Measuring Political Media Slant Using Text Data*

Lucas Shen†

October, 2021

Abstract

Most studies that examine media bias tend to focus on differences in coverage intensity. This study develops a notion of accuracy from existing NLP (natural language processing) methods as the outcome of interest in a simple regression-based analysis. In Singapore, where there are no detectable partisan differences in coverage intensity, this approach detects partisan differences via coverage accuracy. The proposed contribution is methodological, where accuracy can be a supplement to the usual intensity-based measures. Additionally, NLP and ML (machine learning) methods help deal with competing explanations. Finally, the findings are contextualised in institutional-specific machinery that potentially explains away detected differences but, at the same time, necessarily embeds private beliefs about a media that slants.

Keywords: Political media slant, Political economy of mass media, Textual data, Applied machine learning, Natural language processing, Coverage accuracy
JEL codes: C81, L82, N45, P48

*I am indebted to Giovanni Ko for both supervising and seeding the idea to this study. I am grateful to Madhav Shrihari Aney, Wai Mun Chia, Shubhankar Dam, Yohanes Eko Riyanto, Chris Sakellariou, Gaurav Sood, You Yenn Teo, Walter Edgar Theseira, and seminar participants at the Australasia Public Choice Conference and Nanyang Technological University for detailed comments and suggestions at various stages. I also thank a senior journalist from the national print media and a member of parliament for sharing insights into parliamentary logistics and media engagement in the Singapore parliament. All remaining errors are my own. Findings and interpretations in this manuscript are those of the author and not of any affiliated institutions. Click here for latest draft.

†Asia Competitiveness Institute, Lee Kuan Yew School of Public Policy, National University of Singapore. School of Social Sciences, Nanyang Technological University. Email address: lucas@lucasshen.com

1 Introduction

A common strategy in the literature that systematically measures political media slant is to compare differences in *intensity* of coverage (Larcinese et al. 2011; Puglisi 2011; Puglisi and Snyder 2011; Qian and Yanagizawa-Drott 2017; Qin et al. 2018). This paper develops a notion of coverage *accuracy* using a novel dataset on direct quotations of speeches from the Singapore Parliament, which I construct using methods from automated text processing, machine learning (ML), and natural language processing (NLP).

A rich literature already exists on the US media (e.g., in Groseclose and Milyo 2005; Gentzkow and Shapiro 2010) where they find a slight left-of-centre bias. Their findings, however, cannot be directly extended to Singapore—where a monopoly supplies all daily newspapers and where a single party dominates the regular democratic elections. To test for political media slant in Singapore, I assemble a novel panel of direct quotation data by extracting the direct quotation of politician speeches in parliament, as reported in the flagship daily *The Straits Times*, and matching the quotes back to their originating speeches in parliament. This data allows me to compare the coverage accuracy of the ruling party and its opposition.

Specifically, this paper tests whether the *The Straits Times* quotes the opposition speeches from parliament with lower accuracy than their ruling-party counterparts. To quantify *coverage accuracy* any speech-quote pairing, I define two measures of quote accuracy using existing measures of edit distance.[1] The first accuracy measure this paper defines is a *substring* accuracy measure, which scores quote accuracy using the best partial-substring match—this is the quote-length substring within the speech that best matches the quote. The second measure, which is the

[1]Edit distance measures are solutions to the problem of quantifying similarity between two strings (of words in this case). A standard measure is the Levenshtein distance. For example, the Levenshtein distance between *intention* and *execution* is five because five operations (of insertion, deletion, or substitution of characters) is the minimum that is required to convert one string to the other. Applications of edit distance include NLP (e.g., spell-checking, machine translation, and mutual language intelligibility) and computational biology (e.g., similarities in DNA/RNA sequencing).

bag-of-words accuracy measure, on the other hand, scores quote accuracy by using the subset of words common to both quote and speech. The scores go from 0 to 100 (increasing in accuracy).

The empirical approach is straightforward. With quotes as the unit of analysis, I regress quotation accuracy (and intensity) on an opposition dummy using the OLS specification. If there are systematic differences in coverage between ruling-party and opposition politicians, the opposition dummy will pick this up. The main finding is that the parliamentary speeches of the opposition politicians are quoted with lower accuracy relative to the speeches of the ruling-party politicians. Conditional on the observables, the opposition speeches are 1.5 to 2.4 points (by the substring and bag-of-words accuracy measures) less accurate than those of their ruling-party counterparts, which is 11.6% to 22.9% of the standard deviations. Compared to the average accuracy of 91.4 and 96.4 per quote, the opposition get quotes that are 1.6% to 2.5% less accurate.

To deal with identification issues and competing explanations, and as part of the methodological contribution in this paper, I draw on rich textual content. First, I use unsupervised ML to recover the topic distributions from the text data so that each parliamentary speech and each news article have a probabilistic vector for topics (e.g. climate change or crime). Controlling for speech topics helps rule out the interpretation that observed partisan differences in coverage arise from partisan differences in speech content. Second, I use alternative construction of accuracy by removing stop words (e.g. "of", "the", "until") as part of the robustness checks. This helps rule out the case where the token distributions of opposition speeches skew towards the use of more stop words which can be ignored by journalists. A third alternative interpretation is that the opposition speeches are less coherent. To address this, I include controls of language competency from the quantitative linguistics literature, and the baseline conclusions do not turn on these controls.

To deal with any remaining bias, I turn to two bounding arguments. First, controlling for ministerial portfolio introduces a potential compositional bias since

opposition status also determines ministerial portfolio (as a case of bad controls Angrist and Pischke 2009). Here, the bias arises even if opposition status is random, and a set of institution-specific factors predicts that the OLS estimates are biased towards zero. The second bounding argument uses the proportional selection assumption (Altonji et al. 2005; Oster 2017)—and the data suggest that the unobservables drive the results in the same direction as the observables. This does not rely on the random assignment assumption and reinforces the OLS as conservative estimates on the extent to which opposition coverage is less accurate than that of the ruling party.

Overall, there is evidence of a subtle but systematic difference in coverage accuracy. At face value, the lower accuracy and higher fragmentation of quotations for the opposition suggest a media slant towards the ruling party. Through primary research, I contextualise the statistical findings in institutional-specific media-political machinery where the ruling-party politicians grant media agents early access to their speech transcripts, but the opposition party does not. I finally argue that even if this institutional-specific machinery can explain away coverage accuracy, the machinery itself embeds private beliefs about a media that slants.

The proposed contribution of this paper is methodological. Using coverage accuracy to discern bias can be used for any media platform with a known transcript and abstracts away from other media biases (e.g., gatekeeping) that are less testable. Moreover, even if assignment into political parties is determined by unobserved characteristics that also determine coverage, the confounding bias from accuracy as the basis of comparison is less severe than that using intensity since direct quotation should be accurate in any case. In the media of Singapore, there would have been no systematic evidence of slant if the only measure was coverage intensity. To my best knowledge, this paper is the first that narrows down media bias to coverage accuracy using the text data from politician speeches.[2] Another contribution is

[2]A related contribution is on how machine learning has its own place in the applied econometric toolbox (as discussed in Mullainathan and Spiess 2017). Supervised machine learning in this study expedites the data collection, while unsupervised learning affords machine-annotated data free from

the large-scale systematic evidence of political media slant in Singapore, which complements other richer descriptions of media and politics in the same context (e.g., George, 2012).

This paper otherwise relates most to the literature that estimates the size and direction of media bias without relying on anecdotes—in particular, to the literature that uses textual information from congressional speeches (Gentzkow and Shapiro 2010), that takes an incumbent-government dimension (Larcinese et al. 2011; Durante and Knight 2012; Lott and Hassett 2014), that uses the intensity of coverage to discern bias (Larcinese et al. 2011; Puglisi 2011; Puglisi and Snyder 2011; Qian and Yanagizawa-Drott 2017; Qin et al. 2018), that looks at the interaction between the media and a dominant single-party (Enikolopov et al. 2011; Miner 2015; Qin et al. 2017, 2018), and the literature that finds a slight left-of-centre bias in an otherwise centrist US media (Groseclose and Milyo 2005; Bernhardt et al. 2008; Ho and Quinn 2008; Sutter 2012).

More broadly, this paper relates to the rich literature on the political economy of mass media—on how mass media can facilitate accountability (Besley and Prat 2006; Bernhardt et al. 2008; Ferraz and Finan 2008; Snyder and Strömberg 2010; Larreguy et al. 2014; Lim et al. 2015), have welfare effects (Besley and Burgess 2002; Strömberg 2004; Corneo 2006; Eisensee and Strömberg 2007), and impact electoral outcomes (Gentzkow 2006; DellaVigna and Kaplan 2007; Boas and Hidalgo 2011; Chiang and Knight 2011; Enikolopov et al. 2011; Gerber et al. 2011; Adena et al. 2015; Miner 2015).

Section 2 starts by providing a background to the political and media institutions of Singapore. Section 3 details the use of ML and NLP in constructing the panel and accuracy measures. Section 4 begins by discussing the empirical strategy before presenting the results at the article-speech level. Section 5 presents the results at the quote level. Section 6 rules out additional language-based explanations, Section 7 follows with discussion. Section 8 concludes.

a researcher's (my) subjective priors.

2 Background of Political and Media Institutions

2.1 Political Background

Singapore has a unicameral parliament based on the Westminster system. Each parliament has a maximum term of five years, after which a general election must be held within three months of the dissolution of parliament. Political candidates are elected into parliament based on plurality voting.

Group and Single-Member Constituencies. Since the late 1980s, most parliament seats are contested under group representation[3]—voters in group constituencies vote for a group slate instead of any single politician. For instance, in the 2011 general election, only 12 of the 87 parliament seats were contested as single members. Groups must be made of between 3 to 6 candidates, of which at least one must be of an ethnic minority. In the sample, group sizes vary between 4 to 6. The regression analyses control for politician ethnic and group size.

Non-Elected Members of Parliament. Two other schemes were implemented by the early 1990s. One is the non-constituency member of parliament scheme, allowing the best-performing opposition losers to get seats. The other is the nominated member of parliament scheme, where a selection committee appoints non-partisan individuals. In the regression analyses, the ministerial controls include the politician type categorical variable, which includes the above two types of politicians and other types (e.g., minister or parliamentary secretary).[4]

Political Competition in Recent Years. The dominant and ruling political party has always been the People's Action Party (PAP), which forms the government with little legitimate challenge from opposing parties since 1959. The election in 2011, however, cast a different hue. The opposition contested all but five seats. Moreover, the main opposition of the day—the Workers' Party—won a breathtaking

[3]Also known as multi-member districts elsewhere.

[4]Before April 2017, the non-constituency and nominated members have similar but fewer voting rights compared to elected members. From April 2017 onwards, non-constituency members will have the same voting rights as members of parliament, but this does not affect 2005–16 sample.

6 of the 87 seats in parliament. This election marks the first time in Singapore's political history where an opposition party won a (5-member) group constituency, with the sixth seat won in a neighbouring single-member constituency. The margins of victory for the opposition were non-trivial: the group won with a 9.4% margin and the single member by a 29.6% margin. This political background provides a basis for the sample period of 2005–16, approximately five years before and after the landmark 2011 election.

2.2 Mainstream media background

Media-Related Regulations. The political dominance of the ruling party has allowed it to put in place media-related legislation, which directly regulates the publication of print media and potentially influences the content and tone of print media. The most direct regulation is the *Newspaper and Printing Presses Act (NPPA)*, first enacted in 1974. The NPPA requires that no newspapers are to be printed or published without permits. Moreover, the NPPA requires that newspaper companies be publicly listed with two classes of management shares: ordinary shares and management shares. Holders of management shares are entitled to 200 votes to the one vote allotted to the ordinary shareholder (Singapore Statutes Online 1974). These management shares can only be transferred to those with the government's approval, potentially creating the perception of government-vetted nominees.[5] [6]

Indirect media-related legislation also potentially affects media coverage. The *Internal Security Act* bans subversive documents and publications. The *Official Secrets Act* bans publications containing state secrets. The *Sedition Act* encompasses a wide range of offences defined as crimes committed against the state. The *Defamation Act* covers libel. The interpretation of these laws, however, leave plenty of room for ambiguity, lending weight to the perception that journalists practice

[5]This concept of separate share classes is not unique. *The New York Times* for instance, has class B shares for family owners with more voting privileges than class A shareholders.

[6]The Workers' Party, however, have had no issues getting permits for their newsletter—*The Hammer*.

self-censorship.[7]

Dailies in Singapore. Mainstream media in this paper mainly refers to the daily newspapers (or just dailies). The Singapore Press Holdings (SPH) company wholly owns eight of the nine local dailies in Singapore. The ninth daily—the *Today* newspaper—is jointly owned by SPH (40%) and MediaCorp (60%).[8] The flagship daily in Singapore is *The Straits Times*, an English publication with the largest readership (Table A.1).[9]

As per the requirements of the Newspaper and Printing Presses Act, SPH is publicly listed on the Singapore Stock Exchange, with 99.9% of ordinary shares held by the public. The remaining 0.1% are management shares held mainly by financial institutions. A relatively small number of management shares is held by the CEO and directors of SPH (less than 0.1% of management shares).[10] In other words, most management shareholders (those with 200 votes per share) are profit-maximisers rather than owners with ideological agendas (in the vein of Anderson and McLaren 2012; Durante and Knight 2012), though the fact that notable senior management positions are occupied by those with links to the government has also been documented (BBC News 2013; George 2012).

[7]Another media-related legislation implemented in 2019 is the *Protection from Online Falsehoods and Manipulation Act*, which gives the Executive the power to order for a correction of a falsehood shared online, with recourse through an appeal system either through the Executive or through the judicial system. This legislation is more directly related to non-mainstream social media, with its first use on Facebook (see https://www.channelnewsasia.com/news/singapore/brad-bowyer-facebook-post-falsehood-pofma-fake-news-12122952), though it does impose a general and additional restriction on media content.

[8]From a 2004 press release: http://www.sph.com.sg/media_releases/150 (Retrieved: 30 May 2017).

[9]The English daily with the next highest readership—*The New Paper*—has a print subscription of an order smaller (approximately 70 thousand compared to 300 thousand).

[10]http://sph.listedcompany.com/main_shareholder.html.

3 Data

3.1 Downloading and Matching of Quotes to Speeches

Downloading the Textual Data. For *The Straits Times* news articles, I query Factiva using the names of politicians who served in parliament in the years 2005–2016. This returns 62,132 unique news articles (as identified by the title-date tuple), which mention at least one of the active politicians in the sample period. The parliamentary procedure transcripts are from the publicly available and official repository. I automate the extraction of speaker-speech chunks by parsing the HTML tags. The resulting data, however, does not make a distinction between speeches on different types of parliamentary procedures, such as bills or motions. Section A.1 provides some background to the parliamentary procedures and speeches.

Supervised Classification of Articles Containing Quotes. Of the 62,132 articles containing mentions of the politicians, only a fraction will contain relevant direct quotations from parliamentary speeches. To automatically and accurately identify these relevant news articles containing the quotes of politicians, I use the *random forest* algorithm—a supervised learning classifier that takes an initial learning set of 1,419 manually-labelled articles—to predict which of the 62,132 articles contain quotations from parliament.[11] The random forest classifies the articles with 89% accuracy, 83% recall rate, and 70% precision rate. This classification yields 3,425 news articles (5.5% of the 62,132) with at least one quotation of a parliament speech.[12] [13]

[11]The random forest classifier is a bag of decision trees, where individual trees are constructed by bootstrapping over model inputs and the learning sample for a good bias-variance trade-off, and where each tree carries a vote on how to classify an article. I use class weights to mitigate learning biases because of the news articles' class imbalance—that a disproportionally large fraction of the articles does not contain quotes. I tune and evaluate out-of-sample prediction performance using k-fold cross-validation.

[12]It is also possible to use a *dummy* or *naive* classification which classifies as *yes* if a news article contains *parliament*$*$, and *no* otherwise. The naive classifier has an accuracy score of 0.78 (compared to 0.89 from the random forest, a 14% improvement), a better recall score of 0.93 (compared to 0.83), a precision score of 0.47 (compared to 0.70, a 49% improvement), and an F-score of 0.66 (compared to 0.75, a 14% improvement).

[13]For a sense of scale, if the 1,419 manually-labeled news articles took one week to do, parsing

Matching Quotes to Parliamentary Speeches. I define *quotes* as strings within single or double inverted commas or strings after speech colons. From the above 3,425 articles with parliament quotes, I automatically extract the quotations and match them to their originating speeches as follows (implementation is through a simple custom-written GUI, illustrated in Figure A.1).

Let f be the Ratcliff-Obershelp pattern recognition algorithm, an existing edit distance measure. Like other edit distances, it measures the minimum edits required to change a string, of words in this case, into another, providing a measure of string similarity or probability alignment between two sequences. The measure has the well-behaved properties of a metric (triangle inequality, non-negativity, and symmetry (Jurafsky and Martin 2000).[14] For a quote q and a set of speeches $S = \{s_1, \ldots, s_L\}$, the matched speech is $s^* \in S$ which satisfies:

$$s^* = \underset{s_\ell}{arg\,max}\Big\{\mathcal{F}_1(q, s_\ell) \;\Big|\; s_\ell \in \{s_1, \ldots, s_L\}\Big\},^{15}$$

where $\mathcal{F}_1$ is a composite function of f.[16]

This matching process yields 14,903 pairs of quotes and speeches, from 5,227 speeches made by 204 politicians, over the sample period 2005–16.

3.2 Quantifying Quotation Accuracy

The immediate issue with using the baseline measure f directly is that it will likely return an accuracy (similarity) score of zero because the quote needs to be edited (or added to) substantially before it resembles the original speech. In what follows,

the 62,132 news article would have taken approximately 43 weeks to for a single person.

[14]The Ratcliff-Obershelp pattern recognition algorithm is one among many string metrics in the computational literature. This paper uses the Ratcliff-Obershelp because it is an efficient built-in method in the Python ecosystem and that there is no reason to believe that other base string metrics would substantially change the baseline conclusions in this paper.

[15]$\mathcal{F}_1$ is the quote accuracy score (1) defined in the following subsection.

[16]In practice, I manually oversee this matching by looking at the best matches. Virtually all quotes are matched to a parliamentary speech that occurred a day before the print. This is consistent with news being *new* and explains why I do not include a control for the time gap between speeches and media pieces.

I define two accuracy measures based on the pre-existing Ratcliff-Obershelp edit distance. The two accuracy measures are defined specifically for this paper. I am not aware of them as standard measures in the NLP literature.[17] Mnemonic names are given to the two measures for ease of exposition.

Substring Quote Accuracy Measure. The first way I deal with this is to score quote *accuracy* according to the quote-length substring in the speech that best matches the quote. Let q be the shorter string and s the longer string, with lengths m and n, respectively. The substring accuracy measure locates the m-length substring within s that best matches q and scores accuracy according to this best partial substring match. Formally, substring accuracy measure for two strings q and s is:

$$\mathcal{F}_1(q,s) = \max_i \left\{ f\left(q, s^{(i:i+m-1)}\right) \;\middle|\; i \in \{1,\ldots,\ n-m+1\} \right\}, \tag{1}$$

where f is the Ratcliff-Obershelp measure, increasing in accuracy. The superscript $(i:i+m-1)$ indicates the location of the substring of length m. $\mathcal{F}_1$ goes from 0 to 100, increasing in accuracy (because f also goes from 0 to 100).

Concretely, if there is a quote q = "*a fundamental relook*" which is 20-character long and the originating speech recorded is s = "*... is taking a more fundamental relook at this regulation framework and see how best we can support this strategy ...* ", then substring accuracy measure searches for a 20-character string in s that best matches q, which is the 20-character substring [mor]"*e fundamental relook*", and then computes the edit distance between "*a fundamental relook*" and *'e fundamental relook'*. This score then forms the quotation accuracy score between the quote q and the originating speech s.

The substring accuracy measure merely formalises how direct quotations work. If a quote is verbatim, there should always be a quote-length substring in the speech that perfectly resembles the quote (perfect score of 100).[18] Even if the quote is not

[17] They may, however, be developed ad-hoc and use in other private applications relating to fuzzy searching.

[18] Discounting for punctuations which cannot be heard directly. I remove punctuations in pre-

verbatim, there should still be a quote-length substring in the speech that closely resembles the quote, as illustrated above.

Bag-of-words Quote Accuracy Measure. A different concern is when quotes have words taken from other parts of the same speech or when the order of words is not preserved. A concrete example is the quote "*. . . can afford their education in a public university*" which came from the speech "*. . . can afford to attend Government-funded universities . . . .*" Here, misquotation arises because a phrase that should appear in the verbatim quote (*"Government-funded"*) is replaced by another phrase (*"public university"*) that was used elsewhere in the same speech but otherwise refers to the same thing.

To allow an accuracy score more forgiving to type of error above, I define the second quote accuracy measure—bag-of-words accuracy measure. With slight abuse of notation, I first define a new (third) string $c = q \cap s$, as the string containing words common to both q and s. All three strings are then sorted according to alphabetical order to get $\widetilde{q}, \widetilde{s}$, and $\widetilde{c}$. Accuracy score (2) is then the maximum of the scores from all paired combinations of the three strings:[19]

$$\mathcal{F}_2(q, s) = \max\left\{ f\left(\widetilde{q}, \widetilde{s}\right),\ f\left(\widetilde{q}, \widetilde{c}\right),\ f\left(\widetilde{s}, \widetilde{c}\right) \right\}. \tag{2}$$

This second measure is useful for two reasons. Besides edits for syntactic flow and grammatical demands, the bag-of-words accuracy measure is more forgiving when the quote is not verbatim. This could happen because of genuine slips in attention by the reporting journalist during the sitting or because the transcript contains a (slightly) different version of the speech. Another reason for the bag-of-words accuracy measure is because the article-speech level analyses will concatenate all quotes coming from a speech and reported in an article as one observation. The

processing, as is standard.

[19]The max of the three components is taken instead of a weighted average for two reasons. One is that a weighted average requires specification of the weight for each component, and there is no clear way to do this. The other is that the measure is meant to be lenient in the first place—hence the maximum.

substring accuracy measure is no longer sensible in this case. However, the bag-of-words accuracy measure still offers a measure of quote accuracy by comparing the concatenated quotes to the originating speech using words common to both.[20]

Comparison to Other Records of Quotation Accuracy. Based on the quote accuracy scores computed using the substring accuracy measure, 19.8% of the articles contain some objective form of misquotation ($\mathcal{F}_1 < 100$). The proportion of articles with misquotations drops to 14.9% if perfect accuracy is defined as either $\mathcal{F}_1$ or $\mathcal{F}_2$ having a perfect score of 100. These figures are comparable to the 13.1% to 18.2% found in journalism studies of direct quotation accuracy, such as (Berry 1967; Marshall 1977), where they typically send out surveys to persons (most of whom are not politicians) mentioned in the news article and ask if they were accurately quoted.

At the quote level of my data, about 59.6% of the quotes contain some objective form of misquotation. Even with errors in the recording and matching of speeches and quotations in the data (e.g., HTML tags not always consistently used in the backend), this figure is still much lower than the 90% misquotations found through a check with tape recordings in Lehrer (1989). The main explanation for this is likely the context—where coverage of political speeches is more accurate than general. Table 7 lists concrete examples of quote accuracy score for a sample of speech-quote pairings as part of the discussion in Section 7. These examples further provide a sense of how the quote accuracy measures align with human intuitions of quotation accuracy.

3.3 Topic Distribution of Parliamentary Speeches and News Articles

Retrieving Latent Topics. I use the *Latent Dirichlet Allocation* (LDA, Blei et al. 2003) model, an unsupervised learning algorithm, to recover clusters of topics from

[20] As is customary, the computations under the hood remove punctuations and normalise all strings to lowercase (since punctuations and casings can neither be seen nor heard in a speech).

the news articles and parliament speeches. A *topic* in the information retrieval literature refers to a collection of words that occur frequently together. Three examples learned by the topic model trained on the sample parliament speeches are:[21]

1. ⟨*cpf, retirement, minimum_sum, saving, cpf_saving*⟩
2. ⟨*police, home_team, officer, crime, inquiry*⟩
3. ⟨*premium, medishield_life, medishield, insurance, insurer*⟩

Every parliament speech in the sample will have some probabilistic association to one of K topics, where K is a pre-defined total number of topics that a speech can draw from. The sum of the probabilistic topic association of each speech $\sum_K \rho_k$ must by definition be equal to 1. In practice, I set the parameter of the Dirichlet distribution α to be very low, embedding the assumption that parliament speeches are unlikely to have high associations to more than one topic. Similarly, I train a separate topic model for the news article corpus, attributing to each article a topic vector. The topic association of quotes comes from the parliament speech topic model.

No Human Input Required. One major advantage of the LDA is that it is unsupervised—no human input is needed to impose a topic structure before or after the discovery of topics. Besides saving resources on the manual classification of thousands of speeches and news articles, the classification avoids researcher-induced bias on the content of speeches and articles.

Choosing Optimal Number of Topics. A major disadvantage in practice is in choosing the number of topic clusters K—a model hyperparameter that is not optimised within the model. To deal with this, I automate evaluation using topic coherence scores that correlate well with human intuition of topic coherence (Chang

[21]These are taken from the LDA model trained on the parliamentary speech corpus on $K = 92$ topics. The words/phrases in the angular brackets are the top 5 words for topic 4, 18, and 31, with the relevance metric at 0.5. The phrases are words joint by '_', these are word that co-occur frequently and "glued" before the model is trained. The visualisation of the LDA results from the sample textual data is available at https://lsys.github.io/media-lda/ldavis.html.

et al. 2009). I train topic models with K starting from 2, increasing in steps of 2, up till 100, and then select K^* using the highest topic coherence score (favouring lower K's). The baseline results use topic controls from the $K = 92$ parliament speech topic model and the $K = 40$ news article topic model. There is some judgement involved in choosing K. In the specification checks, I show that the baseline conclusions do not change based on the choice of K.

Implementation. Pre-processing of the textual data and implementation of the LDA are done using the *SpaCy* and *gensim* libraries. In pre-processing, a simple algorithm is passed over the text looking for words that frequently co-occur, identifying phrases such as "minimum sum" in topic 1 above, where the meaning is very different if the two words are considered separately. The data appendix provides more details.

Summary Statistics. Table 1 presents the summary statistics. 1,106 of the 14,903 (7.4%) individual quote fragments come from the opposition. Opposition quotes are, on average, from politicians who are younger and have shorter political tenures. The quotes also come from a higher proportion of women. Notably, panel B shows that the opposition has shorter speeches. This might explain the perception of how the opposition seems to get less intense coverage when those perceptions do not account for differences in the base speech length. I address selection issues in the following section.

Overall, the final (unbalanced) panel data includes 14,903 quote level observations over 12 years, 4 parliaments, 3 general elections, 3 by-elections, with 3,425 newspaper articles, 204 politicians, and 5,130 parliamentary speeches.

Table 1. Summary Statistics

	Full sample			Non-opposition			Opposition politicians			Non-opposition − Opposition
	N	Mean	Std Dev.	N	Mean	Std Dev.	N	Mean	Std Dev.	ρ-value
Panel A. Main outcomes										
Quote (word count)	14,903	20.94	38.45	13,797	21.01	39.63	1106	20.05	18.23	0.13
Log of quote	14,903	2.68	0.95	13,797	2.69	0.94	1106	2.57	1.05	0.00
Substring quote accuracy measure	14,903	91.41	12.85	13,797	91.45	12.86	1106	90.80	12.71	0.10
Bag-of-words quote accuracy measure	14,903	96.40	10.52	13,797	96.50	10.26	1106	95.11	13.26	0.00
Panel B. Length controls (word count) for textual data										
Paragraph (of speech)	14,903	103.50	60.79	13,797	104.23	61.69	1106	94.37	47.23	0.00
Speech	14,903	2112.39	1824.96	13,797	2188.47	1858.28	1106	1163.35	909.74	0.00
Article	14,903	625.36	263.42	13,797	619.65	255.48	1106	696.60	339.70	0.00
Panel C. Other control variables										
Age	14,887	51.47	6.96	13,781	51.69	6.70	1106	48.74	9.23	0.00
Age^2	14,887	2697.39	702.22	13,781	2716.36	676.53	1106	2460.97	935.37	0.00
Tenure	14,903	10.97	7.55	13,797	11.11	7.48	1106	9.18	8.10	0.00
$Tenure^2$	14,903	177.22	207.00	13,797	179.42	208.02	1106	149.77	191.88	0.00
Female	14,903	0.17	0.37	13,797	0.16	0.36	1106	0.28	0.45	0.00
Rank	14,903	5.39	3.44	13,797	5.00	3.27	1106	10.28	0.45	0.00
Translations	14,903	0.01	0.12	13,797	0.01	0.11	1106	0.04	0.19	0.00
Group size	14,428	4.77	1.26	13,322	4.87	1.14	1106	3.58	1.89	0.00
Voters	14,428	126,159.37	38,920.13	13,322	128,444.28	36,166.56	1106	98,637.13	56,458.30	0.00
Vote	11,206	74,347.06	28,957.89	10,100	77,327.63	27,531.60	1106	47,128.50	27,440.08	0.00
Vote (%)	11,206	64.42	7.92	10,100	65.73	6.88	1106	52.42	6.60	0.00
Margin	10,900	35,389.49	22,908.53	10,100	37,525.17	22,418.44	800	8426.44	4545.91	0.00
Margin (%)	10,900	30.02	14.37	10,100	31.48	13.78	800	11.52	7.07	0.00

4 Article-Speech Level Results

4.1 Empirical Strategy

The first level of analysis is at the article-speech level, where all quotes originating from a speech s reported in article r are taken as a single observation. For example, if an article contains two quotes (fragments) from the same speech, these two quotes are concatenated and taken as one observation. The combined length of these two quotes is the first (of three) outcome measure—the intensity of coverage. For the second outcome measure—the *count* of quote fragments from a speech—this is recorded as *2*. The third is the bag-of-word quote accuracy measure based on the common set of words between the concatenation of the two quotes and their originating speech.[22] At this level, there are 204 politicians, 3,425 articles, and 5,227 speeches over 12 years, for a total of 7,098 politician-article-speech level observations, of which 640 are from the opposition. Section A.1 provides some background to Parliamentary procedures and speeches. Unfortunately, the data does not distinguish between types of procedures.

To compare the above three coverage measures between the opposition and the ruling party, I estimate the OLS model:

$$y_{irst} = \alpha + \beta\, opp_i + \sum_{k=2006}^{2016} \alpha_k\, year_{kt} + \sum_{\ell=11}^{13} \alpha_\ell\, parl_{\ell t} + \boldsymbol{\gamma}' \boldsymbol{X}_{irst} + \varepsilon_{irst}, \tag{3}$$

where y_{irst} is one of the above three measures of media coverage for politician i's speech s at time t, reported in article r. The estimand of interest is β, the coefficient of the opposition dummy variable that takes on value 1 if politician i is from *any* opposition party. Identification issues notwithstanding, a negative β estimate suggest a systematic political media slant towards ruling-party politicians.

β in equation (3) unfortunately does not carry an unambiguous causal interpreta-

[22] At the article-speech level, the quotation accuracy is as defined by the bag-of-word quote accuracy measure which is computed based on words common to both speech and quote(s), because using best partial string match of a concatenation of multiple quotes is meaningless.

Table 2. Description of Control Sets

A. Individual controls	
Gender	Gender of politician i
Race	Race of politician $i \in$ {Chinese, Indian, Malay, Eurasian/Others}
Age	Age of politician i at time t of speech s given
Tenure	Political tenure of politician i at time t of speech s given
B. Article controls	
Day of week	Dummies for article r publication day of week $\in$ {Mon, Tue, Wed, Thu, Fri, Sat, Sun}
Section	Dummies for article r's section $\in$ {*Singapore, Prime News, Top of the News, Home, ST, Insight, News, Money, Think, Review - insight, Sports, Opinion, World, Others* }
Translation	Dummy for a translation from vernacular to English in speech s
C. Topic distributions	
Speech topics	Topic distribution for speech s from model trained on speeches for $K = 92$
Quote topics	Topic distribution for quote q from model trained on speeches for $K = 92$
Article topics	Topic distributions for article r from model trained on articles for $K = 40$
D. Ministerial controls	
Type	Ministerial type of politician i at time t of speech s given
Portfolio	Ministerial portfolio of politician i at time t of speech s given
E. Electoral controls	
Group size	Number of politicians representing politician i's constituency $\in$ {1, 4, 5, 6}
Voters	Number of eligible voters in politician i's constituency
Votes	Number of votes for politician i's constituency
Votes (%)	Percentage of votes for politician i's constituency
Margin	Number of winner's vote – number of loser's votes in the politician i's constituency
Margin (%)	Ratio of *majority* to the number of valid votes

Notes—Sections appearing no more than 6 times in the sample are collapsed under *Others*. Table A.5 shows the distribution of sections. In the rare instances where a politician holds two different ranks for two different ministry portfolio, the higher of the two ranks is recorded. Portfolios under non-extant ministeries are mapped to their modern day equivalents. For instance, the Ministry of Information, Communications and the Arts (MICA) portfolio is mapped to the current day Ministry of Communications and Information (MCI) portfolio.

tion since a candidate's choice of party is unlikely to be orthogonal to potential media coverage. Nevertheless, I control for a set of covariates that likely determines how the media covers politician speeches and that are also correlated to partisanship, including controls that originate from the richer textual data (Table 2).

First, equation (3) includes year and parliament fixed effects. If there are trends in media consumption or newsroom operations, the year fixed-effect will capture them. Quotes, for instance, exhibit a small but gradual rise in accuracy over time (Figure 1). Parliament fixed effects capture potential trends in political sentiments and changes in parliament composition. The 12th parliament, for instance, had an increase in women representation (even in the absence of binding gender quotas) and an increase in opposition representation.

β in equation (3) above is not identified with individual fixed-effects since no candidate in the sample switched from the opposition to the ruling party (and vice versa). To mitigate concerns about individual characteristics driving observed

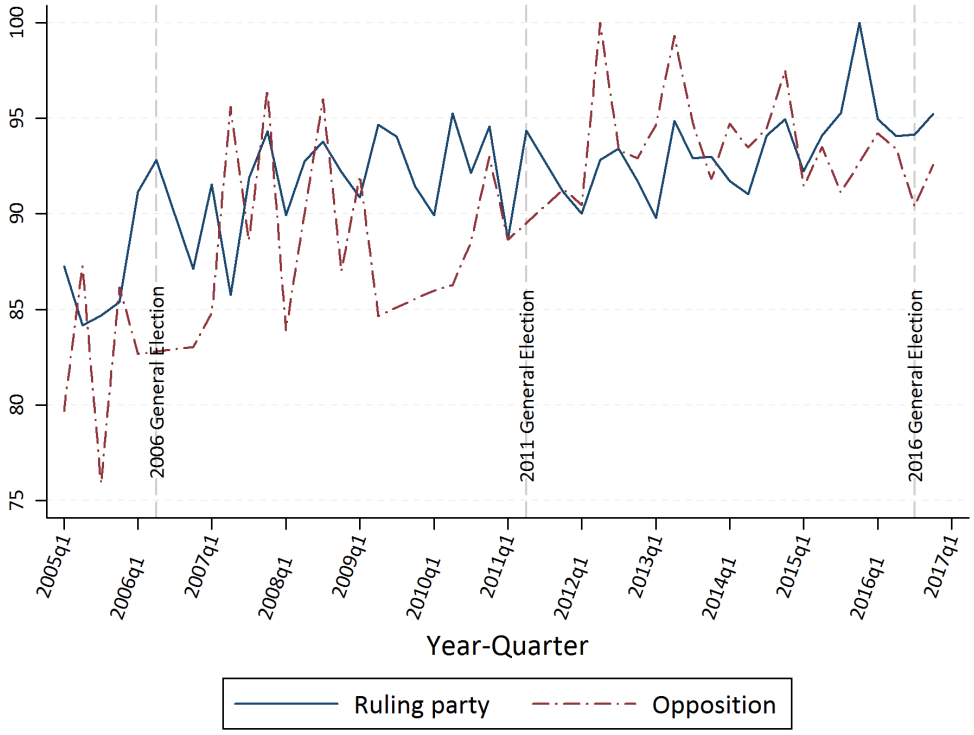


Figure 1. Substring Accuracy Measure of Parliamentary Speeches

differences between opposition, I include the politicians' gender, race, and a quadratic for age and political tenure. $\boldsymbol{X}_{it}$ also includes the full interaction of politician-type (e.g., minister or parliamentary secretary) and ministry portfolio of politicians (e.g., health or education). Certain politicians, for instance, may get more coverage bandwidth because of seniority. The portfolio accounts for some of this effect. These are all recorded as of the date of the speech given.

For the set of news article controls, there are publication day-of-the-week dummies for variations in newsroom operation by day. News section dummies control for the section in which a news article appears. A translation dummy controls for quotes translated from a vernacular to English (as stated in the transcripts).[23] Standard errors are adjusted to allow for clusters within each newspaper article r.[24]

Using Text Data to Deal with Alternative Interpretations. Here I consider the

[23]Speeches of ruling-party politicians, for example, may involve more discussions on geopolitical entities and affairs, and these are usually reported in certain newspaper sections (e.g., *World*). Table A.5 shows the distribution of newspaper sections by partisanship. Opposition politicians, for instance, never appear in *World*, *Money*, or *Opinion* sections in the sample.

[24]In the robustness checks in Sections 4.3 and 5.2 also allows for standard errors to be clustered by parliamentary speech or the reporting journalist.

"*political speech content*" interpretation, "*trivial words*" interpretation, and the "*speech coherence*" interpretation.

One concern is related to partisan ownership of political topics (in the sense of Petrocik 1996; Puglisi 2011). If $\hat{\beta}_{OLS}$ is negative, suggesting that opposition coverage is less favourable, perhaps this arises because of differences in the political focus of the parties. For example, the ruling party may speak more about political issues of greater interest or immediate relevance. Even with quotation accuracy as an outcome measure, perhaps accuracy suffers because journalists take mental breaks during (opposition) less interesting speeches and not because of party status.

To deal with this concern, I control for the topics of the speeches and news articles using the output from the LDA (Section 3.3). Concrete examples of topics learned from the parliamentary speech corpus, represented by the top five words, are:

1. ⟨*cpf, retirement, minimum_sum, saving, cpf_saving*⟩
2. ⟨*police, home_team, officer, crime, inquiry*⟩
3. ⟨*premium, medishield_life, medishield, insurance, insurer*⟩
4. ⟨*student, school, learn, education, teacher*⟩
5. ⟨*fare, bus, public_transport, commuter, operator*⟩
6. ⟨*sport, athlete, community, youth, sports*⟩

The baseline results always include the topic distributions. While the output from the LDA says nothing about partisan ownership of topics, it is sufficient to remove variations in coverage that are correlated with the content of speeches and news articles. This mitigates concerns that differences in the political speech content are driving the results.[25]

The second concern has to do with the usage of words in the opposition speeches. Perhaps the opposition uses more trivial words in their speeches, and journalists can ignore them for efficiency or clarity while retaining the core meaning of the

[25]The clustering of the topic distribution of the speeches and news article is apolitical, so although a natural tendency is to see if there are additional partisan differences in contentious political topics, the output, so the LDA says nothing about which topics are contentious and which are not.

speech. Mechanically, this would imply lower accuracy scores even if the journalist had changed the words for the sake of clarity. To mitigate such concerns, I perform robustness tests (in Sections 4.3 and 5.2) using alternative constructions of the accuracy scores by removing stop words in the pre-processing stage. The baseline conclusions are unaffected.[26]

A third concern is language competency. For example, it may be the case that the speeches of the opposition are less coherent than those of the ruling party. Hence journalists have greater difficulty following and quoting opposition speeches accurately. To address this concern (in Section 6), I tap on the quantitative linguistics literature that estimates English language level based on the distribution of words in a text. Controlling for readability of the speech transcript and language sophistication using lexical richness measures barely attenuates the baseline estimates.

Bounds Using Differences in Ministerial Composition. To deal with any residual bias in the OLS estimates, I turn to two bounding arguments. First is a compositional bias that arises even if party assignment is random. The compositional bias occurs because the opposition status limits the ministerial profile. Opposition politicians, for instance, never get a ministerial rank that is higher than the base rank. In addition, opposition politicians hold no ministerial portfolios (e.g., Education or Health).[27] Hence party status determines both media coverage and ministerial characteristics.

Let the opposition dummy be $D_i \in \{0, 1\}$, and a rank dummy be $r_i \in \{0, 1\}$, and abstract away from the other covariates $\boldsymbol{X}_{it}$. The observed difference in coverage between an opposition politician ($D = 1$) and a ruling-party politician ($D = 0$), with

[26] It may be worth noting that direct quotations should be verbatim even in the presence of errors or language inefficiency, so usage of stop words should have been preserved anyway.

[27] Unlike the typical Westminster system, there is no shadow cabinet.

both of equal base rank ($r = 0$), can then be decomposed as:[28]

$$\mathbb{E}[y_{1i} - y_{0i}|r_{1i} = 0] + \mathbb{E}[y_{0i}|r_{1i} = 0] - \mathbb{E}[y_{0i}|r_{0i} = 0], \tag{4}$$

where the first term is the true causal effect (or slant) as the difference between the potential coverage of a candidate assigned to the opposition party and the ruling party, conditional on potential assignment to the base rank as an opposition.

The compositional bias comes from the second and third terms, and institution-specific factors help sign the bias as follows. The second term is the potential coverage of a ruling-party politician had he been given assignment to the base rank ($r = 0$) as an opposition. Since this assignment would apply to all ruling-party politicians (the opposition only ever gets base rank), this is simply the unconditional average coverage of ruling-party politicians—$\mathbb{E}[y_{0i}|r_{1i} = 0] = \mathbb{E}[y_{0i}]$. The third term is the potential coverage of a ruling-party politician conditional on assignment to the base rank ($r = 0$), and this I argue implies an *inferior* candidate, which means the conditional expectation is below the unconditional one—$\mathbb{E}[y_{0i}|r_{0i} = 0] < \mathbb{E}[y_{0i}]$. Taken together, this decomposition in equation (4) implies an *attenuation bias* if:

(i) the true effect (first term) is negative, $\mathbb{E}[y_{1i} - y_{0i}|r_{1i} = 0] < 0$, which is the case if there is media slant favouring the ruling party;

(ii) the second term is the unconditional expectation of potential coverage of a ruling-party politician, $\mathbb{E}[y_{0i}|r_{1i} = 0] = \mathbb{E}[y_{0i}]$, since assignment to base rank conditional on opposition assignment applies for *all* ruling-party politicians; and

(iii) the third term is lower than the unconditional expectation of potential coverage of a ruling-party politician, $\mathbb{E}[y_{0i}|r_{0i} = 0] < \mathbb{E}[y_{0i}]$, if on average, inferior ruling-party candidates are less likely to take on senior positions, and coverage is increasing in quality of candidate. Hence the second and third terms are

[28]The appendix in Section A.2 provides mode details of this decomposition based on bad controls (Angrist and Pischke 2009).

$\mathbb{E}[y_{0i}] - \mathbb{E}[y_{0i}|r_{0i} = 0] > 0$, which goes in the opposite direction of the true effect.

Bounds Using Proportional Selection. If party assignment is not conditionally random, one can interpret a negative $\hat{\beta}_{OLS}$ as a selection bias, where for whatever institutional reason, candidates that are more competent or hold themselves better in public select into the ruling party, leaving the opposition with an inferior set of candidates. If this is the case, then opposition coverage would naturally be less favourable if coverage is increasing in the calibre of the politicians.[29]

Since calibre is unobserved, amongst other factors, the second bounding argument uses the proportional selection assumption (Altonji et al. 2005; Oster 2017), which provides a formal bound on OLS estimates based on how the movement in OLS estimates when additional observables are added is informative about selection on unobservables, *after* normalising for corresponding movements in the variation explained. The formal bound developed in Oster (2017), assuming equal selection on both observables and unobservables, is:

$$\beta^* = \tilde{\beta} - \left[\overset{o}{\beta} - \tilde{\beta}\right] \frac{R_{max} - \tilde{R}}{\tilde{R} - \overset{o}{R}}, \tag{5}$$

where $\tilde{\beta}$ is the OLS estimate of β with all controls included, and $\tilde{R}$ is the corresponding R^2. $\overset{o}{\beta}$ and $\overset{o}{R}$ corresponds to the OLS results without controls. $R_{max} = 1.3$ as recommended in Oster (2017).[30] I show below in Section 5.3 that the bounds computed using equation (5) has a larger magnitude than the OLS estimates ($\beta^* < \tilde{\beta} < 0$), reinforcing the argument of an attenuation bias even if party selection is not random.

4.2 Baseline Results, Article-Speech Level

Table 3 presents the baseline results at the politician-article-speech level, where the three panels present results for the three outcome variables. Each panel shows

[29] Tan (2014) provides some background on the institution-specific features which confers the ruling party of Singapore an advantage in selecting candidates.

[30] A larger R_{max} only further reinforces the findings.

Table 3. Baseline Results for Political Coverage, Article-Speech Level

	A. Log of quote length by word count		B. Count of quote fragments		C. Bag-of-words accuracy measure	
	(1)	(2)	(3)	(4)	(5)	(6)
Opposition	0.052	0.038	0.328***	0.318***	–2.499***	–2.554***
	(0.064)	(0.074)	(0.089)	(0.109)	(0.785)	(0.832)
Controls						
Time fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Article controls	Yes	Yes	Yes	Yes	Yes	Yes
Topic controls	Yes	Yes	Yes	Yes	Yes	Yes
Ministerial controls	Yes	Yes	Yes	Yes	Yes	Yes
Electoral controls	—	Yes	—	Yes	—	Yes
F-statistics						
F-statistic, year fixed-effects	2.682***	1.613*	1.753**	2.127**	4.589***	2.033**
F-statistic, individual controls	0.397	1.153	1.322	2.775***	2.490**	1.855*
F-statistic, article controls	3.290***	2.538***	5.370***	4.487***	3.189***	2.248***
F-statistic, topic controls	2.762***	2.413***	2.314***	2.164***	1.563***	1.440***
F-statistic, ministerial controls	2.694***	2.722***	5.106***	4.987***	2.360***	1.826***
F-statistic, electoral controls		0.382		1.660		1.476
Mean of dependent variable	3.306	3.301	2.101	2.119	96.326	96.587
R^2	0.181	0.199	0.241	0.256	0.118	0.139
N	7,087	5,143	7,087	5,143	7,087	5,143

Notes—Observations are at the article-speech level, where separate (direct) quotations that originate from the same speech and reported in the same news article are treated as a single observation. Time fixed-effects include both the parliament (10th, 11th, 12th, & 13th) and year (2005–16) fixed effects. Individual controls include the politicians' gender, ethnic (Chinese, Indian, Malay, and Eurasian/Others), and a quadratic in age and political tenure. Article controls include day-of-the-week dummies, newspaper section (e.g. Top of the News, Prime News), and a dummy for whether the quote was translated. Topic controls are vectors of probabilistic association (sum to one) of the newspaper article and parliament speech to topics uncovered using LDA. Ministerial controls include politician type (e.g. Deputy Prime Minister, Parliamentary Secretary) and political portfolio (e.g. Health, Education). Electoral controls include the electorate size that the politician represents, the group size of representation (group representation have sizes of 4–6), vote share, and winning margin in the most recent general election. Politicians who had won by default (no opposition contest) in the most recent election have no electoral data. All regressions also include the intensity measures for the textual data—the (log) length of speech, speech paragraph, and news article. F-statistics report the test statistic for the null that the set of controls are jointly equal to zero. Robust standard errors in parentheses are clustered at news articles.

*** Significant at the 1 per cent level.

** Significant at the 5 per cent level.

* Significant at the 10 per cent level.

two sets of results; without and with electoral controls.

The results suggest some slant, with indications of higher fragmentation and lower quotation accuracy for the opposition politicians. In column (1), conditional on the time fixed-effects (both year and parliament), speech and article lengths, and the individual, ministerial, article, and topic controls, the opposition politicians get quotes that are 5 log points longer, suggesting that the opposition politicians get more coverage. However, this is not statistically significant at the 10 percent level. In column (3), opposition politicians get about a third more quote fragment from each article-speech observation (about 18% of the standard deviation), significant at the 1% level.

In the third panel, the outcome variable is the bag-of-words accuracy measure.

This measure quantifies accuracy based on the subset of words common to both speech and quote. From the result in column (5), opposition politicians get quoted approximately 2.5 points less accurately than ruling-party politicians at the 1 percent significance level (about 26 percent of the standard deviation). The estimates barely change with electoral controls (even-numbered columns).

At the article-speech level, where the concatenation of all quotes from a speech reported in a news article is the unit of observation, there is no evidence that the opposition gets quoted at a lower intensity than the ruling party from their speeches in parliament. This finding does not square with those perceptions of slant where the ruling party gets more coverage.[31] The opposition speeches, however, are quoted using more fragments, and their quotes are less accurate relative to those of the ruling party, providing the first sign that their coverage is more fragmented and less accurate. I discuss the magnitudes of the OLS estimates and implications in section 7 after the quote level results.

4.3 Specification Checks, Article-Speech Level

Table A.6 presents specification checks for the article-speech level result where the opposition gets more quote fragments. Column (1) reproduces the baseline result for comparison (column (3) of Table 3). Column (2) models quote fragments as an over-dispersed count variable using the negative binomial regression, and the result is similar. The baseline result where the opposition gets more fragments of quote is also robust to the inclusion of journalist fixed-effects and beat assignment dummies in column (3), excluding translations in column (5), and adjusting standard errors for clusters within speeches and journalists in columns (6) and (7).

[31] In Figures A.3-A.4, eyeballing the distribution of coverage intensity by year and parliament suggest that intensity for the ruling party is proportional to the share of seats held in parliament. Or, that the ruling party gets more coverage simply because they have more share of seats in parliament. The results at the politician-year level (untabulated) are also consistent with the finding that there is no coverage intensity difference. The opposition politicians get featured in the same number of articles (counted as the number of articles containing at least one quote from parliament) per year compared to the ruling-party politicians.

In columns (8)–(13), the baseline results also survive different assumptions on the topic structure of the textual data. Columns (6) and (7) use the $K = 50$ and $K = 100$ speech topic model. Columns (8) and (9) use the $K = 30$ and $K = 50$ article topic model for article distributions. Column (10) uses the topic distribution of the entire sentence containing the quote, instead of just the quote. Column (11) uses the most parsimonious topic distribution specification available in the sample—the $K = 50$ speech topic model and the $K = 30$ news article topic model.

An exception, however, is in column (4). I exclude ministerial controls on concerns that ministerial rank itself is an outcome of opposition status (as a case of bad controls Angrist and Pischke 2009), and the opposition status is no longer significant. I do not fully understand how to interpret this result—including ministerial controls introduces a compositional bias, and excluding them loses non-trivial heterogeneity among the politicians.

Table A.7 tests the sensitivity of the baseline finding in column (5) of Table 3 where opposition quotes are less accurate, using the same set of robustness test as before (but without the count model). The outcome variable in panel A is constructed in the same way as in Table 3. Panel B uses the alternative construction of quote accuracy where stop words are removed in pre-processing. The results from both panels are robust and never fall below the 1 percent level of significance.

Panel B in particular mitigates concerns that the opposition's lower quotation accuracy can be narrowed down to a quote's trivial contents—that journalists are less careful only with the usage of stop words when quoting the opposition while treating the non-stop words (or the substantial words) the same way for all politicians.

Table 4. Baseline Results for Political Coverage, Quote Level

	A. Log of quote length by word count		B. Substring accuracy measure		C. Bag-of-words accuracy measure	
	(1)	(2)	(3)	(4)	(5)	(6)
Opposition	−0.138***	−0.145***	−1.455**	−1.485**	−2.434***	−2.271***
	(0.047)	(0.055)	(0.701)	(0.753)	(0.707)	(0.675)
Controls						
Time fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
News article controls	Yes	Yes	Yes	Yes	Yes	Yes
Topic controls	Yes	Yes	Yes	Yes	Yes	Yes
Ministerial controls	Yes	Yes	Yes	Yes	Yes	Yes
Electoral controls	—	Yes	—	Yes	—	Yes
F-statistics						
F-statistic, year fixed-effects	3.090***	2.050**	5.229***	3.189***	5.125***	2.501***
F-statistic, individual controls	1.104	1.379	2.027**	0.902	2.055**	0.742
F-statistic, topic controls	2.412***	2.290***	2.106***	2.078***	1.608***	1.459***
F-statistic, ministerial controls	3.901***	3.997***	3.904***	3.964***	2.865***	2.180***
F-statistic, electoral controls		1.096		0.990		1.264
Mean of dependent variable	2.680	2.672	91.405	91.863	96.399	96.647
R^2	0.056	0.068	0.171	0.197	0.113	0.124
N	14,887	10,900	14,887	10,900	14,887	10,900

Notes—Observations are at the quote level; the regressions in this Table considers each quote as separate observations, even if they originate from the same speech or are reported in the same news article. The set of controls are the same as in article-speech results in Table 3. Robust standard errors in parentheses are clustered at news articles.

*** Significant at the 1 per cent level.

** Significant at the 5 per cent level.

* Significant at the 10 per cent level.

5 Quote level Results

5.1 Baseline Results, Quote Level

Here, the unit of analysis is individual quote fragments. At this level, there are 204 politicians, 3,425 articles, and 5,227 speeches over 12 years, for a total of 14,887 quote fragments (with complete controls) of which, 1,106 are from opposition speeches. The model I estimate is:

$$y_{iqrst} = \alpha + \beta\, opp_i + \sum_{k=2006}^{2016} \alpha_k\, year_{kt} + \sum_{\ell=11}^{13} \alpha_\ell\, parl_{\ell t} + \boldsymbol{\gamma}' \boldsymbol{X}_{iqrst} + \varepsilon_{iqrst}, \tag{6}$$

which is synonymous with (3) at the article-speech level, but with an additional indexing of individual quote fragments q, the smallest unit of analysis here. The outcome variable y is alternatively: (i) log of individual quote length by word count, (ii) substring accuracy measure, and (iii) bag-of-words accuracy measure. Standard errors are adjusted to allow for clusters within each news article.

Table 4 reports the results. The first panel confirms the finding at the article-speech level where the coverage of the opposition is more fragmented. Conditional on the time trends and observables, column (1) indicates that the opposition politicians get shorter individual quote fragments. On average, the quotes of opposition politicians are approximately 13% shorter than quotes of ruling-party politicians. Taken together with the article-speech level results, the evidence suggests that the quotes of the opposition are more fragmented—opposition politicians are covered at the same intensity from a speech in an article as their ruling-party counterparts, but the coverage comes from more but shorter quote fragments.

The second and third panels are the results for quotation accuracy. In the second panel, the outcome variable is the substring accuracy measure, which reflects exactly the verbatim nature of direct quotations. Like the bag-of-words measure, it ranges from 0 to a perfect 100. In the third panel, the outcome variable is the bag-of-words accuracy measure, which scores accuracy using the subset of words common to both speech and quote. Compared to the substring accuracy measure, the bag-of-words measure allows for edits of quotes for syntactic flow and grammatical demands. Consistent with the article-speech level results, opposition politicians are quoted less accurately by both measures at the quote level. On average, opposition quotes have approximately 1.5 to 2.4 points lower accurate than quotes of the ruling party (11.6% to 22.9% of the standard deviations in accuracy).[32]

5.2 Specification Checks

Table A.8 presents the specification checks for the results on log of individual quote length (along the same lines as Table A.6 in Section 4.3). Panel B, C, and D use

[32]Including electoral controls can result in either an upward or downward bias. If walkovers typically involve ruling-party politicians sufficiently established to ward off opposition challenges, then excluding these politicians would induce a bias towards zero if these politicians get better coverage because of their prominence. On the other hand, if walkovers are associated with politicians with lower visibility—since they do not need to campaign as much—then leaving them out would artificially accentuate the observed difference between the opposition and the ruling party. The results in columns (2) and (4) of Table 4 are consistent with the latter explanation, though the differences are small in magnitude, which can be attributed to sampling variation.

alternative measures of quote length: (B) by character count, (C) by word count without stop words, and (D) by character count without stop words. The results are mostly robust, with journalist fixed-effects removing most of the observed opposition effect only when stop words are removed.

Table A.9 presents a similar set of specification tests for the results on quote accuracy, which remain statistically significant. Unlike the result for quote length above, including journalists and beat fixed-effects does not move the estimates for quote accuracy towards zero, suggesting that the tendency to quote ruling-party politicians more accurately is more institutional than just the discretion of individual journalists.

5.3 Bounds on OLS Estimates

In Section 4.1 above, I argue that bias in the OLS estimates are attenuating ones, implying that the OLS estimates are conservative. First, an attenuation bias arises when the individual politicians' ministerial controls are included, even if party status is random (as discussed in Section 4.1). Where a selection issue persists, a second bounding argument can be made using the proportional selection assumption—the assumption that the proportion of selection on observables is equal to the selection on unobservables. This assumption implies that all unobserved factors, which affects political coverage and party status, are equal in importance to all the observables available on hand, including the length of speeches and articles, individual and ministerial characteristics, the topics of the speeches and articles, and additional language-based measures (detailed in the following section).

Assuming equal selection between observables and unobservables implies that the treatment effect of opposition status on quote accuracy (or slant) is -2.18 and -3.39, compared with the OLS regression estimates of -1.46 and -2.24 from Table 4. Moreover, for the estimated treatment effect of the opposition status to be zero, the degree of selection on unobservables relative to observables must be high—the

Table 5. Political Coverage of Backbenchers, Quote Level

	A. Log of quote length by word count		B. Substring accuracy measure		C. Bag-of-words accuracy measure	
	(1)	(2)	(3)	(4)	(5)	(6)
Opposition	−0.116**	−0.164**	−1.242	−2.195***	−2.605***	−2.680***
	(0.053)	(0.064)	(0.778)	(0.820)	(0.781)	(0.752)
Time fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Article controls	Yes	Yes	Yes	Yes	Yes	Yes
Topic controls	Yes	Yes	Yes	Yes	Yes	Yes
Electoral controls	No	Yes	No	Yes	No	Yes
F-statistic, time fixed-effects	1.623*	1.262	3.119***	2.268***	2.287***	1.951**
F-statistic, individual controls	1.641	1.872*	3.964***	3.450***	2.241**	0.788
F-statistic, topic controls	2.986***	3.635***	2.252***	2.771***	1.252**	1.476***
F-statistic, electoral controls		0.647		1.823		3.373***
Mean of dependent variable	2.681	2.661	91.553	91.785	96.415	96.533
R^2	0.130	0.161	0.236	0.281	0.190	0.251
N	3,882	3,091	3,882	3,091	3,882	3,091

Notes—Observations are at the quote level, but only for politicians who are backbenchers; the regressions in this Table considers each quote as separate observations, even if they originate from the same speech or are reported in the same news article. The set of controls are the same as in article-speech results in Table 3. Robust standard errors in parentheses are clustered at news articles.

*** Significant at the 1 per cent level.

** Significant at the 5 per cent level.

* Significant at the 10 per cent level.

unobservables must be 58 times and 16 times more important than the observables. Hence, the estimates using equal selection on observables and unobservables further supports the OLS estimates as conservative on the extent to which the opposition politicians receive less accurate coverage.

Nonetheless, to simplify the analysis and directly compare politicians of both parties that belong to the base rank, I repeat the regression analysis only for the backbenchers. Table 5 reports the results. The point estimates vary compared to the baseline regressions, but the sign and magnitudes are largely similar. Focusing on the substring accuracy measure and with electoral controls included in column (4), the point estimate is even larger, and to this extent is consistent with the bounding analyses discussed above.

More generally, Figure 2 plots the point estimates for the quote-level results with substring accuracy as the outcome, from 144 different specifications (= 2 substring accuracy measures, with and without substrings × 8 covariates combinations × 9 possible topic modelling specifications). The estimates are ranked in ascending order, with the chart indicating the specific combinations. Combinations under the

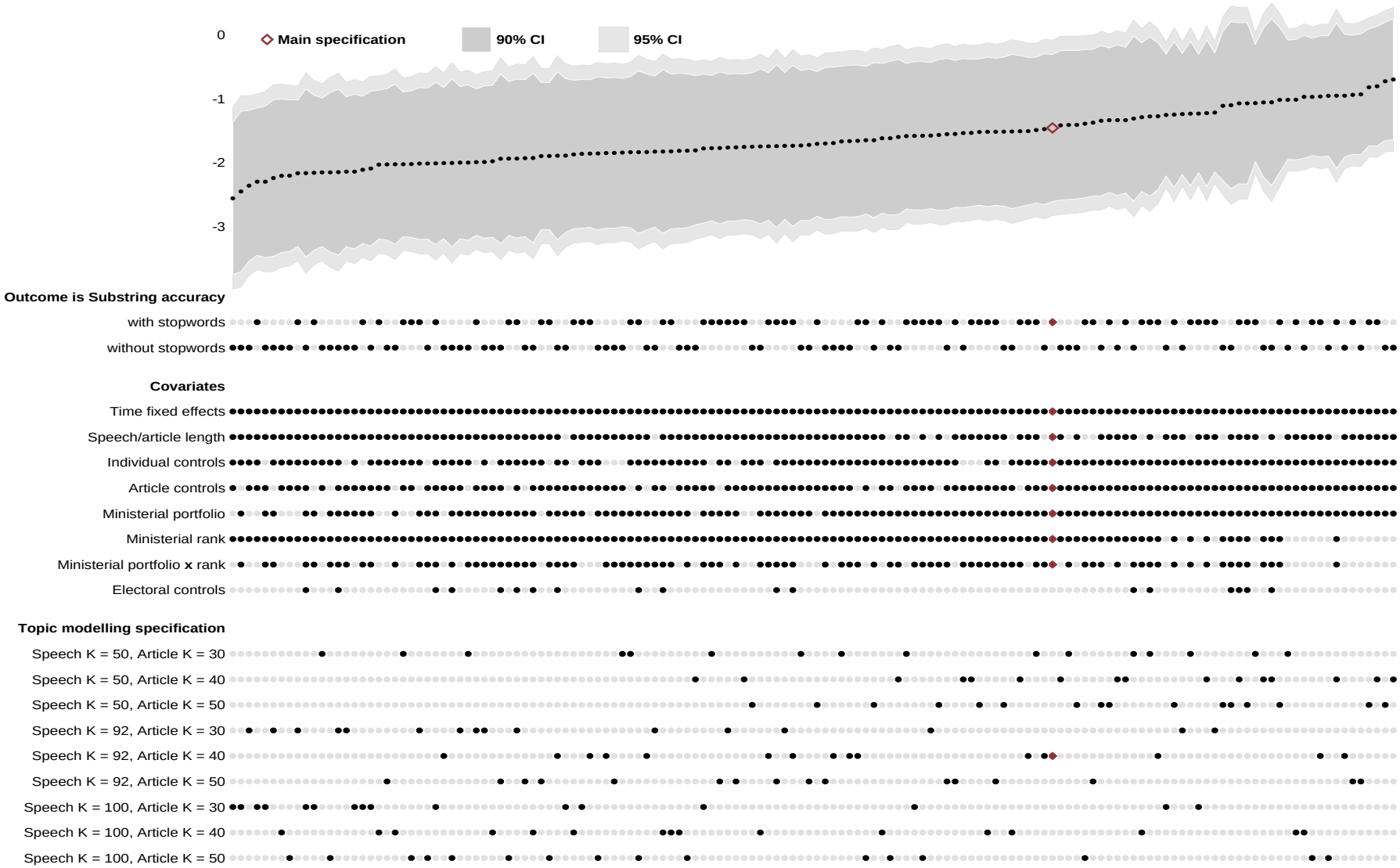


Figure 2. Effect Sizes of Opposition Status on Substring Accuracy

outcome and topic models are mutually exclusive, but those under the covariates are not. Certain specifications have weaker statistical significance, but all have the same signs and approximately the same magnitude. Figure A.2 shows the same chart with bag-of-words accuracy as the outcome measure.

6 Additional Text and Language-Based Measures

6.1 Speech Tone and Language Competency Controls

One concern with the baseline results is that the opposition speeches are less coherent than those of their ruling-party counterparts. A related concern is that the media prefers reporting on speeches of a particular tone (e.g., negative sounding or subjective sounding). To rule these out, I test whether speech: (i) objectivity, (ii) polarity, (iii) readability,[33] and (iv) lexical richness can pick up the differences in media coverage between politicians. If speech competency and tone drive the results, then the opposition estimate should attenuate to zero once speech tone and competency are controlled for. This is not the case.[34]

To generate objectivity (objective vs. subjective) and polarity (positive vs. negative) measures for speeches, I use the *Pattern* sentiment analyser implemented in the *TextBlob* library, which uses part-of-speech tagging so that words in different parts of a speech get different weights. The readability measures are weighted averages of three different pieces of textual information: (i) average word per sentence, (ii) average syllable per word, and (iii) the fraction of text made up of polysyllabic (three or more syllables) words—with readability decreasing in each of them. Lexical richness measures proxy for language sophistication using information on the occurrences of unique words—the more unique words used, the higher the language

[33] *Readability* in the literature refers to how easy it is to read a piece of text, the context in this paper deals with speeches, and so *understandibility* (of a speech) might be a better word. However, I retain the use of the term *readability* since it is a fairly established term.

[34] This section uses data at the quote level. I perform the same set of tests at the article-speech level, and the findings are qualitatively the same.

Table 6. Additional Language-based Measures as Controls

		Additional language and textual controls using 1st principal components of objectivity, polarity, readability, and lexical richness measures				
	Baseline results	Objectivity of textual content	Polarity of textual content	English grade/ readability	Lexical Richness	All
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A. Dep. Var. is Log of quote length by word count					
Opposition	−0.138***	−0.121***	−0.132***	−0.136***	−0.135***	−0.113**
	(0.047)	(0.047)	(0.047)	(0.047)	(0.048)	(0.047)
Objectivity of speech and quote		−0.059***				−0.055***
		(0.005)				(0.005)
Polarity of speech and quote			0.030***			0.017***
			(0.005)			(0.005)
Grade/*readability* score of speech transcript				−0.005		−0.005
				(0.006)		(0.006)
Lexical richness of speech transcript					−0.001	−0.001
					(0.005)	(0.005)
	Panel B. Dep. Var. is substring quote accuracy measure					
Opposition	−1.455**	−1.452**	−1.462**	−1.608**	−1.272*	−1.448**
	(0.701)	(0.701)	(0.701)	(0.702)	(0.703)	(0.704)
Objectivity of speech and quote		0.030				0.021
		(0.064)				(0.065)
Polarity of speech and quote			−0.077			−0.064
			(0.061)			(0.062)
Grade/*readability* score of speech transcript				0.380***		0.355***
				(0.087)		(0.089)
Lexical richness of speech transcript					0.473***	0.451***
					(0.074)	(0.076)
	Panel C. Dep. Var. is bag-of-words quote accuracy measure					
Opposition	−2.434***	−2.411***	−2.425***	−2.443***	−2.379***	−2.352***
	(0.707)	(0.709)	(0.707)	(0.705)	(0.703)	(0.704)
Objectivity of speech and quote		−0.083				−0.074
		(0.052)				(0.054)
Polarity of speech and quote			0.048			0.030
			(0.049)			(0.051)
Grade/*readability* score of speech transcript				0.021		0.003
				(0.087)		(0.091)
Lexical richness of speech transcript					0.206***	0.205**
					(0.078)	(0.080)
N	14,887	14,885	14,885	14,887	14,836	14,834

Notes—Observations are at the quote level. The regressions in this Table are the same as in the baseline specification in Table 4 (replicated in column (1)), but with the additional language measures. Readability and lexical richness are concepts with various proposed measures in practice. The readability and lexical richness measures in this Table are the first principal component of the relevant measures from principal component analyses. The objectivity and polarity measures are first principal components of the different textual components (e.g. speech sentence, speech paragraph). Robust standard errors adjusted for clusters by newspaper article in parentheses.

*** Significant at the 1 per cent level.

** Significant at the 5 per cent level.

* Significant at the 10 per cent level.

sophistication.[35] [36]

[35]The readability measures are computed using the *textatistic* library at `http://www.erinhengel.com/software/textatistic`, while the lexical richness measures are computed using the *lexicalrichness* library I wrote and is hosted at `https://github.com/LSYS/lexicalrichness` (Shen 2022).

[36]Importantly, the measures of lexical richness are computed using measures robust to changes in text length (e.g. Maas, mean segmental type-token ratio (MSTTR), Measure of Textual Lexical Diversity (MTLD), and HD-D (McCarthy and Jarvis 2010; Torruella and Capsada 2013)), since the speeches have varying lengths. The data appendix provides further details.

Table 6 replicates the baseline models at the quote level with the additional text and language controls. In panel A, objectivity and polarity explain some variation in quote length, with more subjective and positive-sounding speeches getting more coverage. The opposition coefficient for quote length, however, remains insignificant. Lexical richness also explains some variation in quote accuracy (panels B and C), with accuracy increasing in the lexical richness of a speech. The baseline conclusion remains otherwise unchanged with all four of the additional language controls, providing evidence against the concern that the opposition speeches are covered less favourable because of the tonality or understandability of their speeches.[37]

6.2 Representation of Speech Tone

Another way in which differences between the opposition and the ruling party might manifest is in the representation of the original speech tone. Specifically, I test whether the opposition dummy can predict how similar the tone of the speech is to the tone of its quotation(s). If there is bias in representation along the dimension of speech tone, then there should be a bigger change in tone for the opposition than the ruling party. While the descriptive statistics indicate that journalists from *The Straits Times* prefer reporting on speeches that sound neutral and positive (see Figures A.12 and A.13), the results from Table A.2 suggests no systematic differences in the representation of speech tone between the opposition and the ruling-party politicians.

[37] Another concern is that linguistic competencies may differ in terms of grammatical errors, and that the mechanical inaccuracies arises because of corrections either in the news article of the official transcripts. Direct quotes, however, should still be verbatim. As are the official transcripts.

Table 7. Examples of Speeches and Quotes

Quote fragment	Originating speech	Quote Accuracy
a fundamental relook	...is taking **a more fundamental relook** at this regulation framework and see how best we can support this strategy ...	95
there will be a need for us to make sure we have regular fare increases of the right quantum	...So **there will be a need for us to make sure that we have regular fare increases of the right quantum**...	95
likely to go back down the slippery slope	...considering that some of them may be drop-outs or expelled from school, they are **likely to go back down the slippery road** ...	90
core Singapore values	...Rather it is an acknowledgement that the **core Singaporean values** of multi-racialism and meritocracy can and should co-exist with each other...	90
resilience in response to ground reaction	...commended the PAP's **resilience in response to the ground's reaction** after the election, and I said this augurs well for Singapore...	85
too employer-focused	...and increase in absentee payroll are **rather employer-focused**...	85
That is the purpose of these amendments	...That is in fact **one of the purposes of this amendment** which we are bringing to Parliament...	80
deviates from the concept of free market	...this Bill **deviates from the idea of the concept of a free market**, where supply of services by companies is set by market demand ...	80
the high end	...we noted that the rates were at **a higher end**, but we had the rates that were charged ...	75
They (said) there are two purposes	...They gave the reasons that they wanted the two budget hotels **to serve two purposes**...	75
look them in the eye	...I want to be able to **look these men and others in the eyes** and say to them ...	70
to help ensure all Singaporeans can afford their education in a public university	...To **ensure that all Singaporeans can afford to attend Government-funded universities** ...	70
temporary or permanent solutions for this important issue	...to propose certain solutions, whether **temporary or permanent, to help resolve this, to me, a very important issue**...	65
completely unwarranted, alarmist, and show fundamental lack of understanding about the law	...I would venture to suggest that such statements are **alarmist and reveal a fundamental misunderstanding as to what this Bill and the law** is all about ...	65

Notes—The speech column shows the portion of the speech which contains the best partial substring match. The text **in bold** is an approximation of the best partial substring match. Quote accuracy is as defined by the substring quote accuracy measure.

7 Unpacking the Findings

7.1 Salience of Quotation Accuracy

One reason this paper focuses on quotation accuracy is that it is arguably a cleaner measure of coverage. Direct quotes are also particularly salient in journalism since they potentially carry more weight in the public eye compared to other types of statements in a news article, such as indirect testimonies (Gibson and Zillmann 1993; D'Alessio 2003). Getting quoted from a speech with more fragments may carry a higher risk of the quotes being taken out of context (intentional or not). Getting less accurate quotes increases the risk of misrepresentation (again, intentional or not).

A concrete example from the sample is the quote "the alternative will be too painful to bear," which originated from the speech

> . . . So, let us not take our harmonious social fabric for granted because **the alternative *may* be too painful to endure**. This is one pillar of success that we must continue to invest in. . .

Though the quote and its originating substring look fairly similar, the quote accuracy score is 84 (out of 100 using the substring accuracy measure). In particular, the quote uses *will be* where the speech uses *may be*—a subtle modulation in assertiveness.

Media coverage of parliamentary speeches may also be important in an institution like Singapore, where campaign periods are in practice very short—ten days or less.[38]

7.2 Magnitudes

From the results at the article-speech level and the quote level analyses, the opposition politicians are covered just as intensely as their ruling-party counterparts

[38]The *Parliamentary Elections Act* allows for a maximum of a 55 day campaign period, the three general elections in the sample period, however, have campaign durations of only ten days or less.

from the parliamentary speeches, conditional on time trends and observables. This finding runs counter to those perceptions of media bias in Singapore, where the ruling party gets higher coverage *rates*.

On average, opposition politicians get quoted from a third more fragment (or 18% of the standard deviation). Compared to the unconditional average of 2.1 quote fragments per politician in an article-speech observation, the opposition politicians are quoted using 14% more fragments. On average, the opposition politicians also get less accurate quotes by 1.5 to 2.4 points (by the substring and bag-of-words quote accuracy measures), which is 11.6% to 22.9% of the standard deviations. Compared to the unconditional accuracy of 91.4 and 96.4 (out of 100) per quote, the opposition gets quotes that are 1.6% to 2.5% less accurate.

Overall, the opposition gets less favourable coverage of their parliamentary speeches. The magnitudes, however, are not especially large. First, though the opposition is quoted from more fragments, this is on average less than one full unit of a fragment. Or, more of the opposition quotes are similar in fragmentation to the ruling party than they are different. Second, though the opposition politicians get quotes that are systematically less accurate, once the differences are accounted for, the opposition does get quotes of *decent* accuracy—conditional on the observables, the opposition still gets quotes that are 90 (out of 100) points accurate. Table 7 provides more speech-quote examples for a sense of how the accuracy scores translate into human evaluation of accuracy.

7.3 Contextualising via the Literature

Here, I contextualise the findings where the opposition politicians get quoted less accurately in the news articles in general. In the subsection that follows, I contextualise under a very institutional-specific logistics difference.

First, the evidence is not in favour of an overt media capture. While inequality (Corneo 2006; Petrova 2008) and media concentration (McMillan and Zoido 2004;

Besley and Prat 2006) increase the risk of capture, and while it is a fact that high-ranking public officials have been placed in senior management positions of *The Straits Times* (George 2012), their effects may have been mediated by the news monopoly's listing on a stock exchange with 99.9% of its shares owned by the public.[39]

The findings are, however, consistent with how the heavy media-related regulations in Singapore prod journalists in such a way that they are just a tad more careful when quoting the ruling party, which has greater resources to bring legal actions to bear. Journalists, especially those working in dailies (as opposed to weeklies), trade off between accuracy and timeliness (Berry 1967). Less accurate quotations of the opposition through the allocation of time spent on checking need not be an intentional partisan slant. The examples in Table 7 suggest that inaccuracies in the quotes come from carelessness (or liberties in words/synonyms); and not from altering quotes with nefarious intent.

Third, the findings are also consistent with the story of demand-driven bias within a spatial competition where: (i) news consumers drive media bias,[40] (ii) social media, with anti-establishment sentiments, offers a marginal substitution to mainstream media, and (iii) media outlets compete along with a political space (as in Hotelling 1929; Mullainathan and Shleifer 2005), with the opposition and the ruling party on opposite ends. In the absence of competitors, the mainstream media monopoly locates in the centre to gain the largest market share possible. But social media, which locates near the extreme opposition end of the political spectrum, leaves a truncated space for mainstream media to optimise market share. The middle of this truncated space is slightly towards the establishment relative to the entire competition space.

[39] The remaining shares are owned by the senior management, which is consistent with incentive compatibility under the principal-agent problems.

[40] Mullainathan and Shleifer (2005) and Gentzkow and Shapiro (2006) model bias as determined from the demand-side, and several empirical evidence supports a demand-driven bias (e.g., Groeling and Kernell 1998; Bernhardt et al. 2008; Gentzkow and Shapiro 2010; Larcinese et al. 2011).

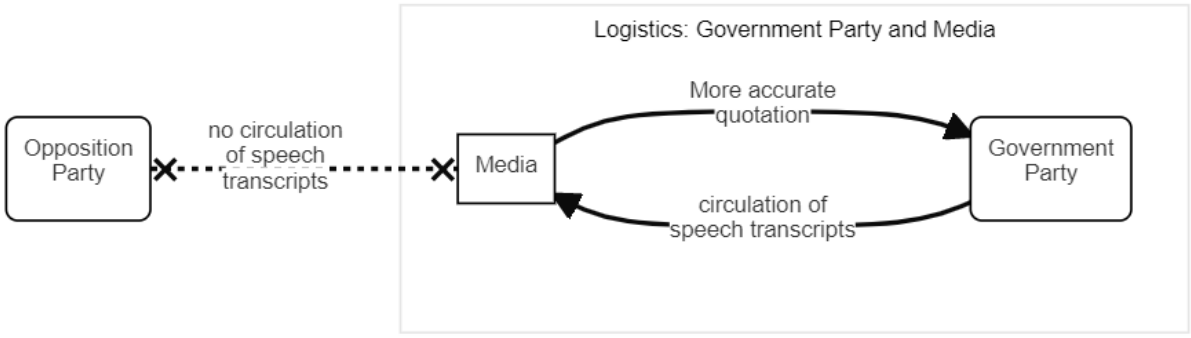


Figure 3. Institutional Logistical Differences

7.4 Contextualising via Institutional Logistical Differences

Here, I situate the findings in a context that emerges from primary and secondary research, from both media and political agents, that accounts for a specific government-to-media communications machinery. For primary research, I conduct a face-to-face interview with a senior journalist at The Straits Times and a member of parliament in the 14th parliament. Secondary research comes from a reviewer who recounts, in detailed writing, their experience as a member of parliament in the 13th parliament.

First, ruling-party politicians circulate their speech transcripts in advance to the media (Figure 3). Politicians who introduce a Bill or Motion, usually of higher ranks or political office holders, not only circulate their speeches to the media in advance but may have also briefed the media agents in advance. This pre-circulation allows the media ample time to prepare their media pieces. Non-senior ruling-party politicians also circulate their speeches to the media in advance, or the media may even initiate the request for the transcript.

Second, and however, opposition politicians do not share the same practices mentioned above. Furthermore, opposition politicians are less likely to send in their transcripts after the sitting voluntarily, and if they do, only send it in late. It also stands to reason that if they are unwilling to share voluntarily, they are also unlikely to respond to media agents who initiate the request, if any, for a copy of the transcript before the sitting begins. Section A.1 provides more details on parliamentary procedures, speeches, and differences in media engagement.

Given that the official speech transcripts are released only a week after sitting, and virtually all news stories are either the same or the next day, media agents have to rely on shorthand notes, video recordings of the sitting, or the pre-circulated speech transcripts. If accuracy increases in access (and timeliness) to the transcripts, the above would suggest that the logistical difference between the ruling and opposition parties would completely explain away the differences in coverage accuracy. The same can be said about the fragmentation of coverage. This is a potential explanation that, unfortunately, cannot be ruled out with the data.

7.5 Rational Choices in a Separating Equilibrium

However, to the extent that the detected partisan difference in quotation accuracy fully reflects logistical differences—the choice on whether to circulate speech transcripts in advance to media agents—the findings in this paper still point to beliefs about media slant as follows.

Allowing media agents early access to speeches leads to two things. First, it increases quotation accuracy ($A_i \in \mathbb{R}_+$). Second, and on the other hand, early access creates disutility from spin ($S_i \in \mathbb{R}_+$). Let c_1 indicate (early) circulation of speech and c_0 otherwise. A political agent i trades off between increasing quotation accuracy and spin by choosing whether to circulate. If the speech is not circulated, there is no spin, and the media reports the speech at face value, or that $S_i(c_0) = 0$. A political agent, therefore, gains from granting early access if

$$A_i(c_1) - S_i(c_1) > A_i(c_0),$$

where the gain in accuracy is greater than the disutility from spin when circulating the speech in advance.

Political agents care about being quoted accurately, and they care about spin because they want control over the framing of their speeches. This precludes the

kind of framing pre-agreed upon by both parties. As an illustration, a quotation about "raising taxes by x% to improve fiscal position" can be misquoted in different ways. One is "raising taxes by y%", another is "raising taxes by x% to stabilise fiscal position".

The story itself can have different spins. It can be framed as the government lacking fiscal prudence in recent years, or as a story about government prescience on future spending needs. It can also be threaded into a pre-existing narrative on social spending (or lack thereof). Spin increases with early circulation because it gives media agents time to develop, or skew, the narrative. To be precise, spin here has no value judgement but captures only magnitude for simplicity. To this extent, spin is undesirable because it takes narrative control away from political agents even if there is a non-zero probability that the political agent ends up 'looking good'. Without early access to develop a richer narrative, the media agent takes the "[improving] fiscal position" narrative at face value.

Here, I rely on the insights drawn from the primary and secondary research. From the institutional fact that a separating equilibrium exists, where the ruling-party politicians ($i = r$) choose to circulate their speeches but the opposition ($i = o$) does not, implies that

$$A_r(c_1) - S_r(c_1) > A_r(c_0)$$

where the ruling-party politicians gain from circulating the speeches, but

$$A_o(c_1) - S_o(c_1) < A_o(c_0),$$

where the opposition politicians do not.

Moreover, if one takes the position where the detected differences in quotation accuracy in this paper are not reflective of slant, but reflective entirely of logistical differences[41]—this implies that $A_r(c_\ell) = A_o(c_\ell) = A(c_\ell)$, where $\ell \in \{0, 1\}$. In which

[41]This position, which is taken by the anonymous journalist and the anonymous member of the 13th parliament from the primary and secondary research, implies that the detected quotation accuracy difference in this paper is capturing $A_r(c_1) > A_o(c_o)$, which is not a fair indication of media

case, the above two equations collapse into

$$A(c_1) - S_r(c_1) > A(c_0) > A(c_1) - S_o(c_1),$$

or that

$$S_o(c_1) > S_r(c_1),$$

which says that even if observed differences in accuracy do not originate from slant but from choices in advance speech circulation, the separating equilibrium, at the very least, still reveals private beliefs about a media that slants towards the ruling party. These beliefs—to the extent that interactions between political and media agents are frequent and that media pieces involving political agents are frequent—are credible given ample evidence for updating.

7.6 Consequences of Quotation Inaccuracy

Finally, even if media slant manifests in quotation inaccuracy, it is an agnostic error in the context of this study in that there is no clear implication of inaccuracy. In Section 7.1, I draw from the communications literature, which suggests that direct quotations carry non-trivial weight in a media piece. Table 7, however, which shows some examples of misquotations, suggests that the inaccuracies in the sample are likely innocuous errors and one, therefore, should not expect significant consequences to the slight misquotations.

Moreover, from the above Section 7.5, the institutional fact that a separating equilibrium exists—where the opposition politicians are willing to trade away quotation accuracy to avoid spin— suggests that the former is less important than the latter. One way to address the consequences of misquotation would be to get participants in a randomized controlled trial to evaluate how their judgement of a speech is affected by misquotation. This is a future avenue of research.

slant because of logistical differences (c).

8 Conclusion

This paper explores a novel methodological approach in that the detection of political media bias is done entirely in a sterile and objective environment. Coverage accuracy as an outcome measure is more plausibly objective. The quantification of further information from the textual data is machine annotated, which, while not free of bias or errors, does not require any subjective human judgements.

In the discussion, I place the findings in institutional-specific media-political machinery. The notion of accuracy developed in this paper can, at the very least, detect differences in media engagement strategy even if media slant cannot be directly detected.

On a final note, the approach in this paper does not suggest that the extant literature that uses coverage intensity to measure slant is lacking. Instead, using coverage intensity in the media of Singapore is insufficient to detect slant. Using coverage accuracy to detect slant can be replicated for any media and is useful in contexts where slant is less overt. Other potential future applications of coverage accuracy include coverage of financial reports and science journalism to assess how accurately the media represents scientific findings.

References

Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015). Radio and the Rise of The Nazis in Prewar Germany. *The Quarterly Journal of Economics 130*(4), 1885–1939.

Altonji, J., T. Elder, and C. Taber (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy 113*(1), 151–184.

Anderson, S. P. and J. McLaren (2012). Media Mergers and Media Bias With Rational Consumers. *Journal of the European Economic Association 10*(4), 831–859.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton : Princeton University Press, c2009.

BBC News (2013). Singapore profile - Media.

Bernhardt, D., S. Krasa, and M. Polborn (2008). Political Polarization and the Electoral Effects of Media Bias. *Journal of Public Economics 92*(5), 1092–1104.

Berry, F. C. (1967). A Study of Accuracy in Local News Stories of Three Dailies. *Journalism Quarterly 44*(3), 482–490.

Besley, T. and R. Burgess (2002). The Political Economy of Government Responsiveness: Theory and Evidence from India. *The Quarterly Journal of Economics 117*(4), 1415–1451.

Besley, T. and A. Prat (2006). Handcuffs for the Grabbing Hand? Media Capture and Government Accountability. *American Economic Review 96*(3), 720–736.

Blei, D., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993–1022.

Boas, T. C. and F. D. Hidalgo (2011). Controlling the Airwaves: Incumbency Advantage and Community Radio in Brazil. *American Journal of Political Science 55*(4), 869–885.

Chang, J., S. Gerrish, C. Wang, and D. M. Blei (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, 288—-296.

Chiang, C.-F. and B. Knight (2011). Media Bias and Influence: Evidence from Newspaper Endorsements. *The Review of Economic Studies 78*(3), 795–820.

Corneo, G. (2006). Media Capture in a Democracy: The Role of Wealth Concentration. *Journal of Public Economics 90*(1), 37–58.

D'Alessio, D. (2003). An Experimental Examination of Readers' Perceptions of Media Bias. *Journalism & Mass Communication Quarterly 80*(2), 282–294.

DellaVigna, S. and E. Kaplan (2007). The Fox News Effect: Media Bias and Voting. *Quarterly Journal of Economics 122*(3), 1187–1234.

Durante, R. and B. Knight (2012). Partisan Control, Media Bias, and Viewer Responses: Evidence from Berlusconi's Italy. *Journal of the European Economic Association 10*(3), 451–481.

Eisensee, T. and D. Strömberg (2007). News Droughts, News Floods, and U. S. Disaster Relief. *The Quarterly Journal of Economics 122*(2), 693–728.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011). Media and Political Persuasion: Evidence from Russia. *American Economic Review 101*(7), 3253–3285.

Ferraz, C. and F. Finan (2008). Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes. *The Quarterly Journal of Economics 123*(2), 703–745.

Gentzkow, M. (2006). Television and Voter Turnout. *The Quarterly Journal of Economics* (3), 931.

Gentzkow, M. and J. M. Shapiro (2006). Media Bias and Reputation. *Journal of Political Economy 114*(2), 280–316.

Gentzkow, M. and J. M. Shapiro (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica 78*(1), 35–71.

George, C. (2012). *Freedom From the Press: Journalism and State Power in Singapore.* Singapore : NUS Press, c2012.

Gerber, A. S., J. G. Gimpel, D. P. Green, and D. R. Shaw (2011). How Large and Long-lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment. *American Political Science Review 105*(1), 135–150.

Gibson, R. and D. Zillmann (1993). The Impact of Quotation in News Reports on Issue Perception. *Journalism Quarterly 70*(4), 793–800.

Groeling, T. and S. Kernell (1998). Is Network News Coverage of the President Biased? *The Journal of Politics 60*(4), 1063–1087.

Groseclose, T. and J. Milyo (2005). A Measure of Media Bias. *Quarterly Journal of Economics 120*(4), 1191–1237.

Ho, D. E. and K. M. Quinn (2008). Measuring Explicit Political Positions of Media. *Quarterly Journal of Political Science 3*(4), 353–377.

Hotelling, H. (1929). Stability in Competition. *The Economic Journal 39*(153), 41.

Jurafsky, D. and J. H. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Upper Saddle River, N.J. : Prentice Hall, 2000.

Larcinese, V., R. Puglisi, and J. M. Snyder (2011). Partisan Bias in Economic News: Evidence on the Agenda-setting Behavior of U.S. Newspapers. *Journal of Public Economics 95*(9), 1178–1189.

Larreguy, H. A., J. Marshall, and J. M. Snyder Jr. (2014). Revealing Malfeasance: How Local Media Facilitates Electoral Sanctioning of Mayors in Mexico. *National Bureau of Economic Research Working Paper Series No. 20697*, 1–57.

Lehrer, A. (1989). Between Quotation Marks. *Journalism Quarterly 66*(4), 902–941.

Lim, C. S. H., J. M. Snyder, and D. Strömberg (2015). The Judge, the Politician, and the Press: Newspaper Coverage and Criminal Sentencing across Electoral Systems. *American Economic Journal: Applied Economics 7*(4), 103–135.

Lott, J. R. and K. A. Hassett (2014). Is Newspaper Coverage of Economic Events Politically Biased? *Public Choice 160*(1), 65–108.

Marshall, H. (1977, mar). Newspaper Accuracy in Tucson. *Journalism Quarterly 54*(1), 165–169.

McCarthy, P. M. and S. Jarvis (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods 42*(2), 381–392.

McMillan, J. and P. Zoido (2004). How to Subvert Democracy: Montesinos in Peru. *Journal of Economic Perspectives 18*(4), 69–92.

Miner, L. (2015). The Unintended Consequences of Internet Diffusion: Evidence from Malaysia. *Journal of Public Economics 132*, 66–78.

Mullainathan, S. and A. Shleifer (2005). The Market for News. *The American Economic Review 95*(4), 1031.

Mullainathan, S. and J. Spiess (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives 31*(2), 87–106.

Oster, E. (2017). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics*, 1–18.

Petrocik, J. R. (1996). Issue Ownership in Presidential Elections, with a 1980 Case Study. *American Journal of Political Science 40*(3), 825–850.

Petrova, M. (2008). Inequality and Media Capture. *Journal of Public Economics 92*(1), 183–212.

Puglisi, R. (2011). Being The New York Times: The Political Behaviour of a Newspaper. *B.E. Journal of Economic Analysis and Policy 11*(1).

Puglisi, R. and J. M. Snyder (2011). Newspaper Coverage of Political Scandals. *The Journal of Politics 73*(3), 931–950.

Qian, N. and D. Yanagizawa-Drott (2017). Government Distortion in Independently Owned Media: Evidence from U.S. News Coverage of Human Rights. *Journal of the European Economic Association 15*(2), 463–499.

Qin, B., D. Strömberg, and Y. Wu (2017). Why Does China Allow Freer Social Media? Protests versus Surveillance and Propaganda. *Journal of Economic Perspectives 31*(1), 117–140.

Qin, B., D. Strömberg, and Y. Wu (2018). Media Bias in China. *American Economic Review 108*(9), 2442–2476.

Shen, L. (2022). LexicalRichness: A small module to compute textual lexical richness.

Singapore Statutes Online (1974). Newspaper and Printing Presses Act.

Snyder, J. M. and D. Strömberg (2010). Press Coverage and Political Accountability. *Journal of Political Economy 118*(2), 355–408.

Strömberg, D. (2004). Radio's Impact on Public Spending. *Quarterly Journal of Economics 119*(1), 189–221.

Sutter, D. (2012). Is the Media Liberal? An Indirect Test Using News Magazine Circulation. *Applied Economics 44*(27), 3521–3532.

Tan, N. (2014). Institutionalized Succession and Hegemonic Party Cohesion in Singapore. In A. Hicken and E. M. Kuhonta (Eds.), *Party System Institutionalization in Asia: Democracies, Autocracies, and the Shadows of the Past*, pp. 49–73. Cambridge: Cambridge University Press.

Torruella, J. and R. Capsada (2013). Lexical Statistics and Tipological Structures: A Measure of Lexical Richness. *Procedia - Social and Behavioral Sciences 95*, 447–454.

A Appendix

A.1 Background to Parliamentary Speeches

This section synthesises parliamentary business facts from the official source https://www.parliament.gov.sg/parliamentary-business/glossary, and from insights drawn from the written experiences of an anonymous reviewer, who is a member of parliament in the 13th parliament. Insights from the anonymous reviewer comes mostly in the last two paragraphs.

Speeches are (mostly) scripted. Most of the speeches in the Parliament of Singapore are jointly planned by the Government in consultation with the Speaker of Parliament (Speaker) who, as a result, is rarely surprised by what Members

of Parliament (Members) intend to say. Bills and Motions, for instance, require at least two day's notice before their introduction, or even longer depending on who introduces the bill. Extemporaneous and impromptu speeches exist but are rare, and likely shorter. A Parliamentary Reporter records all these proceedings in shorthand before archiving in the official repository.

True debates are rare. One common context in which Members speak is on Bills or Motions. The sponsor of the bill, for instance, usually a Political Office Holder, is allowed to speak twice, once to introduce the Bill/Motion and once more to close the debate. Other Members are allowed up to 20 minutes to debate on the introduced Bill/Motion. Still, the interim response to the introduction and the response is likely prepared in advance. A different and rarer context is Ministerial Statements, given by a Minister regarding the Government's policy and decision. Although no notice is required for such statements, it is in all likelihood prepared in advance.

Deviations from scripted exchange. One context in which unscripted exchanges can arise is via Parliamentary Question Time, a period set aside at the beginning of every sitting, where Ministers or Members respond to Questions as filed and accepted by the Speaker, until the end of the allocated Question Time. Since these Questions are filed in advance, responses are scripted. However, once a response is given, any other Member may spontaneously ask supplementary questions relating to the original Question. Here, there is room for unscripted responses, to the extent that the supplementary questions have not been pre-empted and are not made known to colleagues in advance. Deviations from scripted exchange within the incumbent-party members may also arise when Members want to further engage on points of clarification, but these usually occur near the end of a debate.

The main source of unscripted exchanges, therefore, relates to opposition Members who, unlike the incumbent-party Members, do not circulate their speeches and questions in advance. As a result, it becomes difficult for the majority of the incumbent-party Members to prepare scripted responses to speeches from opposition Members.

Overall, the extent to which speeches and their responses are predictable rests on the existing media-political machinery and differs by seniority and party status. Incumbent-party Members are more likely to both circulate their speeches and questions in advance, both among their party colleagues and also to the media. The opposition Members, however, do not.

A.2 OLS Estimate of the Opposition Dummy is a Lower Bound on the True *Magnitude*

The following shows that treating ministerial rank as a bad control suggest that the OLS estimate of the opposition dummy carries an attenuation bias (an upward bias when the true effect is negative), even if assignment into party is random. Specifically, opposition status determines not only political media coverage, but also the ministerial rank of politicians since the opposition politicians never get a higher than base (lowest) rank. Hence even conditioning on rank does not recover the true causal effect because of the composition of the politicians and their rank.

As a simplification, the following abstracts away from other regressors ($\boldsymbol{X}$), or

more specifically assuming that opposition status does not determine $\boldsymbol{X}$ other than rank. In reality, there is also a whole range of ministerial rank/type, but these are all collapsed into a single dummy indicating a higher than base rank, with base rank as the omitted category. Let the opposition status indicator for politician i be

$$D_i = \begin{cases} 1 & \text{if politician } i \text{ is from an opposition party} \\ 0 & \text{otherwise,} \end{cases}$$

and let the higher rank indicator for politician i be

$$r_i = \begin{cases} 1 & \text{if politician } i \text{ holds a higher than base (lowest) level rank} \\ 0 & \text{if base (lowest) rank.} \end{cases}$$

Following Angrist and Pischke (2009)'s treatment of bad controls in the context of this paper, opposition status (D_i) determines both media coverage (y_i) and the rank (r_i):

$$\begin{aligned} y_i &= y_{0i} + (y_{1i} - y_{0i})D_i \\ r_i &= r_{0i} + (r_{1i} - r_{0i})D_i, \end{aligned}$$

where y_{1i} and r_{1i} (y_{0i} and r_{0i}) are the potential media coverage and potential ministerial rank of politician i as an opposition (ruling-party) politician. One can think of rank affecting media coverage as an omitted variable problem, in that rank is increasing in some unobserved characteristics of a politician (e.g. competence, likeability, quotability, public image), and coverage is in turn increasing in these characteristics. By the joint independence of $\{y_{1i}, r_{1i}, y_{0i}, r_{0i}\}$ and D_i, comparing opposition and ruling-party politicians conditional on the base level rank ($r_i = 0$) gives:

$$\begin{aligned} &\mathbb{E}[y_i|D_i = 1, r_i = 0] - \mathbb{E}[y_i|D_i = 0, r_i = 0] \\ &= \mathbb{E}[y_{1i}|D_i = 1, r_{1i} = 0] - \mathbb{E}[y_{0i}|D_i = 0, r_{0i} = 0] \\ &= \mathbb{E}[y_{1i}|r_{1i} = 0] - \mathbb{E}[y_{0i}|r_{0i} = 0] \\ &= \mathbb{E}[y_{1i}|r_{1i} = 0] - \mathbb{E}[y_{0i}|r_{1i} = 0] + \mathbb{E}[y_{0i}|r_{1i} = 0] - \mathbb{E}[y_{0i}|r_{0i} = 0] \\ &= \mathbb{E}[y_{1i} - y_{0i}|r_{1i} = 0] + \mathbb{E}[y_{0i}|r_{1i} = 0] - \mathbb{E}[y_{0i}|r_{0i} = 0] \end{aligned}$$

The observed difference in political media coverage between oppositions and ruling-party politicians can be decomposed into two two parts. The first term in the last line of the equation above is the true causal effect of opposition status on coverage, conditional on having the base ministerial rank.

The bias comes from the second and third terms. The second term is the potential coverage of a ruling-party politician had he been given assignment to a base ministerial rank as an opposition politician. The third term is the potential coverage of a ruling-party politician given that he is assigned to a base ministerial rank.

In theory, the bias can go in either direction. But I propose here that a very

plausible set of assumptions suggests that the bias is likely positive. First, the average of the coverage of a ruling-party politician who would have been assigned to the base rank had he been an opposition politician, is simply the average coverage of all ruling-party politicians, so that $\mathbb{E}[y_{0i}|r_{1i} = 0] = \mathbb{E}[y_{0i}]$.

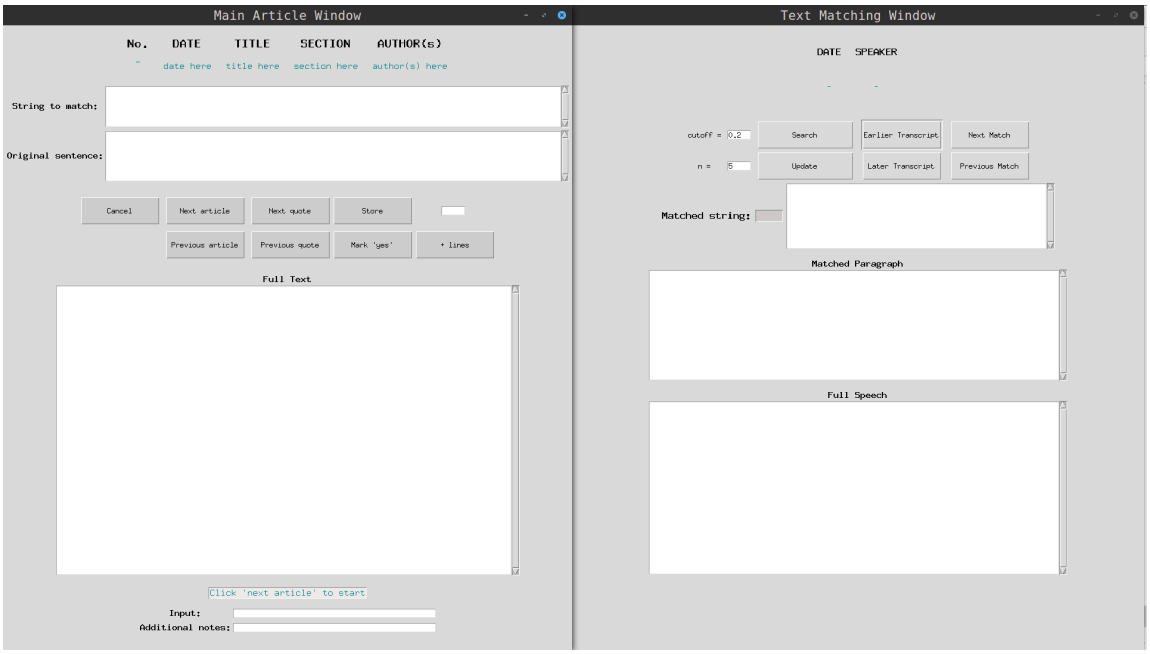
The average coverage of a ruling-party politician assigned to a base ministerial rank however, is likely below average if coverage increases with rank, so that $\mathbb{E}[y_{0i}|r_{0i} = 0] < \mathbb{E}[y_{0i}]$. Together, the assumptions imply that $\mathbb{E}[y_{0i}|r_{1i} = 0] = \mathbb{E}[y_{0i}] > \mathbb{E}[y_{0i}|r_{0i} = 0]$, or that the bias term in the equation $\mathbb{E}[y_{0i}|r_{1i} = 0] - \mathbb{E}[y_{0i}|r_{0i} = 0] > 0$. Hence the bias is positive. When the true causal effect of opposition status on coverage is negative, this becomes an attenuation bias towards zero, and the observed comparison of means from the OLS estimates constitute a lower bound on the true magnitude of the opposition effect.

A.3 Additional Tables and Figures

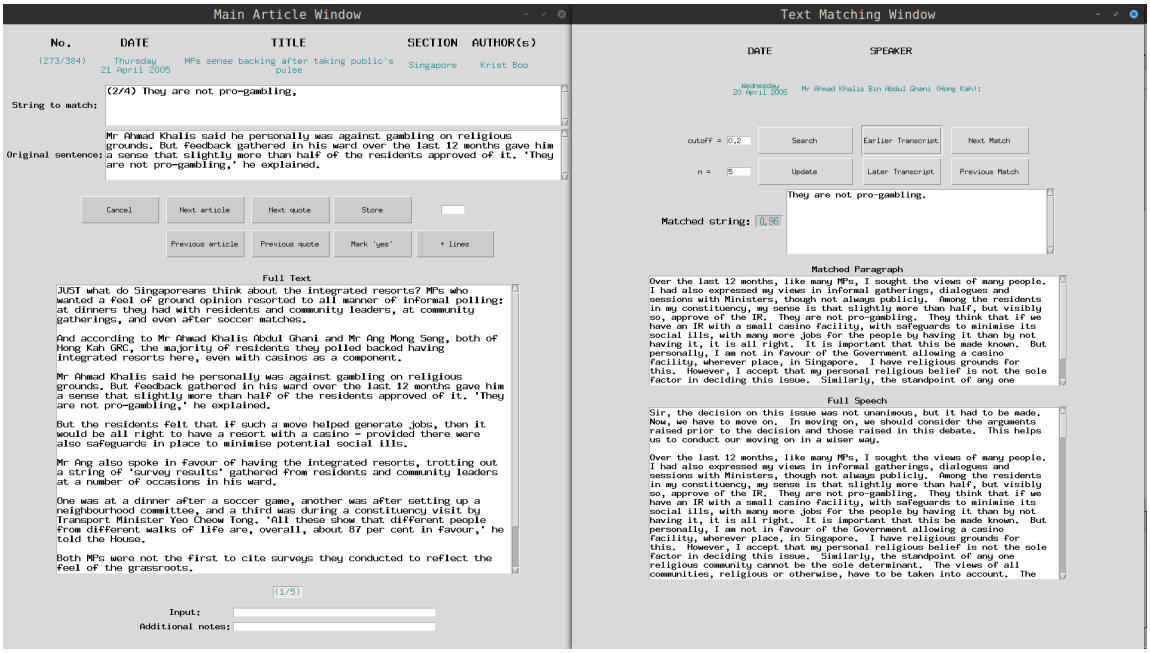
Table A.1. Daily Newspaper Subscription

	Daily Newspaper	Language	Unique Digital Subscription	Print Subscription	Digital Subscription	Total Subscription
1	Berita Harian (Berita Minggu)	Malay	911	44'600	2'500	47'100
2	The Business Times	English	6'658	29'200	18'500	47'700
3	Lianhe Zaobao	Chinese	13'727	148'600	39'300	187'900
4	Lianhe Wanbao	Chinese	1'137	82'500	9'100	91'600
5	The New Paper (The New Paper Sunday)	English	757	70'200	40'400	110'600
6	Shin Min Daily News	Chinese				120'200
7	The Straits Times (The Sunday Times)	English	60'871	304'300	177'400	481'700
8	Tamil Murasu (Tamil Murasu Sunday)	Tamil				12'800

Source: SPH 2015 Annual Report.
a) Subscriptions of the eight daily newspaper wholly-owned by Singapore Press Holdings.
b) Parenthesis indicates the Sunday edition of the daily. For instance, The Straits Times is published as The Sunday Times on Sundays.
c) In addition to total subscriptions, unique digital subscriptions, print subscriptions, and digital subscriptions are shown where available.



(a) GUI example without text



(b) GUI example with text

Figure A.1. Graphics user interface with quote matching and extraction

Table A.2. Differences in Objectivity and Polarity

	Dependent variable is the difference in objectivity/polarity scores from					
	Quote to			Sentence containing quote to		
	Full speech	Speech paragraph	Speech sentence	Full speech	Speech paragraph	Speech sentence
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A. Differences in objectivity					
Opposition	0.003	0.010	–0.011	0.014	0.020*	0.000
	(0.012)	(0.012)	(0.011)	(0.010)	(0.011)	(0.010)
	Panel B. Differences in polarity					
Opposition	–0.003	–0.005	0.002	0.000	–0.002	0.005
	(0.009)	(0.009)	(0.007)	(0.009)	(0.009)	(0.008)
Time fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes
Length controls	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Article controls	Yes	Yes	Yes	Yes	Yes	Yes
Topic controls	Yes	Yes	Yes	Yes	Yes	Yes
Ministerial controls	Yes	Yes	Yes	Yes	Yes	Yes
N	14,887	14,885	14,885	14,886	14,885	14,885

Notes—Observations are at the quote level. The first set compares the quote to each of three speech components: (i) the full speech, (ii) the speech paragraph containing the quote, and (iii) the speech sentence(s) containing the quote; the second set compares the sentence containing the quote to three speech components enumerated above. In panel A, the dependent variable is the difference in objectivity from speech to quote; in panel B, the dependent variable is the difference in polarity from speech to quote. The polarity and objectivity measures are generated using the open-source *TextBlob* Pattern Analyzer (https://textblob.readthedocs.io). Robust standard errors adjusted for clusters by newspaper article in parentheses.

*** Significant at the 1 per cent level.

** Significant at the 5 per cent level.

* Significant at the 10 per cent level.

Table A.3. Political Coverage During Pre-Election Periods

		Log of quote length by word count		Substring quote accuracy measure		Bag-of-words quote accuracy measure	
		Full	Subsample	Full	Subsample	Full	Subsample
Election periods	Variable	(1)	(2)	(3)	(4)	(5)	(6)
		Panel A. 6 & 3-months before a general election					
$E_t = 1$ if 6 months	Opposition	–0.15***	–0.06	–1.19*	–2.99	–2.09***	–5.33
before a general election		(0.05)	(0.22)	(0.69)	(3.34)	(0.69)	(3.28)
$N_{E_t} = 1630$	Opposition$\times E_t$	0.05	.	–3.10	.	–4.08*	.
$N_{opp} = 103$		(0.12)		(2.08)		(2.16)	
$E_t = 1$ if 3 months	Opposition	–0.14***	–0.35	–1.38**	3.58	–2.22***	–1.70
before a general election		(0.05)	(0.29)	(0.70)	(4.48)	(0.69)	(3.66)
$N_{E_t} = 1207$	Opposition$\times E_t$	0.04	.	–0.61	.	–3.05	.
$N_{opp} = 80$		(0.15)		(2.20)		(2.42)	
		Panel B. 3-months before a by-election					
$E_t = 1$ if 3 months	Opposition	–0.13***	–0.04	–1.47**	–3.23	–2.60***	–1.80
before a by-election		(0.05)	(0.19)	(0.73)	(2.37)	(0.74)	(2.29)
$N_{E_t} = 1636$	Opposition$\times E_t$	–0.07	.	0.55	.	2.60***	.
$N_{opp} = 92$		(0.12)		(1.27)		(0.98)	
		Panel C. 3-months & 1-month before any election					
$E_t = 1$ if 3 months	Opposition	–0.14***	–0.01	–1.43**	–0.34	–2.40***	–2.30
before any election		(0.05)	(0.14)	(0.73)	(2.05)	(0.73)	(1.82)
$N_{E_t} = 2843$	Opposition$\times E_t$	–0.01	.	0.03	.	0.06	.
$N_{opp} = 172$		(0.10)		(1.29)		(1.32)	
$E_t = 1$ if 1 month	Opposition	–0.15***	0.15	–1.35*	1.45	–2.43***	–1.33
before any election		(0.05)	(0.43)	(0.71)	(5.40)	(0.71)	(4.33)
$N_{E_t} = 974$	Opposition$\times E_t$	0.06	.	–1.92	.	0.89	.
$N_{opp} = 54$		(0.16)		(1.60)		(1.15)	

Notes—Observations are at the quote-level. The dependent variable in columns (1)–(2) is log of quote length; in columns (3)–(4) it is quote accuracy using measure (1); and in columns (5)–(6) it is quote accuracy using measure (2). Odd-numbered columns uses the full sample, while even-numbered columns uses the subsample observations for the relevant periods just before the elections. N_{E_t} indicates the number of observations in the relevant pre-election period; N_{opp} indicates the number of opposition quote observations in the same period. The models are the same as in the baseline specifications in Table 4, but with an additional Opposition$\times E_t$ term in the full sample model, which estimates indicates whether there is an additional effect during the pre-election periods. In the subsample models, the Opposition term estimates whether there is a difference between the opposition and ruling party politicians in the pre-election periods. Robust standard errors in parentheses are clustered at news articles.

*** Significant at the 1 per cent level.

** Significant at the 5 per cent level.

* Significant at the 10 per cent level.

Table A.4. Minsiterial Rank by party

Ministerial Rank		Non-opposition	Opposition	Total
Rank	Description			
PM	Prime Minister	491	0	491
DPM	Deputy PM	1185	0	1185
Minister	MR1-4 rank	6497	0	6497
SMS	Senior Minister of State	1012	0	1012
MOS	Minister of State	610	0	610
Mayor	Mayor (of 1 of 5 districts)	222	0	222
Sps	Senior Parliamentary Secretary	98	0	98
Parl Sec	Parliamentary Secretary	87	0	87
Speaker	Speaker of Parliament	38	0	38
MP	Member of Parliament (base rank)	3082	800	3882
NCMP	Non-constituency MP	0	306	306
NMP	Nominated MP	475	0	475
Total		13,797	1106	14,903

Table A.5. Newspaper section by party

Newspaper section	Non-opposition	Opposition	Total
Home	961	83	1044
Insight	50	7	57
Money	21	0	21
News	28	0	28
Opinion	11	0	11
Others	22	3	25
Prime News	1929	152	2081
Review - Insight	14	0	14
Singapore	8210	626	8836
Sports	14	0	14
ST	645	34	679
Think	12	3	15
Top of the News	1871	198	2069
World	9	0	9
Total	13,797	1106	14,903

Table A.6. Specification Checks for Quote fragments, Article-Speech level

						Clustering		Topic distributions of textual content					
	Baseline results (1)	Negative Binomial (2)	Journa-list FE (3)	No Ministerial controls (4)	No Translated quotes (5)	Cluster by speech (6)	Cluster by journalist (7)	Speech $K = 50$ (8)	Speech $K = 100$ (9)	Article $K = 30$ (10)	Article $K = 50$ (11)	Sentence topics (12)	Parsimonious topics (13)
Opposition	0.328***	0.167***	0.300***	–0.046	0.300***	0.328***	0.303**	0.318***	0.348***	0.406***	0.341***	0.317***	0.395***
	(0.089)	(0.045)	(0.101)	(0.067)	(0.090)	(0.083)	(0.119)	(0.087)	(0.091)	(0.089)	(0.089)	(0.089)	(0.088)
N	7,087	7,087	6,210	7,087	6,980	7,087	6,210	7,087	7,087	7,087	7,087	7,087	7,087

Notes—This Table reports specification checks for the baseline result in Table 3 where the coverage of opposition politicians are made up of more quote fragments than those of the ruling-party politicians. Column (2) models the count of quote fragments using the negative binomial regression. Column (3) includes journalist fixed-effects and beat dummies. Column (4) excludes ministerial controls on the grounds that ministerial type is a bad control. Columns (5) excludes observations that are recorded as translations (from vernacular to English). Columns (6) and (7) adjusts standard errors for clusters by speech and journalist instead of newspaper articles. Columns (8)–(13) tests various specifications of the topic distributions. Column (12) uses the topical distribution of the sentence containing the quote instead of the quote itself. Column (13) uses the most parsimonious well-performing topic distributions—K=30 for the news articles, and K=50 for the speeches. Robust standard errors in parentheses are clustered at news articles except in columns (6)–(7).

*** Significant at the 1 per cent level.
** Significant at the 5 per cent level.
* Significant at the 10 per cent level.

Table A.7. Specification Checks for Quote Accuracy, Article-Speech Level

					Clustering		Topic distributions for textual content					
	Baseline results (1)	Journa-list FE (2)	No Ministerial controls (3)	No Translated quotes (4)	Cluster at speech (5)	Cluster at journalist (6)	Speech $K = 50$ (7)	Speech $K = 100$ (8)	Article $K = 30$ (9)	Article $K = 50$ (10)	Sentence topics (11)	Parsimonious topics (12)
Panel A. Dependent variable is bag-of-words quote accuracy measure												
Opposition	–2.5***	–2.75***	–1.95***	–2.33***	–2.5***	–2.71***	–2.44***	–2.59***	–2.62***	–2.43***	–2.35***	–2.61***
	(0.785)	(0.899)	(0.594)	(0.778)	(0.779)	(0.944)	(0.784)	(0.799)	(0.784)	(0.765)	(0.773)	(0.789)
N	7,087	6,210	7,087	6,980	7,087	6,210	7,087	7,087	7,087	7,087	7,087	7,087
Panel B. Dependent variable is bag-of-words quote accuracy measure (No stop words)												
Opposition	–3.3***	–3.62***	–2.59***	–3.09***	–3.3***	–3.58***	–3.07***	–3.46***	–3.51***	–3.19***	–3.12***	–3.36***
	(0.947)	(1.07)	(0.721)	(0.931)	(0.937)	(1.21)	(0.936)	(0.966)	(0.945)	(0.92)	(0.938)	(0.937)
N	7,087	6,210	7,087	6,980	7,087	6,210	7,087	7,087	7,087	7,087	7,087	7,087

Notes—This Table reports specification checks for the baseline result in Table 3 where the coverage of opposition politicians are less accurate than those of the ruling-party politicians. Column (2) includes journalist fixed-effects and beat dummies. Column (3) excludes ministerial controls on the grounds that ministerial type is a bad control. Columns (4) excludes observations that are recorded as translations (from vernacular to English). Columns (5) and (6) adjusts standard errors for clusters by speech and journalist instead of newspaper articles. Columns (7)–(12) tests various specifications of the topic distributions. Column (11) uses the topical distribution of the sentence containing the quote instead of the quote itself. Column (12) uses the most parsimonious well-performing topic distributions—K=30 for the news articles, and K=50 for the speeches. Robust standard errors in parentheses are clustered at news articles except in columns (5)–(6).

*** Significant at the 1 per cent level.
** Significant at the 5 per cent level.
* Significant at the 10 per cent level.

Table A.8. Specification Checks for Quote Length, Quote Level

					Subsamples		Clustering		Topic distribution of textual content					
	Baseline results (1)	Negative Binomial (2)	Journa-list FE (3)	No ministerial controls (4)	No translations (5)	No low similarity (6)	Cluster at speech (7)	Cluster at journalist (8)	Speech $K = 50$ (9)	Speech $K = 100$ (10)	Article $K = 30$ (11)	Article $K = 50$ (12)	Sentence topics (13)	Parsimonious topics (14)
Panel A. Dependent variable is log of quote length by *word* count														
Opposition	−0.138***	−0.120***	−0.086*	−0.163***	−0.128***	−0.138***	−0.138***	−0.084**	−0.154***	−0.127***	−0.131***	−0.128***	−0.136***	−0.146***
	(0.047)	(0.036)	(0.048)	(0.039)	(0.049)	(0.048)	(0.048)	(0.042)	(0.047)	(0.047)	(0.047)	(0.048)	(0.048)	(0.047)
N	14,887	14,887	13,513	14,887	14,682	14,423	14,887	13,513	14,887	14,887	14,887	14,887	14,887	14,887
Panel B. Dependent variable is log of quote length by *character* count														
Opposition	−0.123***	−0.112***	−0.074*	−0.147***	−0.117**	−0.121***	−0.123***	−0.072*	−0.139***	−0.113**	−0.118***	−0.113**	−0.120***	−0.132***
	(0.044)	(0.035)	(0.045)	(0.036)	(0.045)	(0.044)	(0.045)	(0.040)	(0.043)	(0.044)	(0.044)	(0.045)	(0.044)	(0.044)
N	14,885	14,887	13,511	14,885	14,680	14,421	14,885	13,511	14,885	14,885	14,885	14,885	14,885	14,885
Panel C. Dependent variable is log of quote length by *word* count (No stop words)														
Opposition	−0.089**	−0.092***	−0.049	−0.100***	−0.084**	−0.091**	−0.089**	−0.047	−0.101***	−0.080**	−0.084**	−0.079**	−0.090**	−0.095**
	(0.039)	(0.035)	(0.039)	(0.032)	(0.041)	(0.040)	(0.040)	(0.034)	(0.039)	(0.040)	(0.039)	(0.040)	(0.040)	(0.039)
N	14,887	14,887	13,513	14,887	14,682	14,423	14,887	13,513	14,887	14,887	14,887	14,887	14,887	14,887
Panel D. Dependent variable is log of quote length by *character* count (No stop words)														
Opposition	−0.081**	−0.084**	−0.038	−0.108***	−0.078*	−0.081**	−0.081*	−0.035	−0.095**	−0.073*	−0.077*	−0.070*	−0.082**	−0.089**
	(0.041)	(0.035)	(0.041)	(0.034)	(0.042)	(0.041)	(0.042)	(0.037)	(0.040)	(0.041)	(0.041)	(0.042)	(0.041)	(0.040)
N	14,872	14,887	13,499	14,872	14,667	14,409	14,872	13,499	14,872	14,872	14,872	14,872	14,872	14,872

Notes—This Table reports specification checks for the baseline result in Table 4 where the quotes of opposition politicians are shorter than those of the ruling-party politicians. In panels C & D, stop words (e.g. "the", "or", "a", "we" "be") are removed before the relevant measures are computed. Column (1) presents the baseline result. Column (2) models quote length as a count variable of words and characters using the negative binomial regression. Column (3) includes journalist fixed-effects and beat dummies. Column (4) excludes ministerial controls on the grounds that ministerial type is a bad control. Columns (5) and (6) excludes observations that are recorded as translations (from vernacular to English) and observations which have low quote accuracy (observations with both similarity measures below 75 are excluded). Columns (7) and (8) adjusts standard errors for clusters by speech and journalist instead of newspaper articles. Columns (9)–(14) tests various specifications of the topic distributions. Column (13) uses the topical distribution of the sentence containing the quote instead of the quote itself. Column (14) uses the most parsimonious well-performing topic distributions—K=30 for the news articles, and K=50 for the speeches. Robust standard errors in parentheses are clustered at news articles except in columns (7)–(8).

*** Significant at the 1 per cent level.
** Significant at the 5 per cent level.
* Significant at the 10 per cent level.

Table A.9. Specification Checks for Quote Accuracy, Quote Level

					Clustering		Topic distribution of textual content					
	Baseline Regression (1)	Journa-list FE (2)	No ministerial controls (3)	No translations (4)	Cluster by speech (5)	Cluster by journalist (6)	Speech $K = 50$ (7)	Speech $K = 100$ (8)	Article $K = 30$ (9)	Article $K = 50$ (10)	Sentence topics (11)	Parsimonious topics (12)
Panel A. Dependent variable is substring quote accuracy measure												
Opposition	−1.455**	−1.953***	−1.570***	−1.302*	−1.455**	−1.928**	−1.223*	−1.516**	−1.736**	−1.515**	−1.569**	−1.488**
	(0.701)	(0.742)	(0.562)	(0.699)	(0.739)	(0.753)	(0.698)	(0.712)	(0.700)	(0.701)	(0.712)	(0.695)
Observations	14,887	13,513	14,887	14,682	14,887	13,513	14,887	14,887	14,887	14,887	14,887	14,887
Panel B. Dependent variable is substring quote accuracy measure (No stop words)												
Opposition	−1.662**	−2.226***	−1.587***	−1.495**	−1.662**	−2.220***	−1.333*	−1.746**	−1.990***	−1.699**	−1.742**	−1.667**
	(0.723)	(0.770)	(0.571)	(0.718)	(0.759)	(0.779)	(0.717)	(0.738)	(0.725)	(0.725)	(0.732)	(0.717)
Observations	14,887	13,513	14,887	14,682	14,887	13,513	14,887	14,887	14,887	14,887	14,887	14,887
Panel C. Dependent variable is bag-of-words quote accuracy measure												
Opposition	−2.434***	−2.594***	−1.802***	−2.321***	−2.434***	−2.573***	−2.200***	−2.284***	−2.660***	−2.349***	−2.301***	−2.446***
	(0.707)	(0.750)	(0.566)	(0.718)	(0.726)	(0.839)	(0.710)	(0.707)	(0.706)	(0.689)	(0.701)	(0.705)
Observations	14,887	13,513	14,887	14,682	14,887	13,513	14,887	14,887	14,887	14,887	14,887	14,887
Panel D. Dependent variable is bag-of-words quote accuracy measure (No stop words)												
Opposition	−3.244***	−3.454***	−2.437***	−3.131***	−3.244***	−3.440***	−2.817***	−3.077***	−3.590***	−3.108***	−3.065***	−3.181***
	(0.902)	(0.967)	(0.713)	(0.914)	(0.927)	(1.174)	(0.893)	(0.900)	(0.905)	(0.881)	(0.901)	(0.892)
Observations	14,887	13,513	14,887	14,682	14,887	13,513	14,887	14,887	14,887	14,887	14,887	14,887

Notes—This Table reports specification checks for the baseline result in Table 3 where the quotes of opposition politicians are less accurate than those of the ruling-party politicians. In panels C & D, stop words (e.g. "the", "or", "a", "we" "be") are removed before the relevant measures are computed. Column (1) presents the baseline result. Column (2) includes journalist fixed-effects and beat dummies. Column (3) excludes ministerial controls on the grounds that ministerial type is a bad control. Columns (4) excludes observations that are recorded as translations (from vernacular to English). Columns (5) and (6) adjusts standard errors for clusters by speech and journalist instead of newspaper articles. Columns (7)–(12) tests various specifications of the topic distributions. Column (11) uses the topical distribution of the sentence containing the quote instead of the quote itself. Column (12) uses the most parsimonious well-performing topic distributions—K=30 for the news articles, and K=50 for the speeches. Robust standard errors in parentheses are clustered at news articles except in columns (5)–(6).

*** Significant at the 1 per cent level.

** Significant at the 5 per cent level.

* Significant at the 10 per cent level.

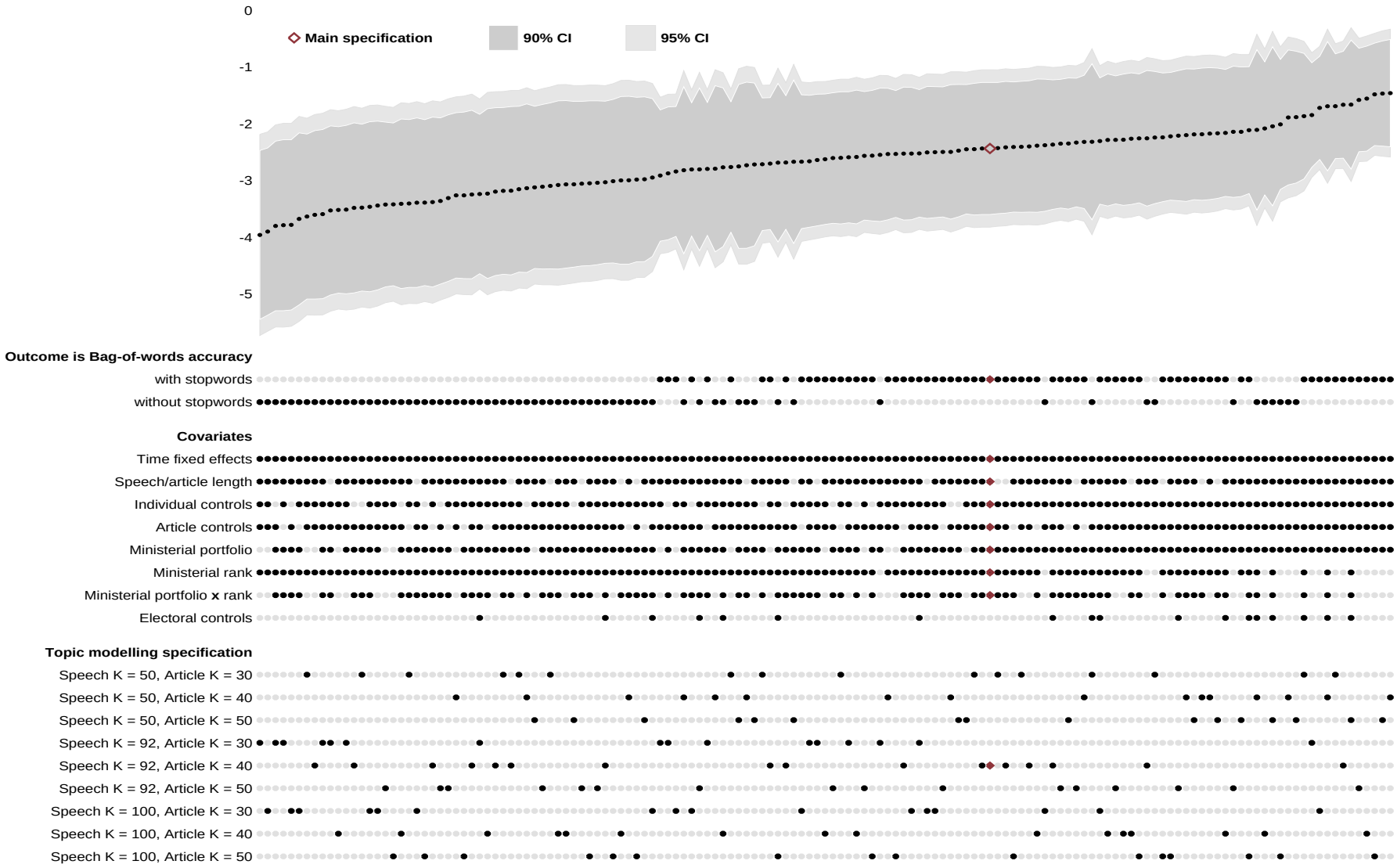


Figure A.2. Effect Sizes of Opposition Status on Bag-of-words Accuracy

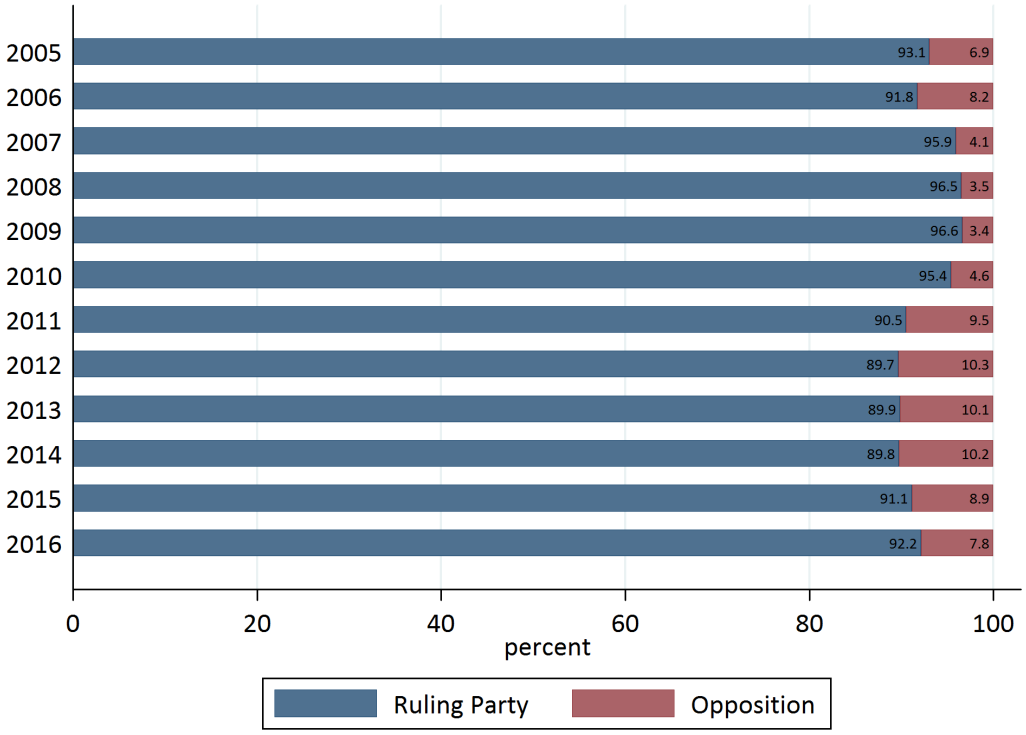


Figure A.3. Count of quotes over the years by partisanship

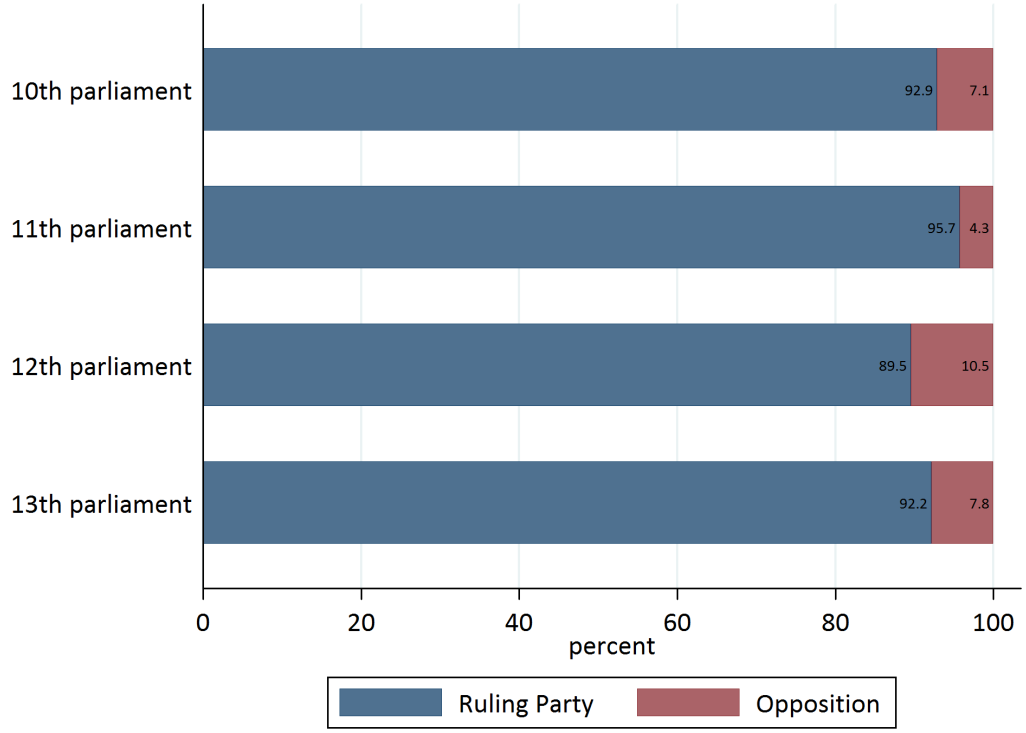


Figure A.4. Count of quotes over parliaments by partisanship

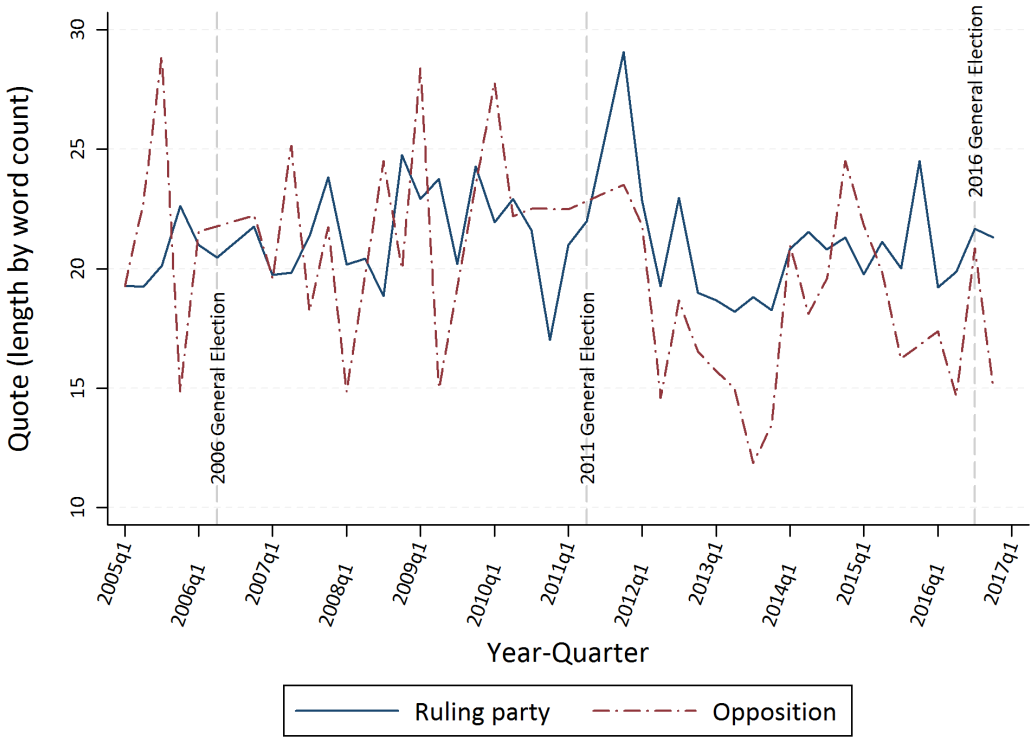


Figure A.5. Quote length over time

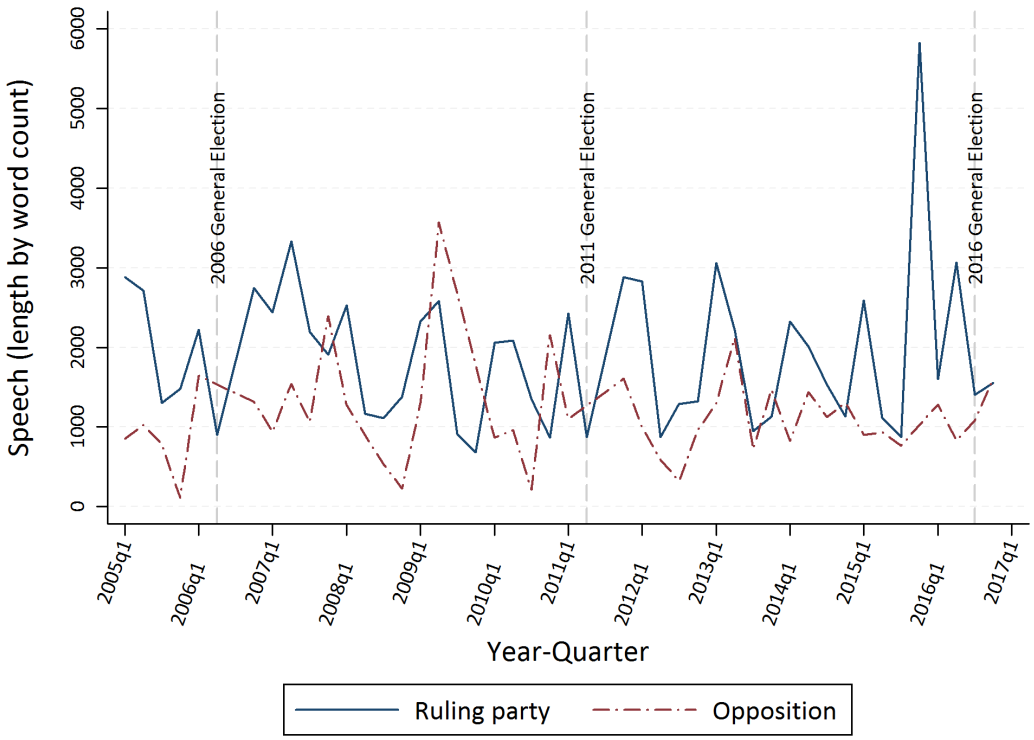


Figure A.6. Speech length over time

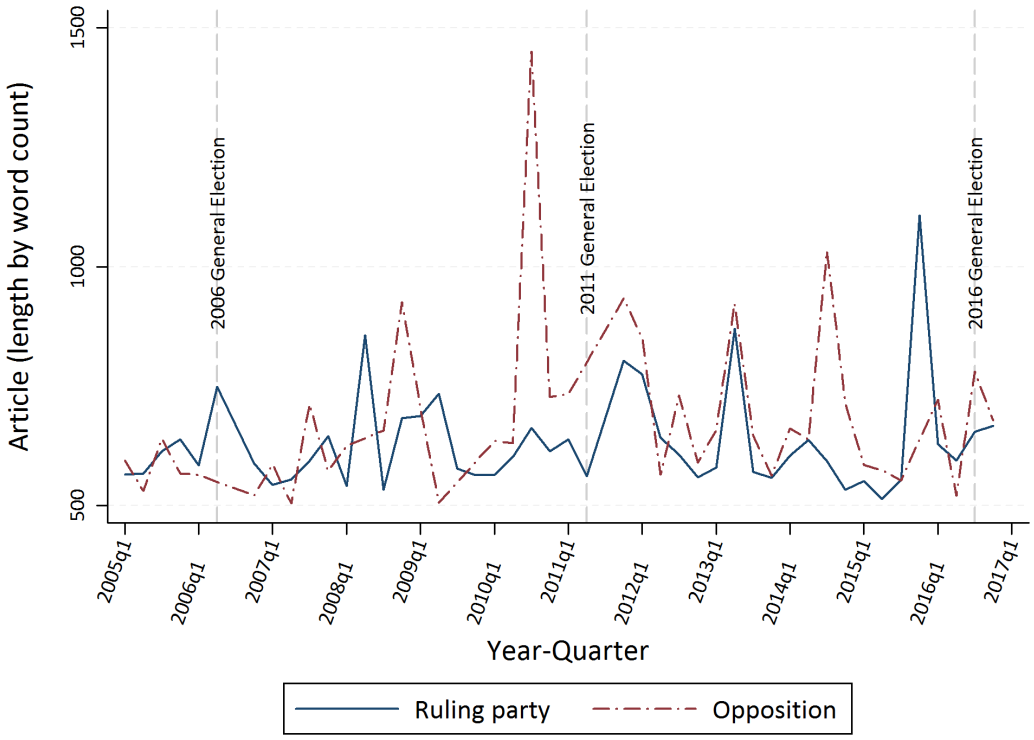


Figure A.7. Article length over time

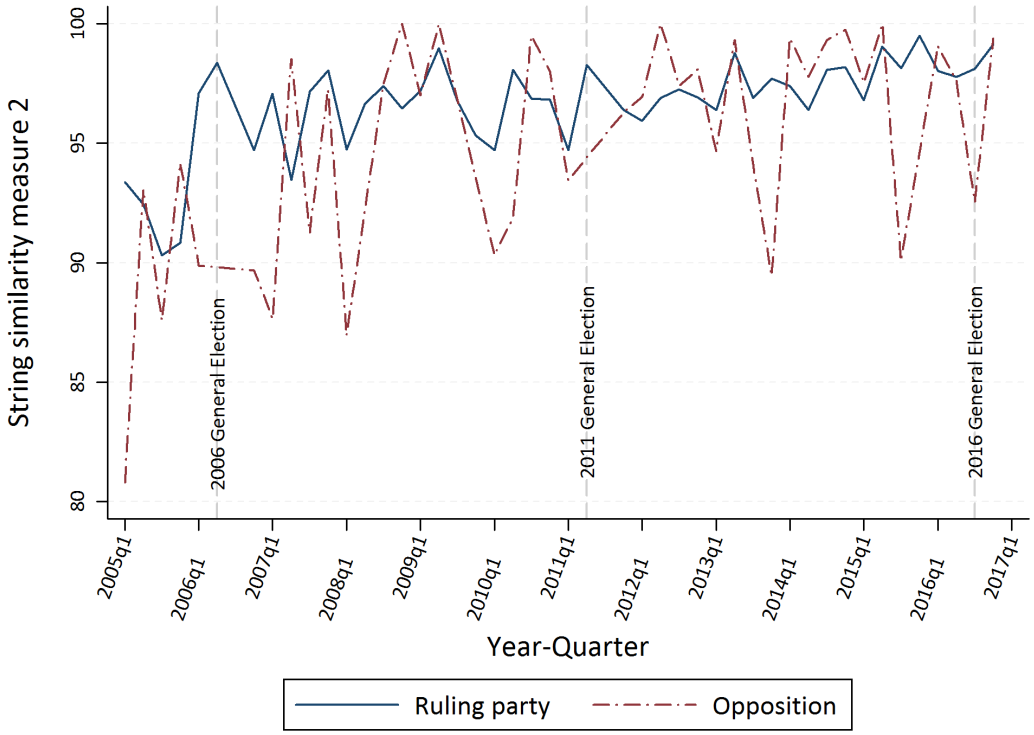


Figure A.8. Bag-of-words accuracy measure over time

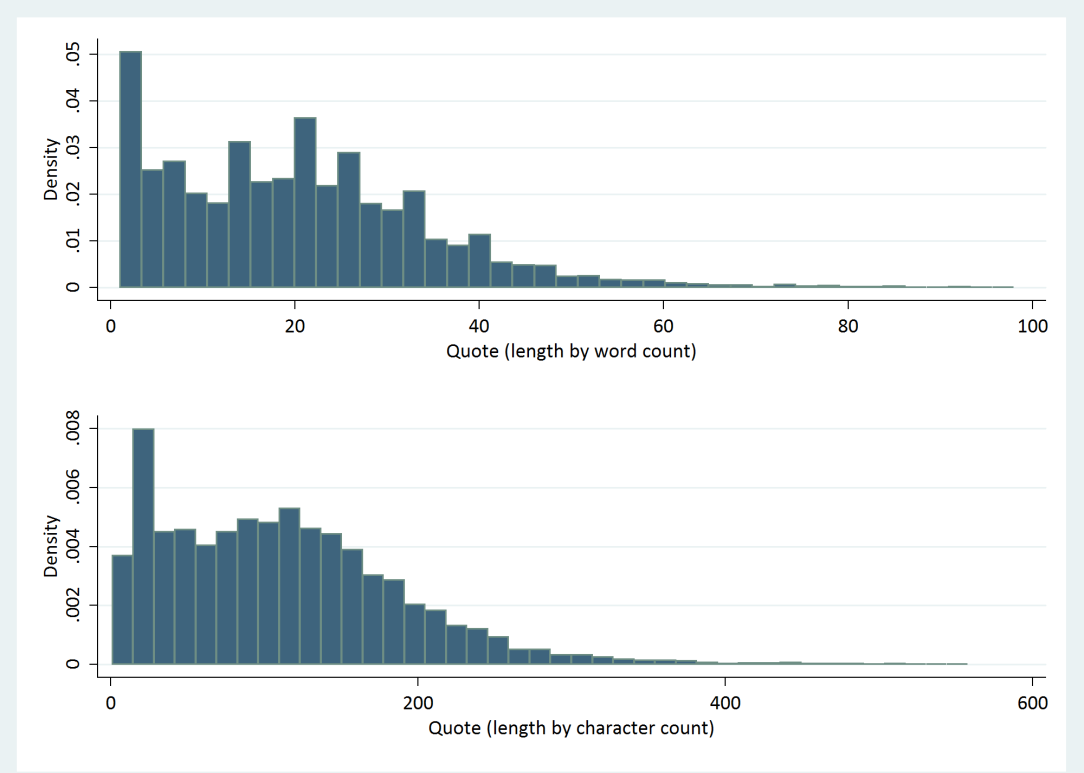


Figure A.9. Distribution of quote length

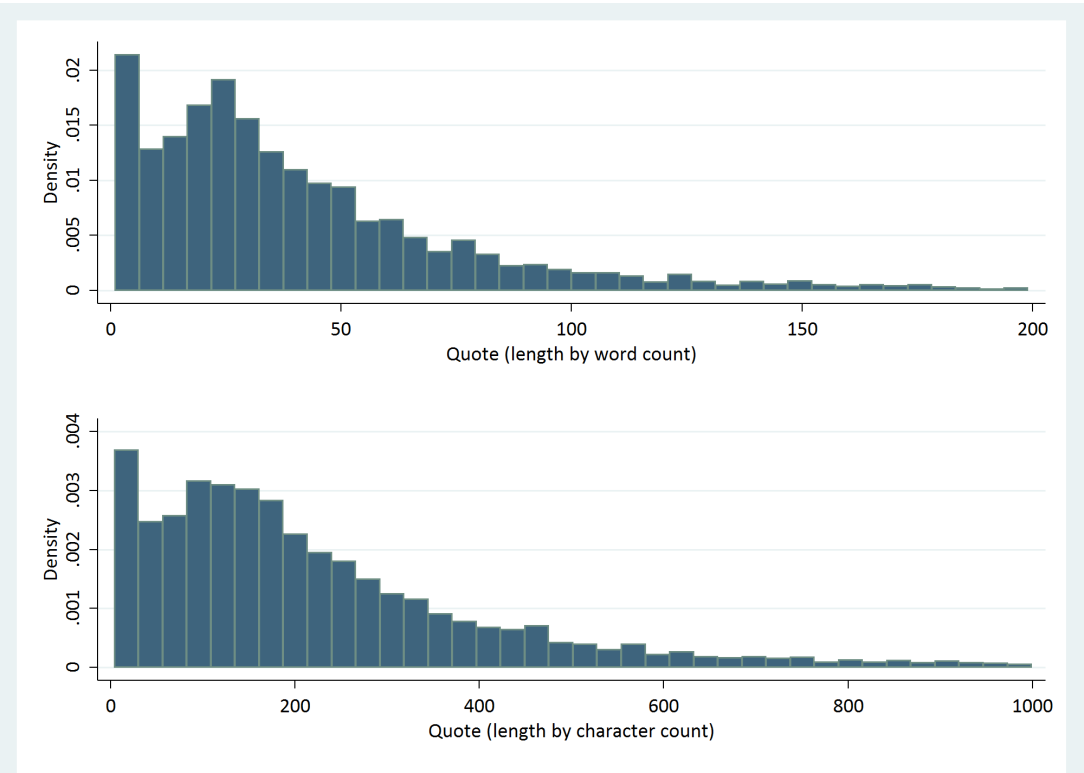


Figure A.10. Distribution of quote length at article-speech level

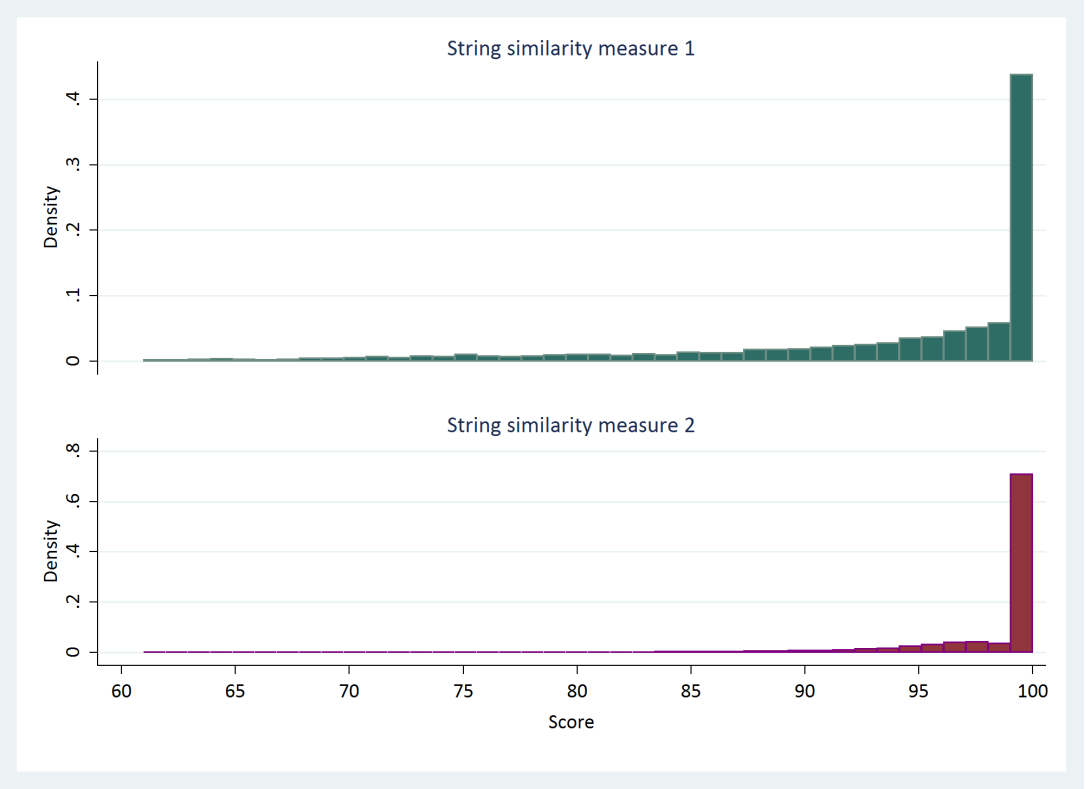


Figure A.11. Distribution of quote accuracy measures

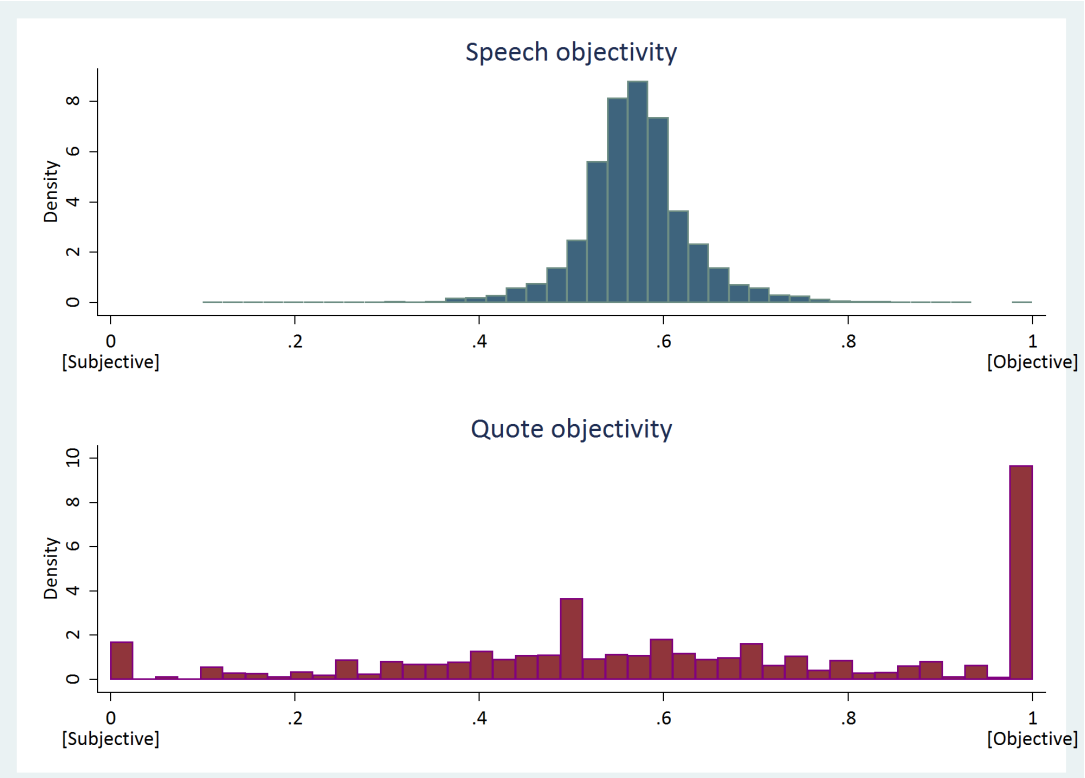


Figure A.12. Distribution of speech and quote objectivity

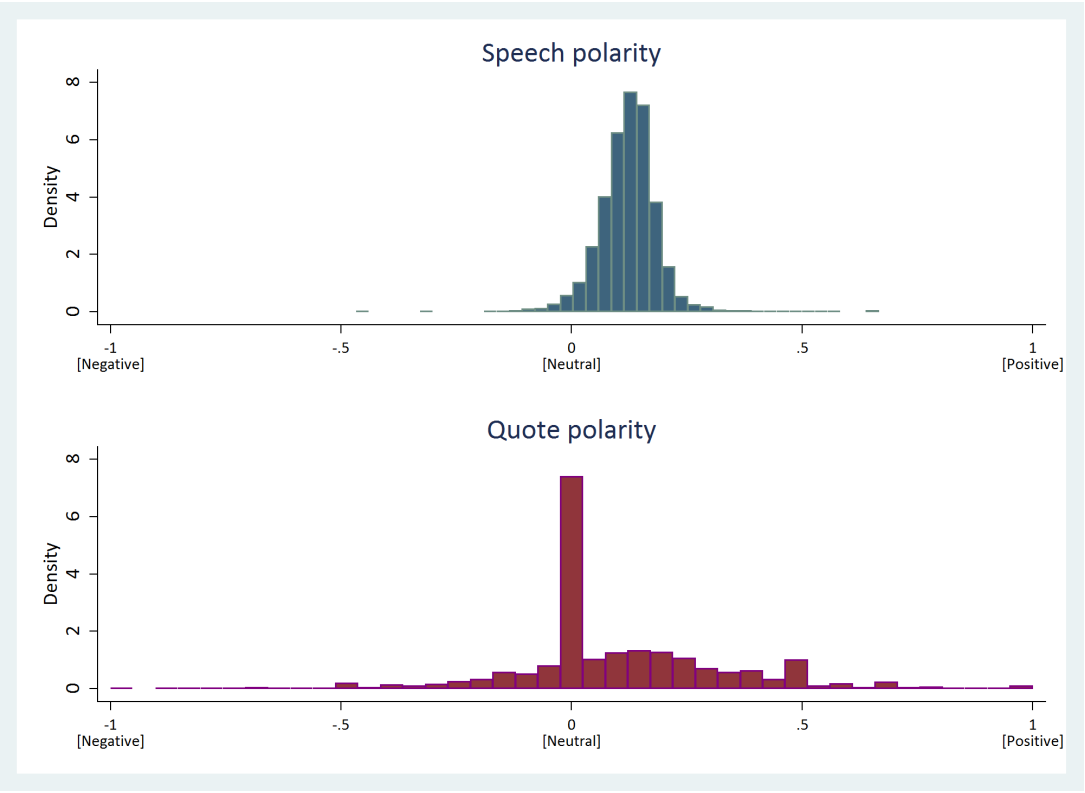


Figure A.13. Distribution of speech and quote polarity