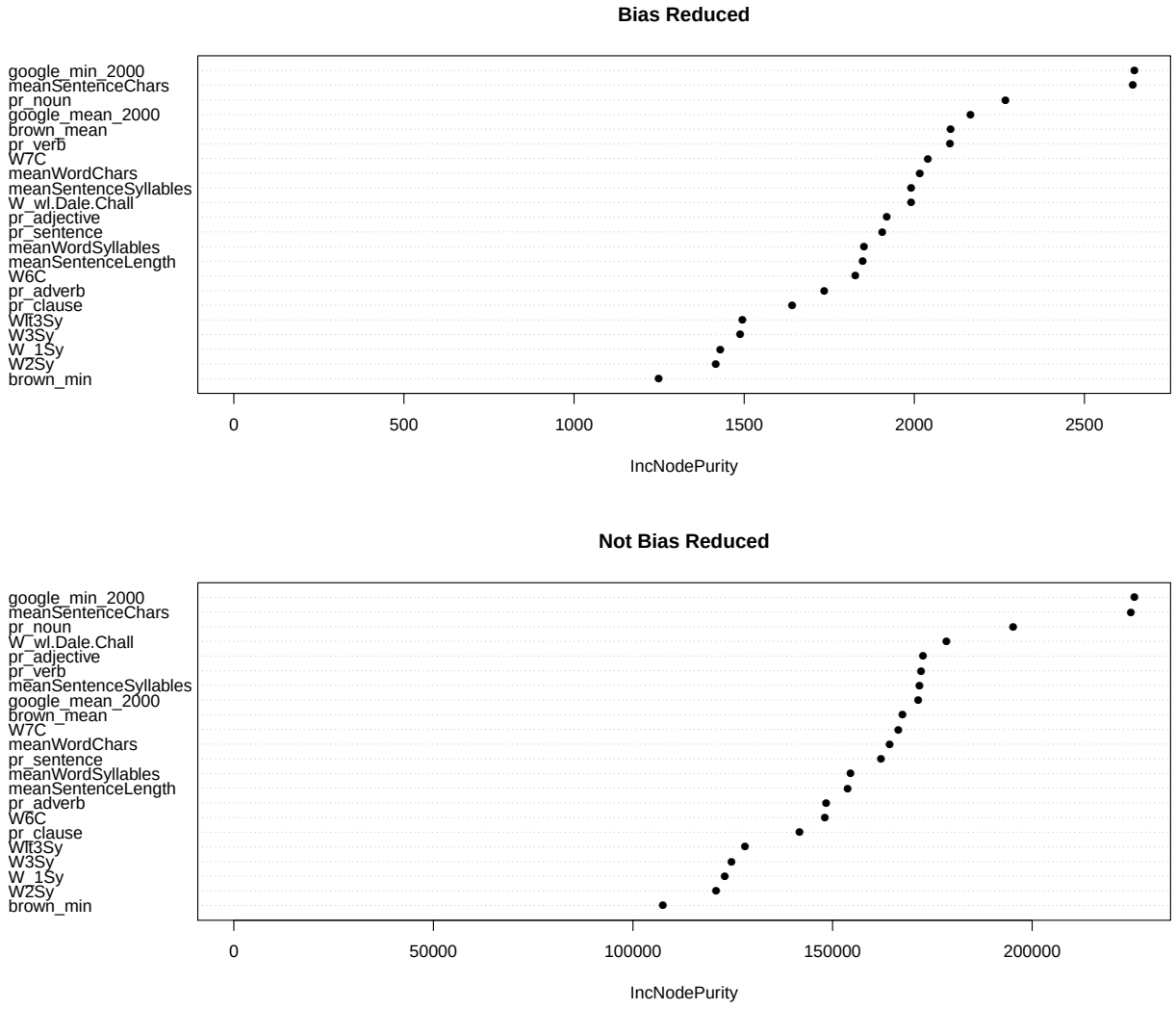

Figure 7: Variable Importance Plots for (unstructured) readability estimates. Note that points further to the right imply more important variables. Top panel is for bias-reduced estimates; bottom panel is for non-bias reduced estimates.

F An alternative continuous measure

There are ways to rescale the λ estimates that may be of greater theoretical appeal. To see this, using Equation 3 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

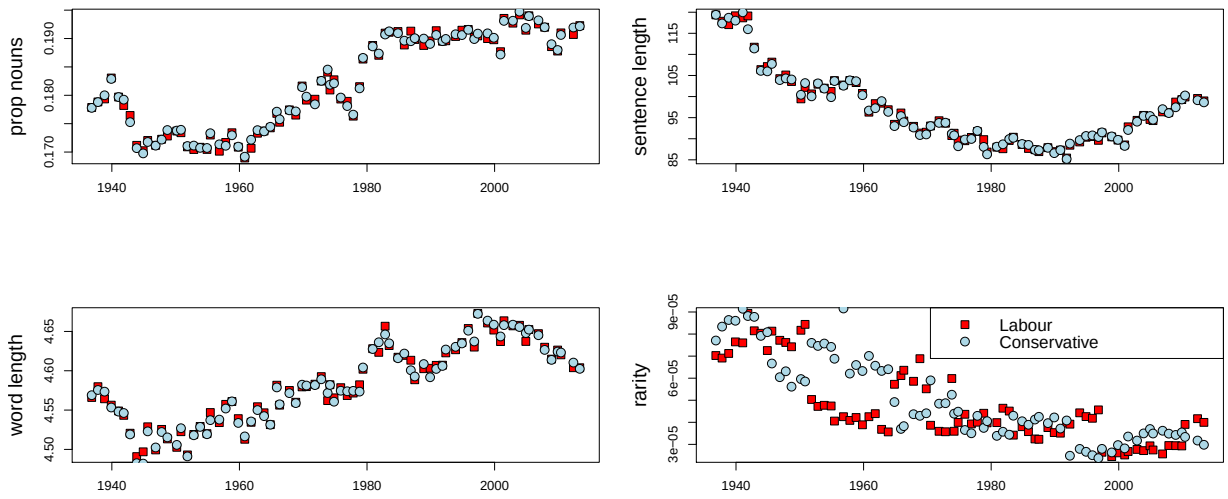

Figure 8: Disaggregation of speech difficulty by party over time (Conservative vs Labour). Note that the parties are essentially similar on all components, except rarity of speech.

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 original interpretation given by Flesch.

G Disaggregation of Conservative vs Labour patterns in government

Above, we noted that there is apparent periodicity in the time series of Conservative vs Labour (mean) speech difficulty. In particular, we noted that when a party is in government, its speeches tend to be harder to understand. To see why this might be, in Figure 8 we disaggregate our measure (for the mean speech) into its component parts, and divide out the data into Conservative and Labour means.

Clearly, the time series overlap: the [blue] Conservative circles overlap with the [red] Labour

squares everywhere with the exception of the bottom right—which is our measure of rarity. Looking at that subplot, we see the following pattern: prior to around 1945, when the Conservatives are in government, their (average) word rarity is larger, meaning they use terms that are more common than those used by Labour. The next five years (when Labour are in government) sees Labour using less rare words. Then when the Conservatives are in government in the 1950s and early 1960s, they use more common words. Labour switches to being the party that uses more common words after that (with the exception of the early 1970s when the Tories are in power for four years). By the 1980s—a period in which the Conservatives are completely dominant—the parties are more similar and almost overlap in word rarity terms; meanwhile, in aggregate, the rarest words used become more rare (the level shifts down over time). This pattern continues until the end of the data, although we note that Labour is generally below the Conservatives everywhere after around 1985 (the very end of the data being an exception).

Analysis of Detrended Data

After detrending the two time series (our estimates and those from the FRE ratio), we fit two linear regressions of the form $Y = \beta_0 + \beta_1 X_{\text{after 1997}}$. Here Y is the relevant measure, and $X_{\text{after 1997}}$ is dummy taking the value 1 if the session occurs after the Labour landslide of 1997, and 0 otherwise. If we think the 1980s was the key period of modernization for Labour, and was also a time of changing recruitment, then it makes sense to investigate the extra effect of time once Labour came to power after a break of 18 years.

Unsurprisingly, given the theory that Labour elites were now more similar to their Conservative peers, for both measures there is a negative effect of Labour gaining power in 1997. However, the coefficient for our (detrended) measure ($\hat{\beta_1} = -0.0167$) is about four times as large (in addition the model fit is better, and the p-value smaller) as that for the FRE (-0.0043). In that sense, then, our measure is more sensitive to the advent of new Labour elites than the most common extant approach.