London School of Economics
and Political Science

Department of Methodology

Essays in Political Text: New Actors, New Data, New Challenges

Tom Paskhalis

A thesis submitted to the Department of Methodology of the London School of Economics and Political Science for the degree of Doctor of Philosophy

September 17, 2020

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

 I declare that my thesis consists of approximately 50,500 words.

Statement of conjoint work

I confirm that Chapter 3 was jointly co-authored with Denisa Kostovicova (LSE), and I contributed 50% of this work.

Маме

Abstract

The essays in this thesis explore diverse manifestations and different aspects of political text. The two main contributions on the methodological side are bringing forward novel data on political actors who were overlooked by the existing literature and application of new approaches in text analysis to address substantive questions about them. On the theoretical side this thesis contributes to the literatures on lobbying, government transparency, post-conflict studies and gender in politics. In the first paper on interest groups in the UK I argue that contrary to much of the theoretical and empirical literature mechanisms of attaining access to government in pluralist systems critically depend on the presence of limits on campaign spending. When such limits exist, political candidates invest few resources in fund-raising and, thus, most organizations make only very few if any political donations. I collect and analyse transparency data on government department meetings and show that economic importance is one of the mechanisms that can explain variation in the level of access attained by different groups. Furthermore, I show that Brexit had a diminishing effect on this relationship between economic importance and the level of access. I also study the reported purpose of meetings and, using dynamic topic models, show the temporary shifts in policy agenda during this period. The second paper argues that civil society in post-conflict settings is capable of high-quality deliberation and, while differing in their focus, both male and female can deliver arguments pertaining to the interests of broader societal groups. Using the transcripts of civil society public consultation meetings across former Yugoslavia I show that the lack of gender-sensitive transitional justice instruments could stem not from the lack of women's

physical or verbal participation, but from the dynamic of speech enclaves and topical focus on different aspects of transitional justice process between genders. And, finally, the third paper maps the challenges that lie ahead with the proliferation of research that relies on multiple datasets. In a simulation study I show that, when the linking information is limited to text, the noise can potential occur at different levels and is often hard to anticipate in practice. Thus, the choice of record linkage requires balancing between these different scenarios. Taken together, the papers in this thesis advance the field of "text as data" and contribute to our understanding of multiple political phenomena.

Acknowledgements

First and foremost, I would like to thank my supervisors, Ken Benoit and Ben Lauderdale. Their personal and professional support for this research, as well as the entire academic path that I have stepped on, goes back to my days as a Masters student at the LSE. I am much indebted to Denisa Kostovicova, who has acted as a co-author, mentor and friend for most of my PhD journey. Arthur Spirling has kindly hosted me at NYU, an experience which proved to be consequential for my career, and for which I am very grateful.

This research would not have been possible without incredible support that I received from the Department of Methodology. Academically, I am particularly indebted to Jouni Kuha and Pablo Barberá. Anna Izdebska, Sam Scott and Esther Sidley have taken great care of the logistical side and I deeply appreciate it. The incredible community of the 7^{th} floor has also been a source of support and good cheer. And I was lucky to have great PhD colleagues who were always there to engage in technical discussions of statistical models or to go to a pub. In particular, I would like to thank Christian Müller, Chris Pósch, Imre Bárd and Thiago Oliveira.

Over my time at the LSE I made many friends from across the school who transformed these years and my world view. Olga Obizhaeva, Marina Nazarova, Ewa Batyra, Tze Ming Mok, Elena Pupaza, Kiwi Ting, Selina Hofstetter, Takuya Onoda, Jan Stuckatz, Tatiana Paredes, Ellie Suh, Toni Rodon, Sarah Jewett, Katharina Lawall, I could have hardly imagined meeting friends from across the world, who, by your own example, proved the universality of so many things about society and human nature. London can be both an exciting and challenging city at times. Without you, this would have been a much less

welcoming place to live in. Especially, I would like to thank Diego Alburez who has been there for me since my first days in London. I would also like to thank all my former colleagues and current friends from YPlan who exposed me to life outside academia.

I am grateful to the LSE for generously funding my studies and the Santander Postgraduate Travel Fund for supporting my research visit to NYU. Part of this research has also benefited from the support provided by the Leverhulme Trust and Arts and Humanities Research Council PaCCS.

Finally, and, most importantly, I would like to thank my mum, Olga Paskhalis, to whom this thesis is dedicated.

Contents

List of Figures

List of Tables

Chapter 1

Political Text and its Analysis

"In the beginning was the Word [...]"

—John 1:1

1.1 Content Analysis

The study of political text traces its origins back to the beginning of the 20th century. Although some authors (Krippendorff, 2004; Neuendorf, 2002) recognise the precursors to content analysis in the Biblical studies and decipherment of ancient languages, such as Egyptian hieroglyphs found on Rosetta Stone by Jean-François Champollion and Thomas Young (which has a decree issued by King Ptolemy V inscribed on it and thus, strictly speaking, falls in the category of political text), I will limit my overview to the developments that happened in social and computational sciences over the last century[1].

Analysis of newspapers and propaganda studies constitute the two major strands of research on political text in the first half of the century. Starting from crude physical measurement of space occupied by articles dedicated to a specific topic on a newspaper sheet, they evolved into simple classification schemes, such as attitudes towards the Neutrality Act (Allport and Faden, 1940), a major policy debate in pre-war America. It is worth noting that in the discussed period there were no rigid boundaries between different domains of social science, and psychological concepts such as attitude in the work by the eminent psychologist Gordon Allport above were frequently applied to public policies and the debates surrounding them. In fact, Harold Lasswell, a pioneer in the study of propaganda and, later, a major figure in the quantitative study of political text is equally acclaimed in political psychology and political science. In his earlier work *Propaganda Techniques in the World War* (Lasswell, 1927), Lasswell explored different propaganda strategies utilised by belligerents in the First World War. This analysis, though very detailed, was done in a rather haphazard manner, without due consideration to such issues as sampling and measurement, and was later criticised by the author himself (Lasswell

[1] Krippendorff (2004) provides an extended historical survey of content analysis.

et al., 1949). The push for a more systematic and quantifiable approach is the marker of all later Lasswell's works on content analysis. With the external political changes, the rise of Nazism and, eventually, with the commencement of the Second World War, the focus of propaganda analysis shifted from describing the strategies to the detection of propaganda in news broadcasts and political speeches[2] In the current statistical learning language this could be described as a shift from classification tasks to building prediction models, though the empirical methods remained unsophisticated. However, this period also brought some important analytical innovations, such as decoupling the message from the author, which led to Lasswell's communication model of *'who says what to whom in what channel and with which effect'* and also formalised such aspects of text analysis as sampling, validity and reliability.

The post-war period saw the integration and further refinement of the quantitative approaches to the analysis of political text (Berelson, 1952; Lasswell et al., 1949). But, most importantly, this is when first computers were introduced in text analysis. The pivotal moment came with the publication of the ground-breaking work *The General Inquirer: A Computer Approach to Content Analysis* by Stone et al. (1966). In addition to compiling extensive content analysis dictionaries and implementing software for mainframes to process large volumes of text, the authors provided a wide selection of case studies with applications ranging from anthropology to political science. Although tabulation of specific words had been applied prior to the General Inquirer, this was the first comprehensive treatment of the analytical categories underlying particular sets of words and a substantial step forward in dictionary methods. The authors made a distinction between an alphabetical enumeration of the terms of interest - 'dictionary' proper and assigning categorical tags to a word list, called 'thesaurus', a name borrowed from lexicography (Stone et al., 1966, pp. 135-139). While the former technique is better suited for looking at the meanings of specific concepts, the latter is useful when a researcher is not interested in the words *per se*, but rather in the analytical categories that they

[2]Perhaps, less directly than in the well-known case of the German tank problem (Ruggles and Brodie, 1947), sometimes this research helped forecast military operations and troops concentration.

define as a group. Furthermore, the General Inquirer was the earliest implementation of key-word-in-context (KWIC) approach (allowing the researcher to explore a given term amongst other words surrounding it in text), stemming (reducing the word to its stem by removing affixes) and automatic syntactic analysis (assigning part-of-speech tags to words) in social sciences. All of these features were implemented as extensions of an elaborate dictionary approach. As a whole, the dictionary approach received a lot of attention in political science and remains a method of choice for many empirical researchers today.

The proliferation of personal computers has made text analysis available to a wider scholarly community. Oftentimes methodological innovation stemmed not from the shifts in research focus, but from the evolution of text analytical programs and word processing software. From the earlier days of the General Inquirer, many software packages functioned as milestones in formalising approaches to working with textual data. WordStat (Péladeau, 1998), a content analysis module for SimStat, used inclusion and exclusion lists for words and categories, which later became known as 'stop-words' list or words deemed uninformative for the ensuing analysis[3]. Most new packages came with embedded wordcount and gradually new features such as readability scores (Klare, 1974) and concordance analysis (KWIC), were added. The metrics for text readability emerged in the educational context (Flesch, 1948), where children's reading assignments and standardised tests in schools had to be evaluated on the their difficulty and accessibility to different grade levels. There is a large number of different scores available[4], but most of them typically include three key components: (1) the sentence length (average or total, in words), (2) the length of words (average or total, in syllables or letters), (3) the number of words of certain type (pronouns, prepositions, difficult words from a pre-specified list), optionally, with applied weights and some combinations and ratios of these numbers. Although mostly confined to its original domains of educational research and psychology, readability scores have recently started receiving more attention in political science

[3]Although, the term 'go-words' list is found occasionally in the literature (Krippendorff, 2004), it was not nearly as widely adopted as 'stop-words' list.

[4]In his review article Klare (1974) provides a number of examples.

(Spirling, 2015; Benoit et al., 2019). For example, Spirling (2015) applied Flesch Reading Ease index to parliamentary speeches in Britain, showing that the Second Reform Act of 1867, which extended franchise to a much larger section of male population, mostly less educated, prompted the parliamentarians, especially those serving in the cabinet, to reduce the linguistic complexity of their speeches.

The development of methods that eventually came to be used in the analysis of political text was happening in parallel and often in disciplines that saw very little interaction with political science. One of the earlier examples of the applications of Markov chain models was the study of the sequences of vowels and consonants in the novel *Eugene Onegin* written by Alexander Pushkin (Markov, 1913). Although it didn't see any immediate successors outside statistics, through information theory the works by Shannon (1948, 1951) and Zipf (1949) it re-entered the study of language and social systems. Zipf's principal contribution, which became known as 'Zipf's law'[5], was that the product between a word's rank in the list sorted by frequency, multiplied by its frequency is, roughly, a constant number:

$$r \times f = C$$

Or, to put it differently, the word with rank 10 is expected to occur three times more often than the word with rank 30. Although not precisely, this relationship usually holds in practice. In the paper on record linkage I provide an illustration of Zipf's law using traditional text (the State of the Union addresses in the US), more idiosyncratic expression of textual data (organization names in the UK) and simulated labels. I use a generalization of Zipf's law proposed by Mandelbrot (1954) to simulate the dataset that exhibits the properties of real-world text.

Another important strand of, primarily, statistical work was the authorship attribution. Mosteller and Wallace (1963, 1983) apply Bayes' theorem to the *Federalist Papers*

[5]Zipf also introduced what he called 'the principle of least effort', as a tendency of humans to minimise the rate of work, that would affect both linguistic and non-linguistic behaviour, this theory had far less impact than his empirical observation of word frequencies.

with disputed authorship and identify Madison as the principal contributor. From the standpoint of classical quantitative text analysis, an interesting aspect of this work is the focus on function words over content words. The idea behind it is that personal writing style that can be used for correctly identifying the author is revealed through the usage of function words, such as prepositions (on, by, of) and conjunctions (while, although), rather than nouns or adjectives that tend to reflect content. Despite the book being, at its core, a statistical work, this theoretical idea of function words having a meaning of their own received further development.

Building upon the General Inquirer, Thematic Apperception Test (TAT) and psychoanalytical theory, Pennebaker and King (1999) developed a large dictionary, representing over a dozen psychological constructs, called 'Linguistic Inquiry and Word Count' (LIWC) (Pennebaker et al., 2001) and a complementary proprietary software for applying the extended version of the dictionary. Despite relative computational simplicity, the inclusion of a large number of theoretically interesting concepts, e.g. power, body and insight (Pennebaker et al., 2015), has led to its wide adoption among social scientists and beyond. Among other, political science applications of LIWC include profiling German party leaders based on their tweets (Tumasjan et al., 2010) and sentiment analysis of Congressional debates (Yu et al., 2008)[6]. The later theoretical framework, however, shifted to the role of function words (particularly, pronouns) in defining linguistic style and personality traits (Pennebaker, 2011). This makes the argument, bar psychological aspect, broadly resemble the one, made by Mosteller and Wallace half a century ago.

Despite being introduced more than a century ago, dictionary methods remain a highly popular method of choice in political science (Young and Soroka, 2012; Soroka et al., 2015; Proksch et al., 2019). While certainly not a go-to approach for every task, dictionary methods provide a useful and quick way of estimating pre-defined concepts from text. The possibility of accessible and flexible machine translation (Lucas et al.,

[6]As these examples suggest, it is also a dictionary of choice for computer scientists working on cross-disciplinary topics. One of the possible reasons for that is that it appears as a recommended lexicon, along with the General Inquirer, to boost training set in sentiment analysis tasks when the data is sparse in landmark natural language processing textbook (Jurafsky and Martin, 2009).

2015; de Vries et al., 2018; Proksch et al., 2019) has further provided an opportunity to extend the analysis to multiple languages (Paskhalis et al., 2019) through translation of the original corpus of documents or through translation of the off-the-shelf or curated dictionaries.

1.2 Classical Applications

As the central contribution of this thesis lies in expanding the range of political actors whose output is analysed using quantitative methods of text analysis, it is helpful to review the classical lines of research. Until recently, those, almost exclusively, focused on parties and legislators. While, undoubtedly very important, they are hardly the only political actors whose actions are consequential for democratic polities. In this work I argue and show empirically that other political actors, such as interest groups and members of the civil society leave behind a trove of data that can be successfully used to generate new insights. Furthermore, I argue that apart from ideological scaling, text analysis can contribute to other important questions about political actors, such as their policy agenda and the state of deliberative democracy. However, moving forward requires careful consideration of the methodological state of the field as well as the advances and mistakes made in prior application of 'text-as-data' approaches.

The single most important frontier in applying text analysis to political text[7] can be described as scaling public policy positions of political actors[8]. The longest and perhaps the largest research project undertaken in political science, the Manifesto Project (or Manifesto Research on Political Representation, MARPOR[9]) (Budge et al., 1987, 2001; Klingemann et al., 2006; Volkens et al., 2013) was designed to cover all free, democratic elections after the Second World War. As the project's name suggests, the key data source

[7] Laver (2014) reviews the application of text analysis as well as other methods to infer policy position.

[8] The theoretical roots of assigning ideal points to political actors on a latent policy dimension lie in seminal economic models of spatial competition (Hotelling, 1929; Downs, 1957).

[9] MARPOR (since 2009) was also previously known as the Manifesto Research Group (MRG) (1979-1983) and the Comparative Manifestos Project (CMP) (1983-2009) and one still comes across these former names in the literature.

for inferring policy positions were party manifestos, issued prior to the parliamentary elections and stating party vision and goals. By being the principal political organizations in democratic systems political parties and their policy positions are of immense importance for both political scientists and voters. Although the precise contractual nature of manifestos is often violated, when parties fail to implement the promised policies, the pledges made in them constitute an important accountability mechanism. The centrality of manifesto promises is most conspicuous in the case of British politics (Kavanagh, 1981). Although this position has been criticised on both theoretical and normative grounds, the amount of attention they receive in academic literature and journalistic reports has never receded. The critics suggested that, first, the bureaucratic model of government suggests that civil servants might have upper hand when it comes to policy delivery over legislators. Second, coalition governments, more frequent in other democratic systems can prevent the promised policies from being implemented. And, third, putting too much emphasis on adversarial party manifestos can promote partisanship, societal fractures and disregard for opposition.

In reality the end goal of public policy scaling is often elusive. The central tenet of every methodological approach is that it is position that is being measured by a given technique. However, the mechanics of text analysis usually involves counting units of choice (be it word, sentence or another category), that give how much emphasis a party puts on specific issue. In Manifesto Project this ingrained controversy has been overcome by the development of saliency theory (Budge et al., 1987, pp.24-28). The core assumption which allows to bridge the gap between saliency and position is that emphasis corresponds to preference and stressing particular issue is indicative of the support for it. Another way to look at it is that parties tend to avoid direct confrontation with other parties and rather than offer different views on the same agenda are more likely to conceal their views on the issues that are less important to them and might prove to be unfavourable with voters. It is worth noting that this theoretical shift in party competition literature from confrontation theory (Robertson, 1976) to saliency theory resulted from a combination

of analytical considerations and methodological caveats. Subjecting saliency theory to independent testing, Dolezal et al. (2014) find support for some of its implications. While parties indeed tend to avoid mentioning each other directly, the level of issue convergence is higher than it is predicted by the theory. The authors' empirical analysis of party manifestos in Austria shows that parties tend to compete over the same issue and engage in direct confrontation over them.

Text analysis constitutes the core of the MARPOR methodology. And this part of the research project attracted most criticism from other scholars. All party manifestos are analysed after manually coding by a single coder into 56 issue categories[10]. This follows the order normally used in manual text analysis, with text (1) being unitised, divided into basic units of analysis; (2) coded by a human coder and (3) aggregated into single score. The first step a coder takes is unitises them in 'quasi-sentences', the basic unit of coding in the MARPOR project. The choice of this unit over others such as word, paragraph or entire text was driven by the realisation that long natural sentences can contain several political ideas (Budge et al., 2001, pp.93-107). This segmentation of grammatical sentences into conceptual quasi-sentences is inherently subjective and has been shown to be a source of inter-coder unreliability, while offering no benefits over splitting text at a pre-defined set of punctuation marks that would produce a more linguistically-sound unit of coding (Däubler et al., 2012). In addition, Dolezal et al. (2016) recently suggested using 'kernel sentences', based on Chomsky's (Chomsky, 1957) syntactic model, as a unit of analysis for manifesto data. This approach, however, still relies on human coders manually dividing the sentences into kernel sentences, albeit the rules are better grounded in linguistic theory and can be expected to be less subjective. After unitising, a coder assigns one code out of the 56-category coding scheme (further subdivided into 7 domains such as external relations and fabric of society) to each quasi-sentence. As the project includes manifestos in multiple languages it makes it very costly to code any document twice. In a coding experiment, Mikhaylov et al. (2012) show that together with complex

[10]The number of issue categories was expanded from initial 21 to 56 categories today.

coding scheme this produces systematic misclassification. Furthermore, coding by a single coder results only in ideal point estimates without any measures of uncertainty (Benoit et al., 2009). This limitation can be partially overcome by bootstrapping confidence intervals, but doing so affects many substantial conclusions as a number of party 'movements' along the left-right scale can be attributed to stochastic noise in textual data rather than actual changes in party position. After the coding is complete, the researchers calculate a single score on a left-right scale, with RILE being the most frequently used for this dataset. As some of the coded categories are neutral, for computing point estimates of party position (θ) the percentages of 13 left categories are summed up and subtracted from the equivalent summation of the percentages of 13 right categories:

$$\theta = \frac{R - L}{N}$$

The choice of summative index assumes constant marginal effect of each additional quasi-sentence and has been criticised by Lowe et al. (2011). Instead, the authors argued for the adoption of log odds-ratios, which would reflect decreasing marginal effect. Put differently, the more pro-right quasi-sentences have been read, the less would be the assumed effect of each additional on the voter and this should be reflected in a shrunken left-right scale. The foundation for this aggregation method would be a well-known psychological regularity, namely, Weber-Fechner's law (Fechner, 1965), which states that equal relative increments of the strength of perceptual stimuli lead to equal changes in sensation. It is reasonable to believe that audial or visual perception of political text would not fall far from other sensory modalities. Despite all the methodological concerns raised by other scholars, only very few of them were adequately addressed by the original MARPOR authors (Volkens et al., 2013). This is rather unfortunate for the advancement of the discipline as manifesto dataset continues to be widely used in comparative politics by researchers who prefer a well-established source for party positions without the need to get bogged down in abstruse methodological debate.

The first attempt to automate coding of party manifestos came with the creation of dictionary by Laver and Garry (2000). This dictionary was based on a new hierarchically-structured coding scheme with simplified policy areas. In this scheme all categories had their antithesis as well as neutral and British 1992 Conservative and Labour manifestos were used to get a pool of words for populating the dictionary. After having 1992 and 1997 manifestos from the UK and Ireland coded by experts, they ran the computer program that calculated left-right score using just the raw word counts and the dictionary. Though the correlation between the two varied across policy areas, face validity was shown when this method correctly identified the key perceived party movement of those electoral campaigns - Liberal Democrats shifting to the left of Labour in 1997. Additionally, the authors contend that the resultant estimates reflect positions rather than issue saliency, however, this opinion was not widely accepted. Overall, this application of dictionary method was an improvement over manual coding, but it still required a rather laborious development of bespoke dictionary, as well as its manual population with words. Implicitly, in this paper as well as other in other applications of dictionary method, researchers adopt 'bag-of-words' approach. The idea is rather simple and straightforward to implement, but conceptually goes against what most people would consider to be linguistically sound. It effectively states that text can be represented as word frequency matrix, entirely leaving out syntax and structure. Dictionary method, in effect, calculates frequencies of pre-specified words and then aggregates them by category. 'Bag-of-words' assumption, however, goes beyond dictionary approaches. It underpins most contemporary text analysis application, including, but not limited to, Wordscores, Wordfish and IRT-type models discussed below, topic models and word embeddings. While not uncontroversial, the empirical research presented in this thesis relies extensively on independence of words within label, sentence and the entire text.

After dictionary method the next logical step was to predict frequencies of all words, possibly, excepting 'stop-words'. This was the approach adopted in Wordscores method developed by Laver et al. (2003). By computing word frequencies of 'virgin' texts, relative

to 'reference' texts with known scores, it assigned them a position on a pre-defined scale. The key feature of this approach, which could also be referred to as supervised scaling[11], is the need to have a set of texts with known positions (training set). It drastically reduced the analysis time, but as with all supervised methods, the requirement to define training set *a priori* can constitute a hurdle if the corpus is new and unexplored or the researcher is dealing with a language other than English, where pre-scored texts might not be readily available. Another, more subtle, assumption is that reference texts contain all the words that are relevant to the position which is being extracted. Wordscores has been criticised by the MARPOR authors (Budge and Pennings, 2007) for the seeming arbitrariness of the choice of reference texts. Lowe (2008), while giving a generally positive assessment and drawing parallels between Wordscores and correspondence analysis, also notes the problem with the lack of underlying statistical model and the inherent need for applying transformation to get interpretable estimates[12].

The next key advancement in automated text analysis came with the adoption of unsupervised scaling (or learning in computer science literature) model instead of supervised learning used in Wordscores. This was the idea developed by Slapin and Proksch (2008) in their Wordfish[13] model. Instead of using reference texts, Wordfish assumes a statistical model that generated the observed word counts. The model draws upon literary style literature (Peng and Hengartner, 2002) and follows above-mentioned work by Mosteller and Wallace (1963). The model is based on Naive Bayes assumption, frequently encountered in text analysis literature. It states that words are distributed independently of each other. In other words, the usage of one word by a text author does not increase or decrease the probability of the usage of any other word from a sample space. Al-

[11] As Benoit and Nulty (2013) show, Wordscores is mathematically equivalent to Naive Bayes classifier, with the principal difference that the former uses individual word-level posterior probabilities to construct an additive scale, while the latter predicts the class using them multiplicatively in the joint probability model.

[12] Martin and Vanberg (2008) provide an extended discussion of several procedures for Wordscores transformation.

[13] The name 'Wordfish' might appear mysterious, but it comes from the English translation of the surname of French statistician Poisson, whose name was given to the distribution used in the underlying statistical model.

though this assumption is clearly violated in all real texts (hence, its 'naivity'), it has been shown to perform well in many text classification tasks, such as spam detection (Sahami et al., 1998). The model is, essentially, a Poisson word count model, similar to the one used in the unpublished manuscript by Monroe and Maeda (2004), with such parameters as party-year(α_{it}) and word (ψ_j) fixed effects, as well as party position (ω_{it}) and word-discriminating parameter (β_j):

$$y_{ijt} \sim Poisson(\lambda_{ijt})$$

$$\lambda_{ijt} = exp(\alpha_{it} + \psi_j + \beta_j * \omega_{it})$$

As is clear from the model specification, the entire right-hand side is unknown and has to be estimated with the expectation maximization (EM) algorithm. The main criticism addressed at Wordfish, and, potentially, other unsupervised scaling models is the assumption that ideology dominates analysed texts. Otherwise, as Grimmer and Stewart (2013) point out, the model will seize upon the primary source of variation across texts. In practice this assumption is only met if the manifestos represent an exhaustive statement of party platform. If, however, a policy position on particular dimension is of interest, Slapin and Proksch (2008) suggest that Wordfish is applied only to corresponding sections of manifesto.

In the pursuit of better reliability and reproducibility (King, 1995) of human coding, researchers have turned to crowd-sourcing platforms. Rather than hiring a handful of coders and subjecting them to rigorous training before they master the coding scheme and can tag units of text analysis with high inter-coder agreement, scholars started exploring the potential of online platforms that match employers to workers, who, in turn, do relatively simple human intelligence tasks, such as tagging pictures, answering surveys and taking part in experiments. This approach has been shown to perform well and be more representative than often-used convenience sampling (Berinsky et al., 2012). Benoit et al. (2016) used Crowdflower to classify sentences with a coding scheme simplified

from 56-category into two scales: economic policy (left-right) and social policy (liberal-conservative). The results show high agreement between experts and crowd-sourced workers judgement about the placement of manifestos on both scales and provide evidence that human judgement and replicable research do not have to be mutually exclusive. Another prospective avenue for further refinement of scaling policy preferences from textual data is extending statistical models. The body of research on policy positions shows that there is some potential in adopting item response theory models to political text (Benoit et al., 2016). Item response theory (IRT), which originated in psychometric studies of human intelligence, has been used prominently in political science for establishing policy positions. At first, it proved to be a highly consistent method for analysing roll call votes in US Congress (Poole and Rosenthal, 1985; Clinton et al., 2004). But later its usage spread to surveys (Bafumi and Herron, 2010), campaign contributions (Bonica, 2013) and Twitter followership networks (Barbera, 2015). Currently, there is also ongoing work on incorporating IRT in text analysis of party manifestos (Däubler and Benoit, 2017). In the chapter on civil society deliberation in post-conflict settings, we adopt IRT model to measuring the quality of deliberation using human-coded speech acts as units of analysis. While we find no difference across genders, the primary focus of our analysis, the varying difficulty and discrimination parameters of individual components (items) provide interesting insight into how hard it is to meet the normative aspects of deliberation in highly contentious environments.

No less important than the evolution of tools for the analysis of party manifestos was the spread of those methods to other sections of political science. In their application of Wordfish scoring to the US Senate and Irish Dail, Lauderdale and Herzog (2016) extend it to build Wordshoal model by subjecting the derived debate-specific point estimates for each speaker to Bayesian factor analysis to discover underlying policy dimensions with speakers placed on these latent scales. Wordfish model has also been successfully applied in the analysis of speeches in the European Parliament (Proksch and Slapin, 2010) and consultation submitted by lobbying groups (Klüver, 2009).

Overall, the body of literature that emerged from measuring policy positions of parties and legislators using textual data provides a robust foundation for extending the types of political actors analysed, as well as bringing in new research questions and new data to address them. In the empirical chapters below I will show that some of the models discussed above can be adopted for analysing data on interest group meetings and civil society discussions.

1.3 New Frontiers

Increasingly, the analysis of political text starts incorporating methods developed in natural language processing[14] and machine learning[15]. Most types of problems that machine learning addresses, can be, roughly, divided into two broad categories: classification and prediction. While the latter is only infrequently a concern for political scientists[16], classification in general and text classification in particular is a problem, often faced by researchers. As the technical entry barrier is often high for social scientists, it is unsurprising that some of the earlier applications of machine learning methods to political questions were done by computer scientists. For example, Yu et al. (2008) applied support vector machine (SVM) and Naive Bayes classifiers to speeches in Congress in order to infer party affiliation of the speakers. Drutman and Hopkins (2013) applied hand coding and SVM to the Enron dataset, that includes over 250,000 internal emails that were released following the corporate investigation, to uncover the key lobbying strategies that were pursued by company's employees. Often scholars do not want to classify each individual document as the universe of certain texts is hypothetically infinite, but rather learn about the proportions of texts of each pre-determined category in the population (Hopkins and King, 2010). Or, conversely, cluster the existing texts in groups that might

[14] Jurafsky and Martin (2009) offer a classic introduction into the field of natural language processing.

[15] Grimmer and Stewart (2013) review the currently available methods of automated text analysis for political scientists with a focus on machine learning and information retrieval methods. Wilkerson and Casas (2017) provide the most recent overview of the state of computerised text analysis

[16] The only obvious exception to that is electoral forecasting, which is a vibrant research area on both sides of the Atlantic (Gelman and King, 1993; Hanretty et al., 2016)

prove to be a new, conceptually interesting, categorisation scheme (Grimmer and King, 2011). These two classification approaches correspond to what is known as supervised and unsupervised learning models in statistical and computer science literature (Hastie et al., 2009). It is also sometimes referred to as computer-assisted clustering (CAC) and fully automated clustering (FAC) in their applications to political texts (Grimmer and King, 2011; Grimmer and Stewart, 2013).

Despite the relatively large selection of new methods for automated or semi-automated text analysis, many of them did not see wide adoption. If this criterion is used to judge the success of a particular method, topic modelling (Blei et al., 2003) is perhaps the most prominent approach in the field of text analysis of those introduced more recently. The roots of this method lie in latent semantic indexing (LSI), a technique developed by Deerwester et al. (1990) to improve the performance of information retrieval systems by deriving latent semantic structure of a set of documents. This is achieved by using singular-value decomposition to estimate a reduced number (50-100 for thousands of words and documents) of orthogonal factors or dimensions and then treating those as latent semantic space to calculate document and word similarity. Hofmann (1999) extends the model by introducing a probabilistic model to the, otherwise, linear algebra transformation. Latent Dirichlet Allocation (LDA), developed by Blei et al. (2003) is a full Bayesian implementation of the probabilistic LSI. Another way to look at is as a hierarchical Bayesian mixture model, where each document i is a finite mixture of underlying topics, with proportions π_i drawn from Dirichlet prior, each topic k comes from an infinite mixture of topic probabilities with multinomial prior and, conditional on the topic, a word w_{ij} is drawn from a multinomial prior:

$$\pi_i \sim Dirichlet(\alpha)$$

$$\tau_k \sim Multinomial(1, \pi_i)$$

$$w_{ij} \sim Multinomial(1, \theta_{ij}|\tau_k)$$

Although it is possible to design a different topic model, based on different priors, LDA remains by far the most widely used. Some of the generalizations of the LDA include hierarchical Dirichlet processes (Teh et al., 2006), where the number of topics can be learnt from data and, more specific to political science, structural topic model (Roberts et al., 2014), where prior distribution for words is expanded to include covariates. This set-up allows to test for significant differences in word distributions conditioning on other parameters of interest. In the first empirical chapter I use a variant of dynamic topic models (Blei and Lafferty, 2006) to explore the evolution of policy agenda from the meetings between the government and interest groups in the UK. In the second empirical chapter we apply structural topic models to estimate the differences in thematic focus across genders.

Most recently, political scientists started adopting deep learning approaches and word embeddings to relax the traditional 'bag-of-words' assumption (Rheault and Cochrane, 2019; Spirling and Rodriguez, 2019). Recall that the main premise underpinning most text analysis models is that words or, rather, wordcounts can be considered in isolation, irrespective of their context. Instead, to estimate embeddings a window around the input word of some size is specified and both the counts of the the word itself as well as the counts of the words occurring within this windows are used as the inputs in the neural network model[17]. As the output, each word can be represented as a vector in some n-dimensional space (embedding), where proximity between the words reflects their semantic similarity in the training corpus[18]. However, despite their promise, producing new insights into the political science phenomena using word embeddings remains a task of the future research.

[17] Rheault and Cochrane (2019) provide a longer treatment of word embeddings in the context of political science

[18] This provides some scope for vector arithmetic as in the classical example of $king + woman = queen$ from Mikolov et al. (2013), the original developers of this approach.

1.4 Conclusions

In the past decade quantitative text analysis has grown out of niche methodology, largely restricted to ideological scaling, into broad and diverse approach to working with any textual data, be it legislative speeches, newspaper articles, public consultations or social media stories. While the range of political actors who are studied using this approach has been gradually growing over the years, there are still many research questions that have seen limited application of text analysis.

The three main challenges that I outline and tackle in this thesis are (1) data challenge, (2) language(s) challenge and (3) computation challenge. Despite the decreasing costs of data collection, obtaining high-quality textual data remains a challenge. Oftentimes, the data remains locked in non-machine-readable format that requires optical-character recognition or scattered over many pages of dynamically generated website that requires writing a sophisticated scraper. What remains even more challenging is ensuring the completeness and consistency of data collection. In the first paper of this thesis I make an attempt to address this challenge by implementing two open-source software packages to assist in compiling data government transparency reports in the UK. Secondly, apart from several notable exceptions[19] most text analysis have been tried and applied to the more common Western languages, such as English and German. At least in part this can be explained by more easily accessible datasets, such as party manifestos or US congress speeches. However, the accessibility of text analysis for other language often varies a lot. This problem can be partially addressed by adopting computational tools from natural language processing. In the second paper of this thesis I use such tools specially designed for Balkan languages to facilitate pre-processing and text analysis in Serbian. And, finally, despite the rapid increase in computational power available to empirical researchers, some applications still require making trade-offs. This is especially true when working with large datasets that do not fit into computer memory. In the third paper I illustrate some of the

[19]See King et al. (2013), for example of text analysis applications in Chinese and Rozenas and Stukal (2019) for text analysis in Russian.

trade-offs involved when linking multiple datasets that contain only textual labels as the common identifier.

Altogether, this thesis identifies the challenges for current applications of text analysis in political science and outlines a few potential avenues of addressing them. Furthermore, it contributes to theoretical discussions in such fields as British politics, lobbying and post-conflict studies.

Bibliography

Allport, G. W. and J. M. Faden (1940). The Psychology of Newspapers: Five Tentative Laws. *Public Opinion Quarterly 4*(4), 687–703.

Bafumi, J. and M. C. Herron (2010). Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress. *American Political Science Review 104*(3), 519–542.

Barbera, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis 23*(1), 76–91.

Benoit, K., D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov (2016). Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data. *American Political Science Review 110*(2), 278–295.

Benoit, K., M. Laver, and S. Mikhaylov (2009). Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. *American Journal of Political Science 53*(2), 495–513.

Benoit, K., K. Munger, and A. Spirling (2019). Measuring and Explaining Political Sophistication Through Textual Complexity. *American Journal of Political Science 63*(2), 491–508.

Benoit, K. and P. Nulty (2013). Classification Methods for Scaling Latent Political Traits. Working paper.

Berelson, B. (1952). *Content Analysis in Communications Research*. New York: Free Press.

Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis 20*(3), 351–368.

Blei, D. M. and J. D. Lafferty (2006). Dynamic Topic Models. In *International Conference on Machine Learning*, pp. 113–120.

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

Bonica, A. (2013). Ideology and Interests in the Political Marketplace. *American Journal of Political Science 57*(2), 294–311.

Budge, I., H.-D. Klingemann, A. Volkens, J. Bara, and E. Tanenbaum (2001). *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998*. Oxford: Oxford University Press.

Budge, I. and P. Pennings (2007). Do they work? Validating computerised word frequency estimates against policy series. *Electoral Studies 26*(1), 121–129.

Budge, I., D. Robertson, and D. Hearl (1987). *Ideology, Strategy, and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press.

Chomsky, N. (1957). *Syntactic Stucture*. The Hague: Mouton & Co.

Clinton, J. D., S. Jackman, and D. Rivers (2004). The Statistical Analysis of Roll Call Data. *American Political Science Review 98*(2), 355–370.

Däubler, T. and K. Benoit (2017). Estimating Better Left-Right Positions Through Statistical Scaling of Manual Content Analysis. Working Paper.

Däubler, T., K. Benoit, S. Mikhaylov, and M. Laver (2012). Natural Sentences as Valid Units for Coded Political Texts. *British Journal of Political Science 42*(4), 937–951.

de Vries, E., M. Schoonvelde, and G. Schumacher (2018). No Longer Lost in Translation. Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis 26*(4), 417–430.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science 41*(6), 391–407.

Dolezal, M., L. Ennser-Jedenastik, W. C. Müller, and A. K. Winkler (2014). How parties compete for votes: A test of saliency theory. *European Journal of Political Research 53*(1), 57–76.

Dolezal, M., L. Ennser-Jedenastik, W. C. Müller, and A. K. Winkler (2016). Analyzing Manifestos in their Electoral Context A New Approach Applied to Austria, 2002–2008. *Political Science Research and Methods 4*(3), 641–650.

Downs, A. (1957). *An Economic Theory of Democracy*. New York: Harper Collins.

Drutman, L. and D. J. Hopkins (2013). The inside view: Using the Enron E-mail archive to understand corporate political attention. *Legislative Studies Quarterly 38*(1), 5–30.

Fechner, G. (1965). Elements of Psychophysics. Sections VII and XVI. In R. J. Herrnstein and E. G. Boring (Eds.), *A Source Book in the History of Psychology*, pp. 66–75. Cambridge: Harvard University Press.

Flesch, R. (1948). A New Readability Yardstick. *The Journal of Applied Psychology 32*(3), 221–233.

Gelman, A. and G. King (1993). Why Are American Presidential Election Campaign Polls so Variable When Votes Are so Predictable? *British Journal of Political Science 23*(4), 409–451.

Grimmer, J. and G. King (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences of the United States of America 108*(7), 2643–2650.

Grimmer, J. and B. M. Stewart (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*(3), 267–297.

Hanretty, C., B. E. Lauderdale, and N. Vivyan (2016). Combining national and constituency polling for forecasting. *Electoral Studies 41*, 239–243.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In *Uncertainity in Artifitial Intelligence - UAI'99*, pp. 289–296.

Hopkins, D. J. and G. King (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science 54*(1), 229–247.

Hotelling, H. (1929). Stability In Competition. *The Economic Journal 39*(153), 41–57.

Jurafsky, D. and J. H. Martin (2009). *Speech and language processing* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Kavanagh, D. (1981). The Politics of Manifestos. *Parliamentary Affairs 1*(7), 7–27.

King, G. (1995). Replication, Replication. *PS: Political Science and Politics 28*(3), 444–452.

King, G., J. Pan, and M. Roberts (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review 107*(917), 326–343.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly 10*(1), 62–102.

Klingemann, H.-D., A. Volkens, J. Bara, I. Budge, and M. D. McDonald (2006). *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990–2003*. Oxford: Oxford University Press.

Klüver, H. (2009). Measuring Interest Group Influence Using Quantitative Text Analysis. *European Union Politics 10*(4), 535–549.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology* (2nd ed. ed.). Sage Publications.

Lasswell, H. D. (1927). *Propaganda Techniques in the World War*. New York: Knopf.

Lasswell, H. D., N. Leites, and Associates (1949). *Language of Politics: Studies in Quantitative Semantics*. Cambridge: MIT Press.

Lauderdale, B. E. and A. Herzog (2016). Measuring Political Positions from Legislative Speech. *Political Analysis 24*(3), 374–394.

Laver, M. (2014). Measuring Policy Positions in Political Space. *Annual Review of Political Science 17*(1), 207–223.

Laver, M., K. Benoit, and J. Garry (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review 97*(2), 311–331.

Laver, M. and J. Garry (2000). Estimating Policy Positions from Political Texts. *American Journal of Political Science 44*(3), 619–634.

Lowe, W. (2008). Understanding Wordscores. *Political Analysis 16*(4), 356–371.

Lowe, W., K. Benoit, M. Slava, and M. Laver (2011). Scaling Policy Preferences from Coded Political Texts. *Legislative Studies Quarterly 36*(1), 123–155.

Lucas, C., R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis 23*(2), 254–277.

Mandelbrot, B. (1954). Structure Formelle des Textes et Communication [Formal Structure of Texts and Communication]. *Word 10*(1), 1–27.

Markov, A. A. (1913). Primer statisticheskago izsledovaniya nad tekstom "Evgeniya Onegina" illustriruyushego svyaz' ispytanyi v tsepi [An example of statistical investigation of the text 'Eugene Onegin' illustrating the connection of trials in chain.]. *Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg 7*(3), 153–162.

Martin, L. W. and G. Vanberg (2008). A Robust Transformation Procedure for Interpreting Political Text. *Political Analysis 16*(1), 93–100.

Mikhaylov, S., M. Laver, and K. Benoit (2012). Coder Reliability and Misclassification in the Human Coding of Party Manifestos. *Political Analysis 20*(1), 78–91.

Mikolov, T., G. Corrado, K. Chen, and J. Dean (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*.

Monroe, B. L. and K. Maeda (2004). Talk's cheap: Text-based estimation of rhetorical ideal-points. In *Annual meeting of the Society for Political Methodology*, pp. 29–31.

Mosteller, F. and D. L. Wallace (1963). Inference in an Authorship Problem. *Journal of the American Statistical Association 58*(302), 275–309.

Mosteller, F. and D. L. Wallace (1983). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer.

Neuendorf, K. (2002). *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage.

Paskhalis, T., B. Rosenfeld, K. Tertytchanaya, and K. Watanabe (2019). Indepedent Media in Electoral Autocracies. *Working Paper*.

Péladeau, N. (1998). WordStat: Content Analysis Module for SimStat.

Peng, R. D. and N. W. Hengartner (2002). Quantitative Analysis of Literary Styles. *The American Statistician 56*(3), 175–185.

Pennebaker, J. W. (2011). *The Secret Life of Pronouns: What Our Words Say About Us*. New York: Bloomsbury Press.

Pennebaker, J. W., R. L. Boyd, K. Jordan, and K. Blackburn (2015). The development and psychometric properties of LIWC2015.

Pennebaker, J. W., M. E. Francis, and R. J. Booth (2001). *Lingusitic Inquiry and Word Count LIWC2001*. Mahwah, NJ: Erlbaum Publishers.

Pennebaker, J. W. and L. A. King (1999). Language Use as an Individual Difference. *Journal of Personality and Social Psychology 77*(6), 1296 – 1312.

Poole, K. T. and H. Rosenthal (1985). A Spatial Model for Legislative Roll Call Analysis. *American Journal of Political Science 29*(2), 357–384.

Proksch, S. O., W. Lowe, J. Wäckerle, and S. Soroka (2019). Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly 44*(1), 97–131.

Proksch, S.-O. and J. B. Slapin (2010). Position Taking in European Parliament Speeches. *British Journal of Political Science 40*(03), 587–611.

Rheault, L. and C. Cochrane (2019). Word Embeddings for the Estimation of Ideological Placement in Parliamentary Corpora. *Political Analysis* (First View).

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science 58*(4), 1064–1082.

Robertson, D. (1976). *A Theory of Party Competition*. New York: Wiley.

Rozenas, A. and D. Stukal (2019). How Autocrats Manipulate Economic News: Evidence from Russia's State-Controlled Television. *Journal of Politics 81*(3), 982–996.

Ruggles, R. and H. Brodie (1947). An Empirical Approach to Economic Intelligence in World War II. *Journal of the American Statistical Association 42*(237), 72–91.

Sahami, M., S. Dumais, D. Heckerman, and E. Horvitz (1998). A Bayesian approach to filtering junk e-mail. In *AAAI Workshop on Learning for Text Categorization*, pp. 55–62.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal 27*(3), 379–423.

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal 30*(1), 50–64.

Slapin, J. B. and S.-O. Proksch (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science 52*(3), 705–722.

Soroka, S. N., D. A. Stecula, and C. Wlezien (2015). It's (Change in) the (Future) Economy, Stupid: Economic Indicators, the Media, and Public Opinion. *American Journal of Political Science 59*(2), 457–474.

Spirling, A. (2015). Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832-1915. *The Journal of Politics 78*(1), 235–248.

Spirling, A. and P. L. Rodriguez (2019). Word Embeddings: What works, what doesn't, and how to tell the difference for applied research. *Working paper*.

Stone, P. J., D. C. Dunphy, and M. S. Smith (1966). *The General Inquirer: A Computer Approach to Content Analysis.* Cambridge: The MIT Press.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

Tumasjan, A., T. Sprenger, P. Sandner, and I. Welpe (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178–185.

Volkens, A., J. Bara, I. Budge, M. D. McDonald, and H.-D. Klingemann (2013). *Mapping Policy Preferences from Texts III: Statistical Solutions for Manifesto Analysts*. Oxford: Oxford University Press.

Wilkerson, J. D. and A. Casas (2017). Large-scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science 20*, 529–544.

Young, L. and S. Soroka (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication 29*(2), 205–231.

Yu, B., S. Kaufmann, and D. Diermeier (2008). Classifying Party Affiliation from Political Speech. *Journal of Information Technology & Politics 5*(1), 33–48.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge: Addison-Wesley Press.

Chapter 2

Interest Group Access and Campaign Spending Limits: Evidence from Brexit

ABSTRACT

Scholars have long been focused on studying lobbying and the potential influence that such activities can have on public policy. The ability to lobby state actors, however, critically depends on having access to them in the first place. So far much of the theoretical and empirical literature on potential mechanisms of acquiring access has been limited to donations or other forms of financial transactions. In this study I argue that in pluralist states with campaign spending limits, the influence of money is more restricted and other mechanisms such as economic importance, long period of interest-government interactions and ideological proximity play an important role in securing meetings with government officials. I use government transparency reports for 2010-2018 from the ministerial departments in the UK to measure the level of access and saliency of policy issues that provide evidence of the importance of these alternative mechanisms.

"'Politics' for us means striving to share power or striving to influence the distribution of power, either among states or among groups within a state."

—Max Weber, *Politics as a Vocation*

2.1 Introduction

There is a widespread concern both in academic (Gilens and Page, 2014; Acemoglu et al., 2015) and popular (Cave and Rowell, 2015) writing that economic elites and business interests receive preferential treatment from the government. If this is the case, adopted policies become biased in the direction of preferences of these actors rather than reflect the position of a median voter. While direct systematic evidence of influence is scarce, much of the literature has been dedicated to access, as a necessary (but not necessarily sufficient) condition for influencing public policy (Wright, 1990; Hansen, 1991; Ainsworth, 1993; Austen-Smith, 1995; Schnakenberg, 2017; Judd, 2019). In his seminal work *The Governmental Process* Truman (1951) put access as the basic interest group objective: "Whichever is operating at a particular point in time, however, power of any kind cannot be reached by a political interest group, or its leaders, without access to one or more key points of decision in the government."(Truman, 1951, p.264) Although in part this focus has been driven by practical considerations, as empirically it is easier to observe some form of access to government officials than the effects they might have on policy, the question, "Which of a plethora of interest groups seeking access get it?" remains the source of a long-standing debate in political science. This pluralist view of liberal democracy goes back to the discussion of 'factions' in *The Federalist Papers* (Hamilton et al., 1787). In the twentieth century this view was further developed by Dahl (1961), who argued that despite the competition, many diverse interests, including those of the general public, get represented. A major theoretical criticism of this view came from Olson's (1965) work on

collective action, who pointed out the 'free rider' problem that riddles widely dispersed interests, as opposed to numerically small, but well-coordinated, groups. Such lack of coordination can lead to what Schattschneider (1960) called "heavenly chorus [...] with a strong upper-class accent" or, later, labelled as 'unheavenly chorus' by Schlozman et al. (2012). In the context of British politics this polemic is paralleled by the introduction of the insider/outsider typology (Grant, 1978) used to characterize the groups that enjoy access to policymakers and those operating away from official cabinets. Subsequent critics (Maloney et al., 1994), argued, however, that most interest groups in fact do not become engaged in consultation process due to highly technical nature of many policies considered by the government.

Most studies to date have focused on the US, which is a specific example of pluralist system where campaign contributions play an oversized role. As I argue below, this situation arises when there is an absence of campaign spending limits. When those are present, however, political candidates have little incentive to raise large sums of money in donations and most interest groups cannot be differentiated on the basis of campaign contributions. This suggests that other access attainment mechanisms should be at work in these systems of government-interest relations. In particular, I focus on economic importance as a key explanatory variable which accounts for preferential treatment of some organizations. I illustrate my argument with the case of Britain, a classical example of pluralist system which also enforces strict campaign spending limits since the introduction of the Corrupt and Illegal Practices Prevention Act in 1883 (Rix, 2008). I further use the referendum on the membership in the EU as an exogenous policy shock that, by imposing additional temporal constraints, has made already limited face-to-face time with policy-makers an even scarcer resource. Despite the anticipated pro-business bias, I find that in the aftermath of the popular vote to leave, upheld by the government, economic importance played a less prominent role in securing meetings with the government.

This study makes five distinct contributions. First, it addresses an important concern about preferential treatment of business over public interests in political systems

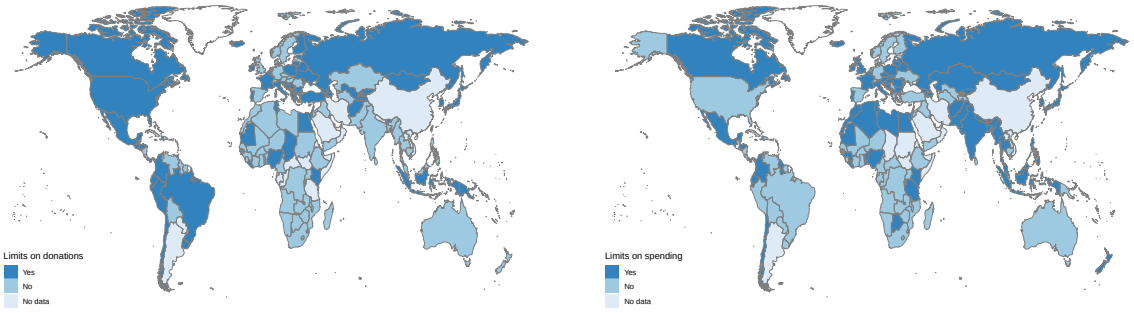
with restricted influence of direct financial investment in lobbying and financing political candidates. Second, I hypothesize and test economic importance and long-running ties between interest groups and state as an access mechanism that might determine achieving government access in pluralist systems with campaign spending limits. Third, I use an exogenous policy shock in the form of Brexit to estimate the joint effects of time constraints and public pressure arising from a highly charged political issue on access attainment. Fourth, given the importance and implications of Brexit that reach beyond national British politics, this work contributes to our understanding of the processes that accompanied the crucial years of negotiating withdrawal agreement. And, fifth, by collecting a novel dataset from transparency reports, I advance the discussion of open government data beyond Freedom of Information (FOI) regulations.

2.2 Lobbying and Campaign Spending Limits

Most empirical and theoretical research on pluralist democracies emphasizes the importance of campaign contributions as a key mechanism of securing access to policymakers (Austen-Smith, 1995; Grossman and Helpman, 2001; Fouirnaies and Hall, 2014; Kalla and Broockman, 2016; Powell and Grimmer, 2016; Fouirnaies and Hall, 2018). The usual argument is structured around strategic considerations on part of the organized interest, that channel money to legislators who enjoy incumbency advantage (Fouirnaies and Hall, 2014), are committee members (Powell and Grimmer, 2016) or possess other procedural powers, such as making committee assignments (Fouirnaies and Hall, 2018), in exchange for meetings and, potentially, influence. Donations, thus, act as a mechanism of either facilitating access through some quid pro quo arrangement (Snyder and Ting, 2008; Kalla and Broockman, 2016) or signalling close policy preferences between interest groups and political candidates (Austen-Smith, 1995; Hall and Deardorff, 2006). However, even scholars of the US politics note that there is less money than one would anticipate to observe given the expected payoffs (Ansolabehere et al., 2003), with very few publicly

listed companies donating money to political campaigns (Fouirnaies and Hall, 2018).

Figure 2.1: **Global Restrictions on Political Finances.** (a) Limits on donations to political candidates and (b) limits on candidate campaign expenditure.



(a) (b)

Note: Further details on data and methodology of evaluation are available in Ohman (2012).

This link between donations and political favours weakens once one considers pluralist systems with low campaign spending limits or corporatist systems with public funding of political parties (Siaroff, 1999), the types of systems dominant across Europe. This gives rise to an important theoretical puzzle. Either such systems create a level playing field where every organization has equal chances of being heard by the government or there are other factors at play driving the biased representation of commercial over public interests. In what follows I focus on economic importance, one of the factors that can shape interest-government relations in systems where donations play only modest role due to imposed campaign spending limits.

It is important to consider the two main regulatory restrictions that can affect this relationship between organized interest and political actors. First, it is possible for government to impose limits on the amount of money donors can give to political candidates and parties. Figure 2.1a shows countries across the world that have some limits on donations to candidates. Second, it is possible to restrict the sums that political candidates can spend on their campaigns (figure 2.1b). While there is some literature assessing the impact of the former restriction (Barber, 2015), no study has yet looked at how the mech-

anisms determining access differ under the latter[1]. In other words, do policymakers treat business interests as a more welcoming party to the negotiation table than civil society groups or non-profit organizations? And, if so, what drives this preferential treatment when almost no organization coaxes elected politicians through donating money to their political campaigns?

There are several implicit assumptions in these questions that need to be spelled out and tested. First, campaign spending limits have a more profound impact on the shape of interest-government relations than restrictions on donations. While the former completely changes the incentive structure by drastically limiting the demand for campaign resources, the latter merely necessitates some adjustments on how money-raising is organized. Indeed, as the Figure 2.1 shows, the most studied case of the US illustrates this point very well. Rather than levelling the playing field, diverse restrictions set at both state and federal level, arguably, shape how corporations channel their money through political action committees (PAC) and individual donations. Conversely, imposing limits on campaign expenditure would lead candidates in more competitive races to spend very close to the maximum amount permitted. At the same time candidates in safe seats have little incentive to raise money in donations given their confidence in electoral outcome. Second, in line with the previous literature, politicians' time is a limited and valuable resource that requires competition between interest groups. As was argued in previous studies on Britain only some groups acquire 'insider' status, while many remain excluded from the policy-making process (Grant, 1978, 2000)[2]. Or, to put it in other words, the *demand-side* for access is present in all polities. And, third, politicians are not omniscient and require information about the implications of potential policy changes, thus, *supplying access* to some interest groups but not others.

[1]There is emerging literature on the effects of campaign spending limits on electoral competition (Avis et al., 2017; Fouirnaies, 2018), however, this study is the first one that looks at how campaign spending limits shape interest-government relations.

[2]See also Maloney et al. (1994) for the critique of this dichotomy.

2.3 The British Case

When looking at the global variation in different forms of regulation affecting the relations between political candidates, their electoral campaigns and organized interest, the potential combinations are almost innumerable. However, in expanding our understanding of the interest-government relations it is helpful to explore a case which both builds upon the studies that looked at the US as a pluralist system par excellence and at the same time differs from it in the important respect of having strict campaign spending limits. While some authors looked at the patterns of access in countries that can more accurately be described as corporatist (Binderkrantz et al., 2015), this is the first study to provide an in-depth analysis of access acquisition in pluralist system where campaign contributions do not play a prominent role.

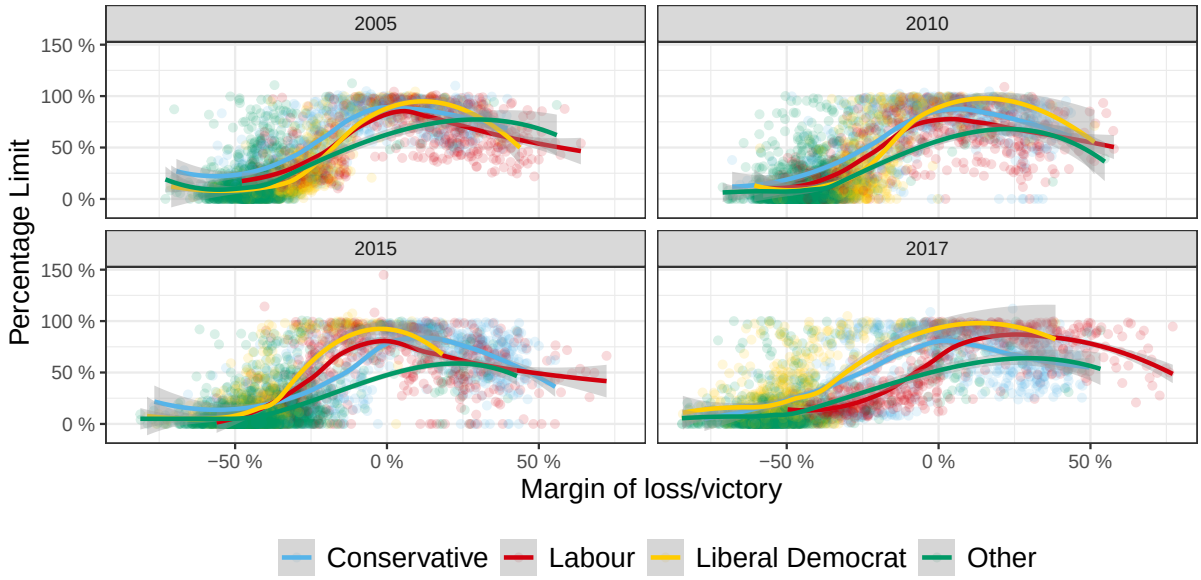
Britain, by being a classic example of the pluralist system in Europe and also a country with a long history of regulating political campaigns (Rix, 2008) provides an ideal case for such study. It is worth noting that despite the particularities of Westminster-type democratic systems, in those two respects– a diverse range of interest groups seeking access and restrictions on the amount of money candidates can spend on the electoral campaign, the UK is not entirely different from a range of other states, such as Canada, Ireland or Italy, to name but a few.

Interest (or pressure) groups featured prominently in the earlier studies of British politics (Beer, 1956; Eckstein, 1960), but their presence in academic literature declined considerably over the years. This is unfortunate as there is no indication that their political clout has diminished correspondingly[3]. Even more surprisingly, this has been happening against a backdrop of increasing scholarly attention to interest groups in American politics (Kalla and Broockman, 2016; Schnakenberg, 2017; Fouirnaies and Hall, 2018; Li, 2018; Judd, 2019) in the wake of the *Citizens United v. FEC* court ruling. Despite

[3]The consistent emergence of 'cash-for-X' controversies with 'cash-for-questions affair' in the 1990s and 'cash-for-influence scandal' in the 2000s is a testimony to the continuous concern about outside influence on politicians. David Hencke, "Tory MPs were paid to plant questions says Harrods chief," *The Guardian*, October 20, 1994, and Stephen Byers, "Ex-ministers suspended from Labour party over lobbying allegations," *The Guardian*, March 23, 2010

certain similarities in the interest-government relations, notably, the pluralistic nature of access seeking, not all insights from literature on the US are directly applicable to the case of Britain. Crucially for my argument here, faced with low spending limits individual political candidates have little incentive to raise substantial amounts to finance their electoral campaigns. This is especially the case for candidates, who do not run in marginal constituencies. These low investment in spending and raising campaign finances on the part of the candidates manifests in low number of donations[4] and decreased return on investment (access in return for campaign funding) on the part of organized interest.

Figure 2.2: **UK Candidate Spending 2005-2017.**



Note: Percentage of spending limits for short campaign reached by candidates in 2005-2017 Parliamentary General Elections and their margin of loss/victory. Candidates above the 100% level are those who spent above their legally permissible threshold. Data obtained from the spending returns reported to the Electoral Commission.

To illustrate the spending behaviour of candidates in the parliamentary General Elections, I analyzed spending reports for the past four elections that took place in 2005, 2010, 2015 and 2017. Figure 2.2 shows the distribution of levels of campaign spending limits reached by candidates in General Elections in the UK and their election results. The curves for all parties peak around 0, which indicates that candidates running in

[4]For more details on donations see Appendix A.

close races tend to reach the highest permissible levels of campaign spending. On the contrary, nonviable candidates concentrated in lower left corners of the panels, as well as candidates running in the safe seats (on the right) that they win with high margins spend considerably less and often do not go above 50% of the threshold[5]. In other words, winning candidates running in very safe areas can be expected to spend well below their permissible threshold. Taking into account a substantial number of safe seats for major parties in the UK, spending considerations and fund-raising appear to play a role only for a small number of candidates and MPs. Which, in turn, attracts fewer donors and interest groups, who anticipate limited returns on political donations.

2.4 Access Mechanisms

What are the potential sources of differential treatment of interest groups when candidates do not expand efforts to raise campaign funds and very few organizations donate to politicians? As politicians seek re-election, they are likely to be more attentive to the interests that help them secure one. In this sense, raising additional campaign funds can be viewed as one of the strategies to increase the chances of victory (Jacobson, 1978; Jacobson and Carson, 2015). However, as Gelman and King (1993) have argued, election campaigns are important insofar that they permit voters to learn about fundamental issues, such as the state of the economy. Thus, in order to increase the chances of re-election incumbent politicians operating in polities with strict campaign spending limits can instead focus on improving their standing on those fundamental issues. In performance-oriented retrospective voting model (Barro, 1973; Ferejohn, 1986) voters decide on the candidate who maximizes their well-being subject to constraint that politicians pursue their own interest. Furthermore, operating in the world of imperfect information, voters do not observe the actions of the politicians' directly. To put it in the context of interest group access, voters do not know how frequently the government meets with conservation organizations

[5]In practice these are relatively small amounts of money. With mean spending limit threshold being roughly £12,000, the winners, on average, spend about £9,000.

or trade unions, they only learn about the changes in their well-being linked to environmental or labour regulation and cast their ballot accordingly. However, to understand the complexities of individual policy issues, politicians require information about them in the first place. As interest group scholars have long argued (Milbrath, 1964; Baumgartner and Leech, 1998) information provision is one of the key activities that interest group engage in, often targeting both politicians as well as mass public. It is important to note that this information cannot be assumed to be entirely objective and unbiased. Indeed, as scholars have shown a considerable part of information provision involves agenda-setting, issue definition, and framing (Baumgartner and Leech, 1998). What is important, however, is that politicians either consider some interest groups to be the best sources of information or, simply, lack the capacity to get a more unbiased sources, thus being restricted to the pool of interest groups bidding for access.

It is hard to argue that of all fundamental issues that politicians might choose to deliver on to increase their chances of re-election, the economy is not the most important one. As has been aptly written by the Bill Clinton's political strategist for 1992 presidential campaign: '[It's] the economy, stupid'. Overall, the link between economic performance and voting has been studied extensively in the literature (E.g. Soroka et al. (2015); for reviews, see Lewis-Beck and Stegmaier (2000, 2007)) and found to be highly predictive of the electoral outcome. With this consideration in mind, politicians can be expected to be more likely to meet interest groups whose potential effects on the economy are larger. In other words, the economic importance of some groups can make their position and policy input essential to decision-makers. Tax contributions to the state economy, employment of voters, infrastructure projects are all important concerns for politicians seeking re-election that oftentimes cannot be addressed without consulting business organizations. Indeed, anecdotal evidence suggests that changes in tax regulation in the UK are often preceded by meetings with representatives of major corporations (Cave and Rowell, 2015). Thus, re-election concerns and the focus on economic performance could lead politicians to prefer meeting with business associations over civil society groups and

with larger corporations over local firms.

Second, in addition to economic importance politicians might prefer meeting organizations with whom they or their party had developed a strong connection over time. This mechanism is especially relevant to the British case, which is defined by long-established connections between some groups and political parties. Notably, until the present day sitting MPs are permitted to take jobs outside of parliament. Despite the requirement to disclose any financial interests or income since 1975, the practice of keeping second jobs continues to be a source of heated debate[6]. In comparison to other countries, a higher proportion of firms in the UK have been shown to have some political connection. Faccio (2006) estimates that 39% of firms (by market capitalization) have at least one of their large shareholders or top officers who is an MP, minister or is closely related to a top official. Historically, these close connections developed largely along the party lines. The history of the Labour party until very recently was intertwined with trade unions, while business interests and large business associations had close connections with the Conservative party. Eggers and Hainmueller (2009) find that Conservative candidates winning the elections leave behind larger estates, while Labour MPs do not seem to gain financially from their political careers. This finding can also be attributed to the differences in the nature of historical links between politicians and interest groups. Despite the decline of some of those affiliations in recent years, certain organizations might still be perceived as more reliable sources of policy-related information due to close historical or personal links.

While in the empirical section below I will focus on these two mechanisms, namely, *economic importance* and *long-running ties* with the government as the key explanatory variables, there are several other potential mechanisms that I outline below without explicitly testing them. As has been argued in some theoretical work before (Austen-Smith, 1995; Hall and Deardorff, 2006), preferential treatment arising from campaign contribu-

[6]One recent example includes George Osborne, a former Chancellor of the Exchequer, taking the editorship of *Evening Standard*, while retaining his position as an MP. "George of all trades," *The Economist*, March 23, 2017.

tions could be a result of closely-aligned policy positions, with donations merely signalling of this proximity. In many circumstances, however, it is plausible that policy positions of interest groups are known to politicians in advance. This would make signalling redundant, while still giving more saying to groups with preferences similar to those of the government. For example, one can imagine a conservation group lobbying government which includes a Green party as part of the ruling coalition. Fourth, as some issues become more salient, interest groups connected to these areas, e.g. operating in a particular sector of the economy, even if they do not, in general, are dominant for the economy at large, can enjoy periods of better access to government. Fifth, interest groups can promise lucrative career opportunities to politicians after they retire from the office (Eggers and Hainmueller, 2009; Blanes I Vidal et al., 2012; Palmer and Schneer, 2019). This mechanism is similar in spirit to quid pro quo financial transactions that occur during tenure in the office, apart from its delayed effect. While this 'revolving door' path is certainly a possibility, I argue that getting a chance to promise profitable post-retirement positions and ensuring that politicians perceive this as a genuine and trustworthy transaction, given the inherent delay and almost no legally-binding instruments, hinges upon having pre-existing established relationship that should be manifest through having consistent access in the first place.

2.5 Data and Research Design

In this study I use transparency reports released by ministerial departments of the British government. I collected and collated 1,193 individual files published quarterly on the official government website[7]. Despite a high variability in the specific release format, the standard fields that are contained in the vast majority of transparency files provide sub-

[7]To facilitate data systematic data collection from `https://www.gov.uk/government/publications`, I wrote a special R package. Furthermore, I identified any inconsistencies or missing periods in the automatically downloaded data, by checking the yearly distribution of meetings within each department. This, as well as the details of compiling a consolidated dataset from individual files are available in Appendix C.

stantial information about the contacts between government and organized interest. The observed variables include *official*, with whom the meeting took place, *date* when it happened (usually, up to month and year, as well as day for the latest periods), *organization*, one or several entities that participated in the meeting and a short description of a *purpose* of that meeting[8].

The collected data covers the period from 2010 to 2018 and includes information on all 24 current government departments. Overall, in the specified period government officials[9] had over 60,000 meetings[10] with over 30,000 organizations. To the best of my knowledge, this is the first study which uses the dataset on all reported government meetings in Britain[11]. Before proceeding to the research design, it is worth describing the context around the adoption of regulation that made it obligatory for the government departments to issue transparency reports.

The pledge to implement measures of greater government transparency was part of the Conservative party's electoral campaign for the 2010 General Election. It came as a response to 2010 'cash for influence' scandal, when a number of influential (and largely) Labour MPs, including 5 cabinet ministers, were recorded offering their services as lobbyists to an undercover reporter[12]. The scope of the reforms that happened when the coalition of Conservatives and Liberal Democrats took the office was greater than is being exploited in the present study. The establishment of the Open Data Institute, headed by the inventor of the World Wide Web, Tim Berners-Lee, or the creation of

[8]See the detailed instructions used by civil servants to fill in the reports in Appendix B.

[9]The data covers ministers, junior ministers, as well as, in some cases, permanent secretaries and special advisers. The meetings of the latter were not required to be reported when they were accompanying their respective ministers.

[10]In addition to meetings, the departments are required to report the details of ministers' overseas travel, gifts and hospitality. However, this data is far smaller than information about meetings and, arguably, less informative. With the vast majority of meetings taking place in London, trips are relatively rare and, often coincide with diplomatic missions abroad. Gifts and hospitality might contain information about extra efforts that interest groups make to persuade policymakers' of their position. But given that both parties know ex ante that this information will become public, they are very likely to avoid any transactions that might compromise them, leaving only ceremonial gifts that are retained by the department.

[11]See, however, Dommett et al. (2017) for a descriptive analysis of the data covering coalition years from 2010 to 2015

[12]"The great stink," *The Economist*, March 25, 2010

a dedicated open data repository[13] received considerably more public attention. While establishing the precise rationale for making this data publicly available is beyond the scope of the present study[14], it is important to consider the incentives of policymakers. The requirement to disclose meetings with external organizations forms part of the Ministerial Code[15], which sets out guidelines and standards for government ministers (Blick, 2014). While no sanctions are specified for breaking the 'rules', given public nature of the document and scrutiny by the opposition and media, re-election concerns are likely to be the primary force driving compliance with the code. This lack of explicit enforcement of a new regulation, however, led to uneven and protracted adoption by the departments. The quality of the data released by the government departments gradually improved and reports started to be published more systematically (once per quarter) and in proper machine-readable format (CSV or well-formed Excel file[16]). Overall, it can be said that by 2013-2014 the reporting standards were sufficiently standardized and largely followed[17].

In addition to transparency reports, I use data provided by Bureau van Dijk on more than 10,000,000 organizations registered in the UK[18]. This dataset includes both private and public companies, as well as non-profit organizations. The linking of transparency data to information on interest groups allows to create a uniquely rich dataset for studying the level of government access and different mechanisms that can increase or decrease the chances of attaining it. All linked organizations were further manually labelled by the author whether they are regular private companies or charities or represent a multitude

[13]In fact the website in question, https://data.gov.uk/ contains only a fraction of the data files used in this investigation, while the main https://www.gov.uk/government/publications, from which the data was obtained, stores the full universe of department transparency reports.

[14]It could be a result of the electoral uncertainty - while winning a plurality of seats in the Parliament, Conservatives still required the support of Liberal Democrats to form a government. Thus, it would be in the interests of the incumbent to pass relevant legislation, such that in case of a future loss, transparency data could be exploited for criticizing the new incumbent and fighting the following electoral campaign. It is also plausible that they, simply, perceived the preference for greater transparency to be shared by large parts of the electorate. See Berliner (2014) for more detailed discussion of the adoption of transparency regulations.

[15]Further details of the provisions specified in the Ministerial Code are available in Appendix B.

[16]For earlier years is is not uncommon to find reports in PDF or even DOC format.

[17]See Appendix D for descriptive statistics across years and departments.

[18]In the present analysis I restrict the entire available dataset of 300,000,000 organizations across the world to those legally registered in Britain for both conceptual and computation reasons.

of interests by functioning as umbrella organizations. An examples of those would include trade associations, chambers of commerce and multi-academy trusts.

In this study I exploit the unexpected results of the EU membership referendum (Brexit in common parlance) as a sudden exogenous policy shock, that could have a dramatic impact on the types of interests that get access to government ministers during the crucial two-years phase of negotiating withdrawal agreement. While a definitive identification of the effects of Brexit on access attainment is impossible due to composite nature of shock and the lack of well-defined control group, I approximate it by comparing the association between different mechanisms and the level of access before and after the EU referendum.

In what follows I provide a brief overview of the EU referendum in the UK and the theoretical expectations about the effects it might had on the level of access acquired by more or less economically powerful interest groups. Next, I show the changes that the referendum made in the policy agenda of government-interest discussions by looking at the changes in keywords usage and using a variant of dynamic topic model to model the overall evolution of government-interest agenda over time. After that I outline the modelling strategy for measuring economic importance of different groups over time. And, finally, I present the results of the main analysis of access mechanisms, comparing the relationship between different mechanisms and the level of access achieved before and after the referendum.

2.6 EU Referendum and Time Constraints

In June 2016 voters in the United Kingdom took part in the referendum on membership in the EU. With 51.9 % of them voting for the Leave option , the government started the process of withdrawing from the European common market. While there is an ongoing debate about the reasons for this outcome (Hobolt, 2016; Clarke et al., 2017; Colantone and Stanig, 2018; Carreras et al., 2019), it is important to note that the referendum

result, although a clear possibility according to some polls[19], took most members of the government, public and organized interest by complete surprise. Very few, if any, at that time took any measures to alleviate the potential economic damage caused by the sudden and dramatic change in political governance, tariff policy and labour market. Furthermore, despite the referendum being legally non-binding, the government decided and publicly announced that it will not renege on the 'will of the people' and deliver Brexit, whatever the costs might be for economic actors. In March 2017, it invoked article 50 of the Treaty of the Union and started a two-year countdown process to the actual withdrawal[20]. All of these points to a situation, in which business interests were faced with high uncertainty about the future that was further exacerbated by time constraints imposed by the triggering of the article 50. Under these circumstances, we would expect an increase in the demand for access, while the supply side, namely, policymakers' time, is even more limited than usually.

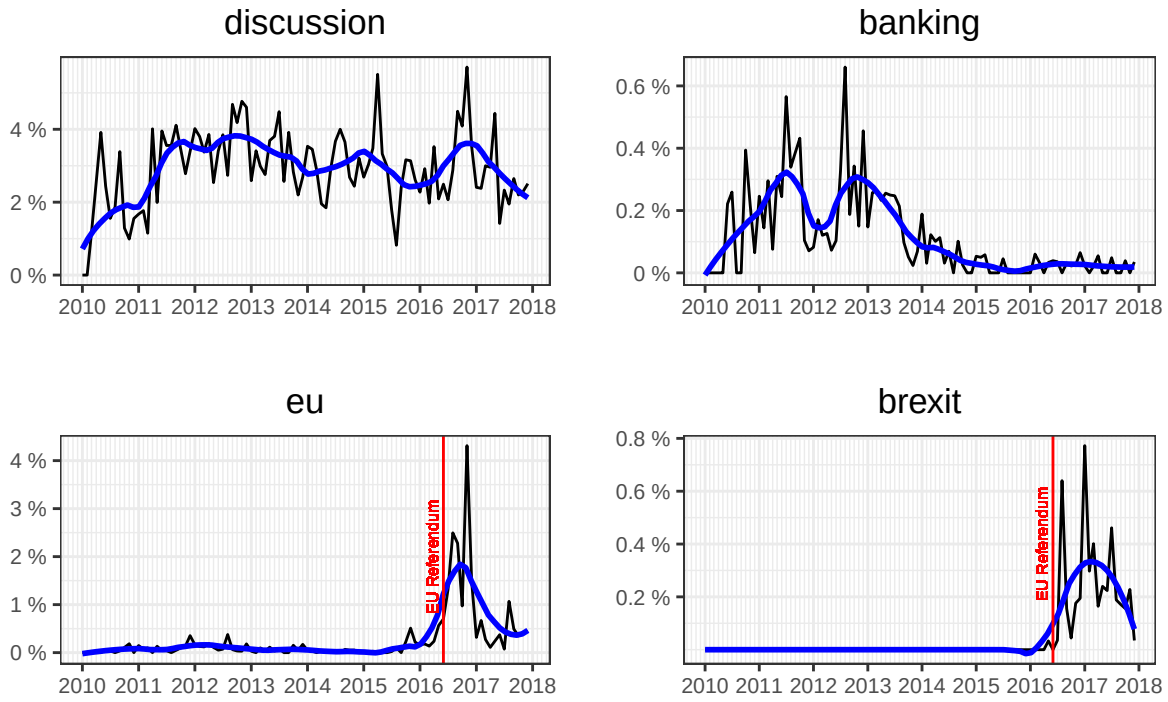
The collected transparency reports permit a rough test of this supposition by including a stated goal of the meeting between interest groups and government officials. Indeed, as the lower panels of figure 2.3 show, there was a sharp rise in the proportion of words 'eu' and 'brexit' used to describe the purpose of meeting with the government officials. It is worth noting that these two words start from different baselines. While the term 'brexit' is, essentially, absent prior to some time earlier in 2016, the discussions regarding 'eu' were always part of government's agenda. This, however, jumps to over 4 % of all words some time in late 2016. The two additional terms, 'discussion' and 'banking', are plotted as a reference. As the former is used throughout the period under analysis to describe a general or introductory meeting its usage is nearly constant across time. On the other hand, the discussion of banking regulations featured in the aftermath of the financial crisis of 2008, but tails off over time.

When time is in short-supply, one would expect political actors to be more selective

[19]E.g. https://yougov.co.uk/topics/politics/articles-reports/2016/06/28/online-polls-were-right

[20]The date for triggering article 50 surfaced in October 2016, which could be another important signal to organized interests that time is in short supply. More detailed timeline of the UK's withdrawal from the EU can be found in Evans and Menon (2017)

Figure 2.3: **Purpose of Meetings.** The monthly share of the words *discussion*, *banking*, *brexit* and *eu* in the reported purpose of meeting with a loess smoothed line.



about whom do they grant access to. Whatever the mechanisms driving preferential treatment of some interests over other, they are likely to become more pronounced during the period of extreme uncertainty. In other words, a reasonable expectation would be that only a selected few interest groups get audition. However, it is important to note another side of the Brexit crisis, which is the popular nature of the vote that set off this chain of events. This and great level of public scrutiny (considerably higher than for any other policy issue , that accompanied the process of negotiating new terms with the EU make the usual expectation that large business interests will prevail far from given. In what follows, I use this unique case of increased demand for access under the conditions of reduced supply, accompanied by high level of pressure exerted by the voters, to test empirically the mechanisms and biases in acquiring access in pluralist systems.

2.7 Measurement of Issue Salience

The imposition of strict time constraints on policy-making process when the status quo is not a viable option creates a lot of uncertainty for all interests affected by the change. First, it increases the number of affected parties. Under standard conditions when a new bill is introduced into parliament, groups that are content with the status quo only need to engage in the consultation process to the extent to which they believe a draft bill has chances of replacing the current status quo. However, when the status quo is not among potential options, this should mobilize all groups no matter where their preferences lie. Second, the level of uncertainty depends on how narrow the potential changes could be. Broader changes to multiple policy areas imply higher current and future costs of planning and implementation. Facing such costs coupled with ambiguous payoffs, groups are likely to respond by investing resources into discerning any credible signal about the future from the policymakers. Both of those aspects are characteristic of the bargaining process between Britain and the EU in the aftermath of the referendum. The status quo, remaining a member of the single market, was ruled out by the government from the very beginning of negotiations. Furthermore, it is hard to conceive a more sweeping set of reforms that would simultaneously affect tariffs and labour market than exiting the EU.

These two points should provide theoretical underpinning for the expected increase in the demand for access on the side of interest groups. As even in the best case scenarios we can only observe groups that acquire access rather than the entire universe of those that seek it, it is typically impossible to measure demand empirically. However, for changes of such magnitude it might be feasible to derive observable implications from the data. One such indirect test could be done by looking at the shifts in the focus of discussion by the interest groups that acquire access. Re-balancing the focus from a more niche policy area (e.g. environment) to the one that substantively affects a larger population of organized interest could indicate an increase in competition among different organizations. A tentative indication of that can be seen in the Figure 2.3 above. However, the question

remains about how one policy area fares against others. For that we would need to model jointly the salience of various policy issues indicated in the reported purposes of meetings as well as their co-evolution over time.

Until recently the most widely used approach to modelling the salience of policy issues over time is manual coding of multiple data sources, such as adopted by the *Comparative Agendas Project* (Baumgartner and Jones, 1993; Baumgartner et al., 2019), the largest and most well-known cross-country collaboration of its kind. While human coding has its merits, such as fine-grained categorization and nuanced understanding of context by the coders, similar cross-time cross-country projects, such as *Comparative Manifesto Project* (Budge et al., 1987; Volkens et al., 2013), designed to capture shifting party policy positions over time has been shown to suffer from major methodological drawbacks, such as lack of uncertainty estimates (Benoit et al., 2009) and unreliability of many theoretically informed categories (Mikhaylov et al., 2012). In light of this critique, here I adopt an alternative approach to modelling policy agenda over time, the one based on unsupervised learning from text (Grimmer and Stewart, 2013). More specifically, I use a variant of dynamic topic modelling (Blei and Lafferty, 2006b) to capture cross-time variation in declared policy interests. While here this method is applied to the case of Britain, it is easily extendable to other national contexts[21].

Topic models in their original form proposed by Blei et al. (2003) are Bayesian hierarchical mixed-membership models, where each document in a corpus is given prior Dirichlet distribution over a fixed number of topics and each word in a document is drawn from multinomial distribution conditioned on a randomly chosen topic. The appeal of this method is the possibility to model documents as a mixture of different topics as opposed to ascribing each of them to a single class (Quinn et al., 2010). Another advantage of unsupervised methods, to which topic models belong, is learning from the data without the need to pre-specify the policy areas and label individual documents as

[21]See Greene and Cross (2017) for the recent application of dynamic topic models for exploring the policy agenda of the European Parliament.

containing a subset of them[22]. One of the disadvantages of the original implementation of topic models, Latent Dirichlet Allocation (LDA), was both cross-sectional and temporal independence of topics. These shortcomings of the original model were later addressed in Blei and Lafferty (2006a) and Blei and Lafferty (2006b), respectively. As the primary concern of policy agenda literature and the main observable implication of changes in demand for access are shifts in issue salience over time, dynamic topic models are a natural choice of modelling strategy.

The key downside of the original Dirichlet prior over topic proportions is independence of individual components of the resultant vector from each other. This severely constrains our ability to model any endogenous dependency structure that might appear more appropriate for the problem at hand. To address this limitation, I follow Blei and Lafferty (2006b) and replace Dirichlet prior with sequentially chained normal distributions. To recap the original formulation, let D be a collection of documents $\{d_1, ..., d_M\}$, each of them being a mixture of K different topics. Every document contains W_{d_i} words, drawn from a common vocabulary V. Thus, each observation is a word $\{w_1, ..., w_N\}$ that is conditioned on the topic z and word probabilities $\beta_{ij} = P(w_j = w^v | z_i = z^k)$. Topic z here has the following prior structure:

$$z_N \sim \text{Multinomial}(\theta)$$

$$\theta \sim \text{Dir}(\alpha)$$

where α is a pre-specified concentration parameter of the Dirichlet distribution. Instead of using Dirichlet, in the dynamic topic model this prior structure is replaced by a simple state space model:

[22]Further details on the differences between supervised and unsupervised models are available in Hastie et al. (2009).

$$\theta \sim \text{Multinomial}(\alpha)$$

$$\alpha \sim \text{Multinomial}(\eta_t)$$

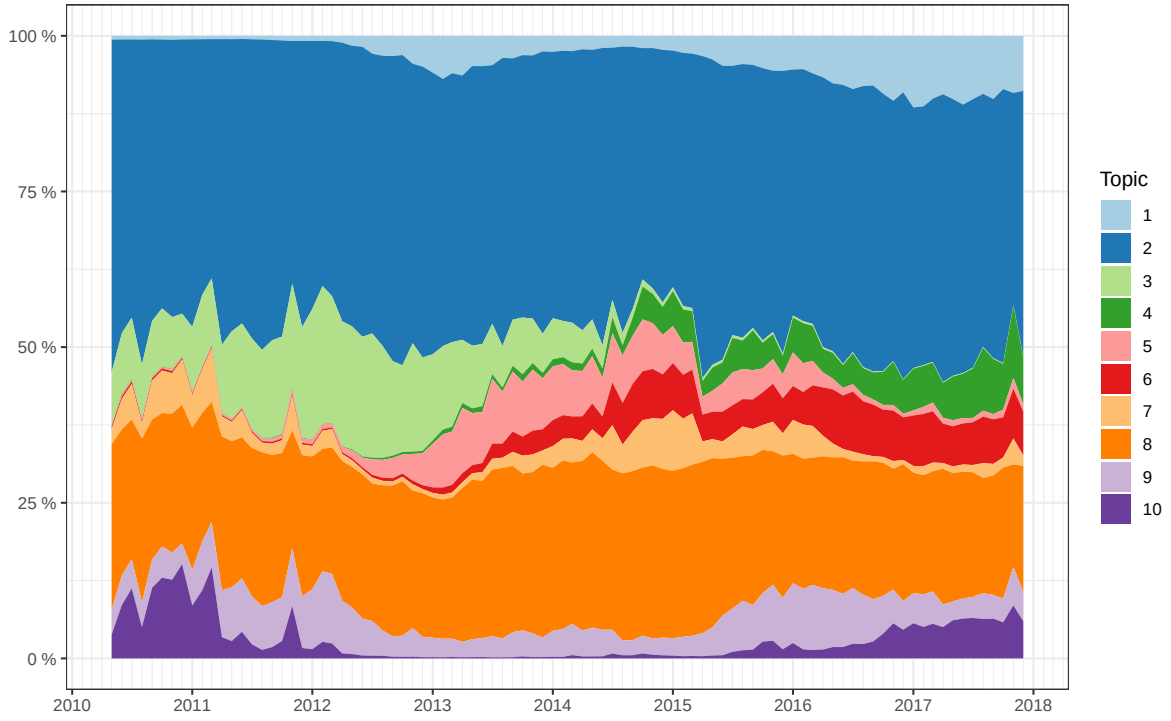
$$\eta_t \sim \text{N}(\eta_{t-1}, \sigma^2 I)$$

where η_t is a topic proportion parameter at period t is conditioned on the equivalent parameter at $t-1$[23]. In contrast to the model proposed by Blei and Lafferty (2006b), I assume word probabilities β_z to be fixed across time. In other words, while permitting topics to evolve over time, the composition of topics remains constant. This assumption is adopted for substantive reasons as well as with computational considerations in mind. Given the relatively short time span of the observed data (2010-2017), it is highly likely that the lexicon used to describe different policy areas remained constant. At the same time, chaining β parameters requires the computation of covariance matrix of $V \times V$ size, which becomes prohibitively large even for middle-size vocabularies. In addition to that manual inspection and qualitative assessment of the estimated topics becomes problematic when word probabilities change over time.

As the individual descriptions of meetings are very short (typically, not more than a single sentence) they have to be aggregated to make the estimation of topic model feasible. Monthly aggregation was chosen to balance the trade-off between granularity of observed changes and the amount of data needed for model fitting. In other words, a composite string containing purposes of all meetings in a given month, is taken here a single document for modelling purposes. After some experimentation, the model with 10 topics was selected to describe the evolution of policy issues in Britain between 2010 and 2018. Figure 2.4 shows how salience of different policy areas changed over this period. Table 2.1 further lists the 10 words with the highest probability of being generated by a given topic.

[23]The state of η at time period 1 is modelled by taking a random draw from a normal distribution: $\eta_1 \sim \text{N}(0, 5)$

Figure 2.4: **Evolution of policy issues over time.** Modelled by fitting dynamic topic model to the reported purposes of meetings between interest groups and government aggregated by month.



Much of the discussion with interest groups throughout this time is dominated by the general industrial policy (*'Topic 2'*), which is characterized by such words as 'general', 'business' and 'meeting'. It is noteworthy that the proportion of the agenda dedicated to the general business meetings visibly declines from around half to about 40% by the end of the observed period. Another noticeable change is the proportional rise of the topics that are related to Brexit (*'Topic 4'* and *'Topic 10'*). Each of them reaches 4% and 5% of the overall agenda by the end of 2017[24]. Other words within those topics ('teacher', 'steel', 'academies') indicate that, apart from a general upward trend, Brexit also encroaches upon a diverse set of policy issues.

Although circumstantial, the resurgence of Brexit-related topics in the agenda of government-interest discussions and its interplay with other policy areas points to the

[24]While both of those topics feature prominently at the beginning of the perioc, around 2010, this can be attributed to the new coalition government of Conservatives and Liberal Democrats and the manifesto promise of the former to re-negotiate the UK's relationships with Europe, which resulted in the referendum six years later.

Table 2.1: **Top 10 words in each policy area.** Estimates are derived by fitting dynamic topic model to the reported purposes of meetings between interest groups and government.

Education (1)	General (2)	Coalition Summits (3)	Brexit I (4)	Geopolitics (5)	Northern Powerhouse I (6)	Military (7)	Charities (8)	Northern Powerhouse II (9)	Brexit II (10)
exiting	policy	mesothelioma	exiting	antimicrobial	powerhouse	antimicrobial	campaigning	powerhouse	exiting
brexit	for	ï	dit	wwi	antimicrobial	resistance	esp	steel	brexit
powerhouse	introductory	table	brexit	mr	resistance	your	awe	charter	exit
steel	update	unemployment	termly	premium	official	ebola	deals	relevant	reading
charter	business	1/2	exit	transatlantic	steel	awe	wwi	northern	regime
teacher	meeting	g8	do	airport	charter	attendance	friends	benefit	contribution
northern	discussion	localism	victims	de	teacher	army	streets	localism	phonics
across	general	dairy	apprentices	pupil	northern	command	animal	children's	renegotiation
children's	issues	waste	supported	1/2	children's	get	general	devolution	initial
devolution	discuss	big	sponsorship	army	devolution	academies	discuss	big	academies

expansion of the population of interest groups that could be affected by the looming policy changes.

2.8 Measurement of Economic Importance

In the absence of direct financial transactions between interest groups and policymakers, economic importance could be a primary mechanism that explains advantageous position of large corporations. Measuring this importance that different companies have in the politicians' eyes is not straightforward. Some of it can be captured by usual economic indicators, such as assets and number of employees, but it seems rather improbable that elected officials always consult those numbers before inviting relevant stakeholders or granting access to those who made an inquiry. What the politicians are more likely to have in mind is some noisy aggregate picture of how important a particular organization might be. As most organizations exist for some period of time, prior economic history might also play a role in determining their current perception. In principle, we are interested in measuring a subjective component of economic importance. Or, in other words, how different objective economic indicators might be weighted by politicians when making a decision about access. However, this measurement problem currently presents a challenge that is hard to directly address empirically. Instead, I model economic importance as a latent variable with time-dependency structure. While this model does not explicitly include data that would allow to capture subjectivity directly, through aggregating multiple economic indicators and incorporating temporal change, it should sufficiently

approximate the types of information that politicians might take into account.

In order to measure economic importance of interest groups I build a dynamic factor model that uses time series of objective economic indicators, while also incorporating lagged latent component[25]. More formally, each interest group i at time point t is said to have economic importance λ_{it}, which is a function of its economic importance at $t-1$:

$$\lambda_{it} = \theta + \gamma\lambda_{i,t-1} + \upsilon_{it} \tag{2.1}$$

Here, θ is the mean of economic importance of all groups across all years and γ is the auto-regressive coefficient, which is assumed to be constant across time. Intuitively, this first-order auto-regressive part on latent factor means that companies with exactly the same performance figures in one year can have different importance based on their indicators for the previous year. In other words, subjectively politicians give weight not only to the current situation, but also to their recollections of past status of the company. It is possible that former industrial titans still carry a lot of importance in the politicians' eyes, while newly emerging technological start-ups are overlooked despite parity in the economic indicators. By assuming γ to be constant across time, this 'persistency effect' is taken to have universal effect for each year-to-year transition.

One important caveat of this dynamic model is that initial observations has to be modelled separately, as no data is available for preceding time points[26]. To approximate the first latent state I follow Heckman (1978) and Stegmueller (2013) and model λ_{i1} as:

$$\lambda_{i1} = \psi\xi_i + \upsilon_{i1} \tag{2.2}$$

Here initial observations are shaped by unobserved organization-specific characteristics

[25]The proposed model is similar in spirit to the approach adopted previously by Martin and Quinn (2002); Stegmueller (2013) and Fariss (2014).

[26]Although the data for previous years is not incorporated in the model, one cannot simply assume that organizations did not exist before certain universal moment in time. In fact for this model I am using data for 2009-2018 to model the parameters for 2010-2017. However, adding extra time periods simply shifts the problem to an earlier point and, potentially, attenuates the assumptions, but does not resolve it.

ξ_i with a scale parameter ψ.

All parameters in equations 2.1 and 2.2 are unobserved. To estimate the parameters in the structural part of the model, I use the observed economic measures: *assets*, *employees* and *revenue*, that are linked to the latent economic importance through the following equation:

$$y_{ijt} = \alpha + \beta_j \lambda_{it} + \epsilon_j \tag{2.3}$$

Where y is the logarithm of the indicator j for company i in year t[27]. Each indicator is assumed to have unique loading β_j on the latent factor λ_{it}[28]. To estimate the specified model I use MCMC No-U-Turn sampling that was implemented in Stan (Carpenter et al., 2017)[29]. Table 2.2 shows the regular and umbrella organizations that had most meetings with the government between 2010 and 2018 and their respective economic importance parameters averaged over this period. As these estimates were produced using registration data from UK, for some of the multinational groups (e.g. EDF Energy) economic importance might appear underestimated. However, this reflects the fact that its UK subsidiary is smaller, according to the indicators used, than for some other corporations (e.g. KPMG). Amongst the umbrella organizations with most meetings we can find traditional actors, such as Universities UK, representing the interests of higher education and the British Medical Association, which unites doctors, other medical professionals and organizations within the National Health Service. It is worth noting that all umbrella groups fall behind corporate entities in terms of their estimated economic importance. By and in themselves they rarely employ a comparable number of people or hold significant assets, as opposed to their constituent organizations, whose interests they assumed to represent.

Estimated economic importance parameters ($\hat{\lambda}_{it}$) can now be used as an explanatory variable for modelling level of access acquired by each of the organizations. Another

[27]Taking the logarithm of the indicators converts otherwise heavily right-tailed distributions of the number of employees, assets and revenue to a one closer to normal. This adjustment substantially improves the fit of the latent factor analysis model.

[28]For identification purposes the variances of the stochastic errors $\epsilon_j \sim N(0, \sigma_\epsilon^2)$ and $\upsilon_{it} \sim N(0, \sigma_\upsilon^2)$ are fixed: $\sigma_\epsilon^2 = \sigma_\upsilon^2 = 1$. Furthermore, errors are assumed to be independent: $Cov(\upsilon_{is}, \upsilon_{it}) = t \forall s \neq t$

[29]I run 10,000 iterations and discard the first 5,000 before estimating the parameters. Convergence diagnostics are available in Appendix F.

Table 2.2: **Economic Importance.**

(a) Individual Organizations

Organization	Incorporation	N Meetings	Importance
bae systems	31 December 1979	428	58.00
bt	2 February 1988	291	47.86
kpmg	22 February 2002	250	52.66
network rail	22 March 2002	232	57.55
bp	NA	183	62.01
centrica	NA	179	57.83
tesco	NA	169	60.85
vodafone	7 January 1980	138	55.25
lloyds banking group	11 January 1995	125	61.48
edf energy	1 April 1989	118	53.44

(b) Umbrella organizations

Organization	Incorporation	N Meetings	Importance
universities uk	29 June 1990	189	40.83
which?	13 December 1960	144	44.15
national housing federation	22 June 1935	113	41.96
business in community	2 March 1982	101	42.67
eef	28 September 2006	88	44.37
cancer research uk	20 November 2001	74	50.07
thecityuk	26 November 2009	68	38.13
british medical association	21 October 1874	58	46.71
country land and business association	28 February 2007	57	41.27
scotch whisky association	22 April 1960	57	38.63

Note: Top 10 organizations by the number of meetings and their estimated average economic importance over 2010-2018.

consequence of including the temporal lag in the model is smoothening of any abrupt changes in the indicators of economic performance. In the Appendix I further show the individual trajectories of economic importance over time to check for any changes in the overall dispersion of estimated parameters over time.

2.9 Results

To measure the level of access enjoyed by interest groups if varying economic importance and how it changes before and after the vote I take subsets of the full data with the month of the referendum in the middle and one or two years window on each side[30]. First, I compare the effects of hypothesized mechanisms on the level of access to government before and after the EU referendum. The level of access is measured here as the number of meetings each interest group gets during this period. Apart from the referendum itself in June 2016, another potential break point could be the triggering of article 50 in March 2017[31]. More formally, here I estimate the following quantities:

[30] In addition to that I test the break point with a placebo referendum date by shifting it two years and testing the same model on this subset of the data.

[31] There are, however, a couple of considerations that suggest otherwise. First, given the then potential level of policy uncertainty, it is reasonable to expect that interest groups' demand for access increased immediately after the referendum results were announced. Second, as Figure 2.3 indicates the first spike in the mentions of EU-related terms in the description of meetings' purposes happens around the time

$$Meetings_i = exp(\delta_1 I\{Post - Referendum\} + (\sum_{j=1}^{J} \delta_{2j} m_{ij} + \delta_{3j} m_{ij} I\{Post - Referendum\}) + (\sum_{k=1}^{K} \eta_k x_{ki}) + \epsilon) \tag{2.4}$$

Or the number of meetings the group i had before and after EU referendum took place in June 2016 with m_{ij} characterizing mechanism j and x_{ki} capturing other organization-level covariates. While δ_1 is not of direct interest here, it represents any potential increase in the overall number of meetings for the post-referendum period. Coefficients δ_{2j} capture a more interesting dynamic, as they show the relationship between different hypothesized mechanisms and the level of access, while δ_{3j} indicate how this relationship was shaped by time constraints resulting from the abrupt decision to leave the European common market. The exponentiation part is present here as the outcome of interest is the count of meetings, which is modelled with Poisson regression.

Table 2.2 shows the estimates of this model. To capture the potential historical links that interest groups might have with the government, I use the number of years that passed between group's incorporation and having a meeting to create what is labelled as 'age'. While it is likely that the specific nature of long-running ties between certain group and the government depends on the incumbent, with Conservatives favouring business associations and Labour tilting towards trade unions, the lack of variation in the party controlling the government during the analyzed period, prevents from testing it empirically[32].

In this specification age and importance are averaged over years before and after the referendum, in which each group had meetings. In other words, if a group had meetings with the government in 2014 and 2015 (before the referendum took place) and in September 2016 and some time in 2017 (after), its economic importance and age before the cutoff is estimated as the mean of economic importance and years of existence in

of the referendum.

[32]Although in 2015 there was a shift from a coalition of Conservatives and Liberal Democrats to Conservative government, this change was idiosyncratic for post-war era in British politics.

2014 and 2015. Equivalently, for the post-referendum period the average of importance and age is calculated for 2016 and 2017. To ensure comparability I estimate the model on the restricted subset, with only meetings happening in the year or two years before and after the referendum included.

The results of estimating this model on two subset of the data are presented in Table 2.3. There are no significant changes in the overall number of meetings for the shorter one-year window, however, looking at a longer time span of two years we find that in the post-Referendum period the same organization can be expected to have 36% ($\hat{\delta}_1 = 1.311$, model 6) more meetings, holding other organization characteristics constant. While that politicians' time is inherently limited and cannot be dramatically increased irrespective of circumstances, when deciding between activities it is plausible that gathering additional information on the potential effects of withdrawal was among their priorities, which led to an increase in the number of meeting they held with organized interest. The age variable appears to have minimal effect on the level of access across all specifications. At the same time economic importance is consistently positive and statistically significant. Furthermore, the interaction term between the post-referendum coefficient and economic importance is negative (and also statistically significant for longer time window), indicating that an organization of similar economic importance can be expected to have 1% ($\hat{\delta}_{3\_importance} = -0.0076$, model 6) fewer meetings in the aftermath of the referendum than before it. I also find that umbrella organization, representing multiple interests have substantially more meetings. On average they can be expected to have 75% more meeting before the referendum ($\hat{\delta}_{2\_umbrella} = 0.5545$) and 90% more after ($\hat{\delta}_{2\_umbrella} + \hat{\delta}_{3\_umbrella} = 0.64$).

In Figure 2.5 I also show the expected number of meetings before and after the referendum given organization's economic importance. The marginal estimates are based on model 6 from the Table 2.3[33].

[33]Further robustness tests using placebo Brexit date can be found in the Appendix.

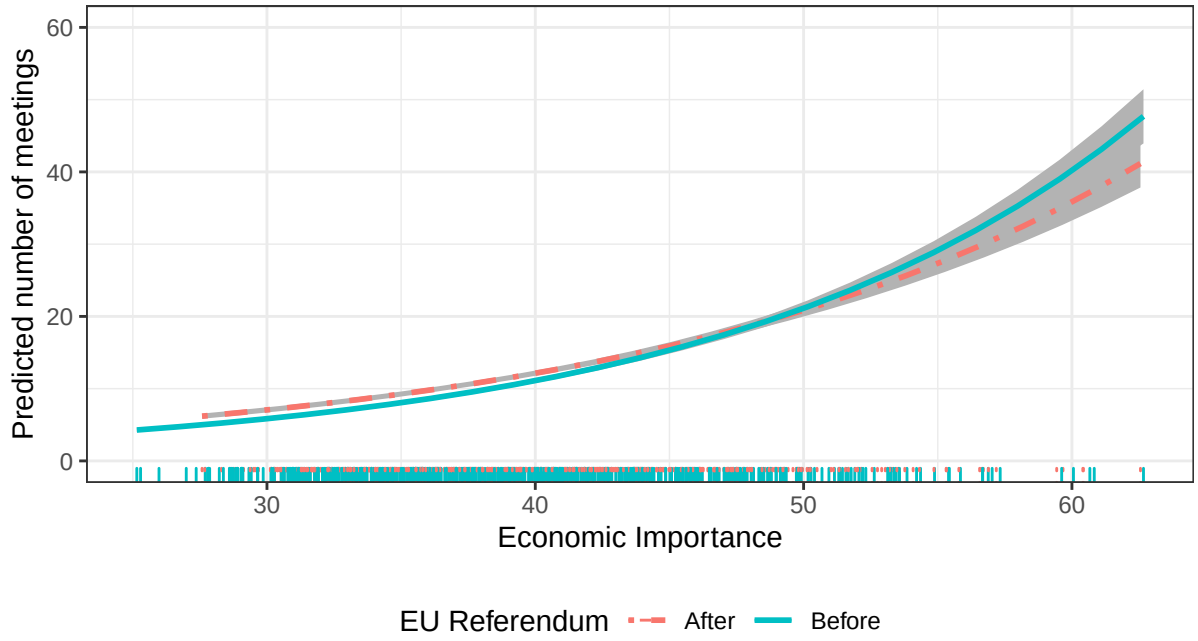
Table 2.3: **Models of the level of access.**

	Meetings (12 months pre/post)			Meetings (24 months pre/post)		
	(1)	(2)	(3)	(4)	(5)	(6)
Post-Referendum	0.0056	-0.0593	0.0307	0.0367*	0.3044*	0.3111*
	(0.023)	(0.1742)	(0.1742)	(0.0169)	(0.1276)	(0.1275)
Economic Importance	0.0585***	0.0589***	0.0578***	0.0758***	0.0799***	0.0772***
	(0.0019)	(0.0026)	(0.0027)	(0.0014)	(0.0019)	(0.002)
Age	7e-04*	-3e-04	6e-04	7e-04**	-3e-04	5e-04
	(3e-04)	(5e-04)	(5e-04)	(2e-04)	(4e-04)	(4e-04)
Umbrella	0.4976***	0.4651***	0.4199***	0.6355***	0.5836***	0.5545***
	(0.0284)	(0.0398)	(0.0411)	(0.0207)	(0.0302)	(0.0314)
Post-Referendum * Importance		-3e-04	-0.002		-0.0079**	-0.0076**
		(0.0038)	(0.0038)		(0.0028)	(0.0028)
Post-Referendum * Age		0.0019**	0.0019**		0.002***	0.0016**
		(7e-04)	(7e-04)		(5e-04)	(5e-04)
Post-Referendum * Umbrella		0.0662	0.0364		0.0959*	0.0874*
		(0.0568)	(0.0578)		(0.0415)	(0.0423)
Sector			✓			✓
Constant	-0.5661***	-0.5457***	-0.7934***	-1.0833***	-1.222***	-1.2763***
	(0.0869)	(0.1159)	(0.155)	(0.064)	(0.0885)	(0.1104)
Observations	894	894	870	1266	1266	1236

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Note: Access is measured by the number of meetings that took place between any government minister and a given interest group. Poisson link is used to model the dependent variable.

Figure 2.5: **Predicted number of meetings.** Economic importance secures fewer meetings in the aftermath of the EU referendum.



Note: Estimates are based on Poisson GLM model. Ticks along the x-axis show the observed values in the data.

The results demonstrate that in accordance with the theoretical expectations, economic is a positive and significant predictors of the number of meetings between interest groups and the government. A more surprising finding is that after the referendum this association becomes weaker. At the same time there is no indication that age plays a significant role in getting extra time with policy-makers or that Brexit affected the link between organizational history and an associated level of access.

2.10 Discussion

As the results above indicate the direct association between economic importance and long history is in line with theoretical expectations. However, the effect of Brexit and concomitant time constraints is not immediately clear. Why did the relevance of organizations' economic importance for securing access to the government decline in the

aftermath of the referendum? While a well-grounded answer to this question will require further investigation, it is possible to outline a few possible explanations. First, it is worth noting that this empirical result is largely in line with the perception of business circles as well as journalistic accounts of the negotiation process that emphasized that many business interests were ignored and need to be vigorously defended[34]. Second, it is worth highlighting another aspect of the Brexit process - high salience of the issue and, therefore, high visibility of much of the process. This distinguishes the process from any typical bargaining over policy changes that interest groups usually engage in. As many of these discussions can be highly technical in nature, usually they attract far less public scrutiny than a sweeping set of changes that the departure from the EU common market would necessitate. This finding also speaks to 'scope of conflict', proposed by Schattschneider (1960). In this view business interests are more likely to prevail if the scope is narrow and the visibility of the case is limited. As the exact opposite is true of the debate surrounding Britain's relationships with the EU, it is, perhaps, less surprising that less important groups, but who could be assumed to represent public preferences get more preferential treatment and higher level of access to the government. Third, a, perhaps, less normatively appealing explanation for this finding could be that the government in negotiating the withdrawal agreement with the EU had to rely on the input from a number of consultancies, which tend to be smaller on average than manufacturing and retail firms or non-profit organizations or umbrella associations.

2.11 Conclusions

In this paper I propose and test mechanisms that drive differential access to policy-makers in pluralist systems. In particular I focus on economic importance of organizations as an explanatory variable that substitutes signalling through campaign contributions in those pluralist systems which set limits on electoral spending by parties or individual can-

[34]"When the gloves come off," *The Economist*, June 24, 2017.

didates. In empirical analysis I illustrate this theoretical proposition with the case of Britain. Using transparency reports I show that higher economic importance is associated with securing higher level of access. Umbrella groups that represent the interests of multiple organizations are also more likely to have more meetings with the government. However, a sudden shock of increased policy uncertainty in the aftermath of the EU referendum, weakens the association between economic importance and level of access. In other words, organizations with less economic importance appear to get more access after the referendum than prior to it. This finding can be a result of a high public scrutiny of the negotiation process. In addition to theoretical contributions, I make several methodological advancements. First, I compile a novel dataset that allows to empirically test government-interest relations with unprecedented precision. Second, I propose a new approach to modelling issues salience using recorded textual data. And, third, I apply a latent variable modelling strategy to estimate economic importance at the level of individual organizations. These findings advance our understanding of interest group access in pluralist systems and can be relevant for many other national contexts beyond Britain.

Bibliography

Acemoglu, D., S. Naidu, P. Restrepo, and J. A. Robinson (2015). Democracy, Redistribution, and Inequality. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution*, Volume 2B, pp. 1885–1966. Amsterdam: Elsevier B.V.

Ainsworth, S. (1993). Regulating Lobbyists and Interest Group Influence. *The Journal of Politics 55*(1), 41–56.

Ansolabehere, S., J. M. de Figueiredo, and J. M. Snyder (2003). Why Is There So Little Money in U.S. Politics? *Journal of Economic Perspectives 17*(1), 105–130.

Austen-Smith, D. (1995). Campaign Contributions and Access. *American Political Science Review 89*(3), 566–581.

Avis, E., C. Ferraz, F. Finan, and U. C. Berkeley (2017). Money and Politics: Estimating the Effects of Campaign Spending Limits on Political Entry and Selection. *NBER Working Paper 23508*.

Barber, M. J. (2015). Ideological Donors, Contribution Limits, and the Polarization of American Legislatures. *Journal of Politics 78*(1), 296–310.

Barro, R. J. (1973). The control of politicians: An economic model. *Public Choice 14*(1), 19–42.

Baumgartner, F. R., C. Breunig, and E. Grossman (2019). *Comparative Policy Agendas: Theory, Tools, Data*. Oxford: Oxford University Press.

Baumgartner, F. R. and B. D. Jones (1993). *Agendas and Instability in American Politics*. Chicago: University of Chicago Press.

Baumgartner, F. R. and B. L. Leech (1998). *Basic Interests: The Importance of Groups in Politics and Political Science*. Princeton, NJ: Princeton University Press.

Beer, S. H. (1956). Pressure Groups and Parties in Britain. *American Political Science Review 50*(1), 496–515.

Benoit, K., M. Laver, and S. Mikhaylov (2009). Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. *American Journal of Political Science 53*(2), 495–513.

Berliner, D. (2014). The Political Origins of Transparency. *Journal of Politics 76*(2), 479–491.

Binderkrantz, A. S., P. M. Christiansen, and H. H. Pedersen (2015). Interest Group Access to the Bureaucracy, Parliament, and the Media. *Governance 28*(1), 95–112.

Blanes I Vidal, J., M. Draca, and C. Fons-Rosen (2012). Revolving Door Lobbyists. *American Economic Review 102*(7), 3731–3748.

Blei, D. M. and J. D. Lafferty (2006a). Correlated Topic Models. In *Advances in Neural Information Processing Systems 18*.

Blei, D. M. and J. D. Lafferty (2006b). Dynamic Topic Models. In *International Conference on Machine Learning*, pp. 113–120.

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

Blick, A. (2014). The Cabinet Manual and the Codification of Conventions. *Parliamentary Affairs 67*(1), 191–208.

Budge, I., D. Robertson, and D. Hearl (1987). *Ideology, Strategy, and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, P. Li, and A. Riddell (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software 76*(1).

Carreras, M., Y. I. Carreras, and S. Bowler (2019). Long-Term Economic Distress, Cultural Backlash, and Support for Brexit. *Comparative Political Studies* (Forthcoming).

Cave, T. and A. Rowell (2015). *A Quiet Word: Lobbying, Crony Capitalism and Broken Politics in Britain*. London: Vintage.

Clarke, H. D., M. J. Goodwin, and P. F. Whiteley (2017). *Brexit: Why Britain Voted to Leave the European Union*. Cambridge, UK: Cambridge University Press.

Colantone, I. and P. Stanig (2018). Global Competition and Brexit. *American Political Science Review 112*(2), 201–218.

Dahl, R. (1961). *Who Governs?: Democracy and Power in an American City*. New Haven: Yale University Press.

Dommett, K., A. Hindmoor, and M. Wood (2017). Who meets whom: Access and lobbying during the coalition years. *British Journal of Politics and International Relations 19*(2), 389–407.

Eckstein, H. (1960). *Pressure Group Politics: The Case of the British Medical Association*. Stanford, California: Stanford University Press.

Eggers, A. C. and J. Hainmueller (2009). MPs for Sale? Returns to Office in Postwar British Politics. *American Political Science Review 103*(4), 513–533.

Evans, G. and A. Menon (2017). *Brexit and British Politics*. Cambridge, UK: Polity.

Faccio, M. (2006). Politically Connected Firms. *American Economic Review 96*(1), 369–386.

Fariss, C. J. (2014). Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability. *American Political Science Review 108*(2), 297–318.

Ferejohn, J. A. (1986). Incumbent Performance and Electoral Control. *Public Choice 50*(1-3), 5–25.

Fouirnaies, A. (2018). How Do Campaign Spending Limits Affect Electoral Competition? Evidence from Great Britain 1885-2010. *Working Paper*.

Fouirnaies, A. and A. B. Hall (2014). The Financial Incumbency Advantage: Causes and Consequences. *Journal of Politics 76*(3), 711–724.

Fouirnaies, A. and A. B. Hall (2018). How Do Interest Groups Seek Access to Committees? *American Journal of Political Science 62*(1), 132–147.

Gelman, A. and G. King (1993). Why Are American Presidential Election Campaign Polls so Variable When Votes Are so Predictable? *British Journal of Political Science 23*(4), 409–451.

Gilens, M. and B. I. Page (2014). Testing Theories of American Politics: Elites, Interest Groups, and Average Citizens. *Perspectives on Politics 12*(3), 42.

Grant, W. (1978). Insider groups, outsider groups and interest group strategies in Britain. *University of Warwick Department of Politics Working Paper no. 19*.

Grant, W. (2000). *Pressure Groups and British Politics*. Basingstoke: Macmillan Press.

Greene, D. and J. P. Cross (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis 25*(1), 77–94.

Grimmer, J. and B. M. Stewart (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*(3), 267–297.

Grossman, G. M. and E. Helpman (2001). *Special Interest Politics*. Cambridge, Massachussets: The MIT Press.

Hall, R. L. and A. V. Deardorff (2006). Lobbying as Legislative Subsidy. *American Political Science Review 100*(1), 69–84.

Hamilton, A., J. Madison, and J. Jay (2008 [1787]). *The Federalist Papers*. Oxford: Oxford University Press.

Hansen, J. M. (1991). *Gaining Access: Congress and the Farm Lobby, 1919-1981*. Chicago: University of Chicago Press.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Heckman, J. J. (1978). Heterogeneity and State Dependence. In S. Rosen (Ed.), *Studies in Labor Markets*, pp. 91–140. Chicago: University of Chicago Press.

Hobolt, S. B. (2016). The Brexit vote: a divided nation, a divided continent. *Journal of European Public Policy 23*(9), 1259–1277.

Jacobson, G. C. (1978). The Effects of Campaign Spending in Congressional Elections. *American Political Science Review 72*(2), 469–491.

Jacobson, G. C. and J. L. Carson (2015). *The Politics of Congressional Elections* (9th ed.). Lanham, Maryland: Rowman & Littlefield.

Judd, G. (2019). Access and Lobbying in Legislatures. *Working Paper*.

Kalla, J. L. and D. E. Broockman (2016). Campaign Contributions Facilitate Access to Congressional Officials: A Randomized Field Experiment. *American Journal of Political Science 60*(3), 545–558.

Lewis-Beck, M. S. and M. Stegmaier (2000). Economic Determinants of Electoral Outcomes. *Annual Review of Political Science 3*(1), 183–219.

Lewis-Beck, M. S. and M. Stegmaier (2007). Economic Models of Voting. In D. Russell and H.-D. Klingemann (Eds.), *The Oxford Handbook of Political Behavior*, pp. 518–37. Oxford: Oxford University Press.

Li, Z. (2018). How Internal Constraints Shape Interest Group Activities: Evidence from Access-Seeking PACs. *American Political Science Review 112*(4), 792–808.

Maloney, W. A., G. Jordan, and A. M. McLaughlin (1994). Interest Groups and Public Policy: The Insider/Outsider Model Revisited. *Journal of Public Policy 14*(1), 17.

Martin, A. D. and K. M. Quinn (2002). Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999. *Political Analysis 10*(2), 134–153.

Mikhaylov, S., M. Laver, and K. Benoit (2012). Coder Reliability and Misclassification in the Human Coding of Party Manifestos. *Political Analysis 20*(1), 78–91.

Milbrath, L. W. (1964). *The Washington Lobbyists*. Chicago: Rand McNally.

Ohman, M. (2012). *Political Finance Regulations Around the World. An Overview of the International IDEA Database*. International Institute for Democracy and Electoral Assistance.

Olson, M. (1965). *The Logic of Collective Action : Public Goods and the Theory of Groups*. Cambridge, Massachussets: Harvard University Press.

Palmer, M. and B. Schneer (2019). Postpolitical Careers: How Politicians Capitalize on Public Office. *Journal of Politics 81*(2), 670–675.

Powell, E. N. and J. Grimmer (2016). Money in Exile: Campaign Contributions and Committee Access. *The Journal of Politics 78*(4), 974–988.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science 54*(1), 209–228.

Rix, K. (2008). ' The Elimination of Corrupt Practices in British Elections'? Reassessing the Impact of the 1883 Corrupt Practices Act. *English Historical Review 123*(500), 65–97.

Schattschneider, E. E. (1960). *The Semisovereign People: A Realist's View of Democracy in America*. New York: Holt, Rinehart and Winston.

Schlozman, K. L., S. Verba, and H. E. Brady (2012). *The Unheavenly Chorus: Unequal Political Voice and the Broken Promise of American Democracy*. Princeton: Princeton University Press.

Schnakenberg, K. E. (2017). Informational Lobbying and Legislative Voting. *American Journal of Political Science 61*(1), 129–145.

Siaroff, A. (1999). Corporatism in 24 industrial democracies: Meaning and measurement. *European Journal of Political Research 36*(2), 175–205.

Snyder, J. M. and M. M. Ting (2008). Interest groups and the electoral control of politicians. *Journal of Public Economics 92*(3-4), 482–500.

Soroka, S. N., D. A. Stecula, and C. Wlezien (2015). It's (Change in) the (Future) Economy, Stupid: Economic Indicators, the Media, and Public Opinion. *American Journal of Political Science 59*(2), 457–474.

Stegmueller, D. (2013). Modeling Dynamic Preferences: A Bayesian Robust Dynamic Latent Ordered Probit Model. *Political Analysis 21*(3), 314–333.

Truman, D. B. (1951). *The Governmental Process: Political Interests and Public Opinion*. New York: Alfred A Knopf.

Volkens, A., J. Bara, I. Budge, M. D. McDonald, and H.-D. Klingemann (2013). *Mapping Policy Preferences from Texts III: Statistical Solutions for Manifesto Analysts*. Oxford: Oxford University Press.

Wright, J. R. (1990). Contributions, Lobbying, and Committee Voting in the U.S. House of Representatives. *American Political Science Review 84*(2), 417–438.

Appendix

A Money in the UK Politics

Campaign Spending

In Figure 2 of the main text I show graphically the relationship between the marginality of electoral loss or victory and the level of campaign spending. To investigate this relationship further and also focus on the most competitive candidates, I model the level of spending limits reached by winning candidates as a function of the margin of their victory. For this I use the same data covering the UK General Elections from 2005 to 2017 as in the main text. The results in Table A.1 show a sizable effect of marginality even when controlling for candidate's party and different characteristics of the constituency. For example, a candidate, who received 10% higher vote share than a runner-up can be expected to have reached 5% less of the campaign spending threshold, holding everything else constant. In other words, winning candidates running in very safe areas can be expected to spend well below their permissible threshold.

Patterns of Donations

To investigate the effect of reduced incentives to engage in fund-raising reflected on the patterns of donations, I analyse the donations data released by the UK Electoral Commission[35]. As political candidates are less invested in raising campaign funds, fewer interest groups are likely to donate and their donations can be expected to be smaller. Figure A.1 shows the monthly trends in the number of donors and amount of donations to sitting MPs from 2001 until 2017. Despite a few noticeable spikes, especially for Labour in the run-up to the 2015 General Election, the trend remains almost flat over the entire period. Notably, no more than 80 unique donors are recorded within any given month, with a typical number well below 20 even for the two main parties.

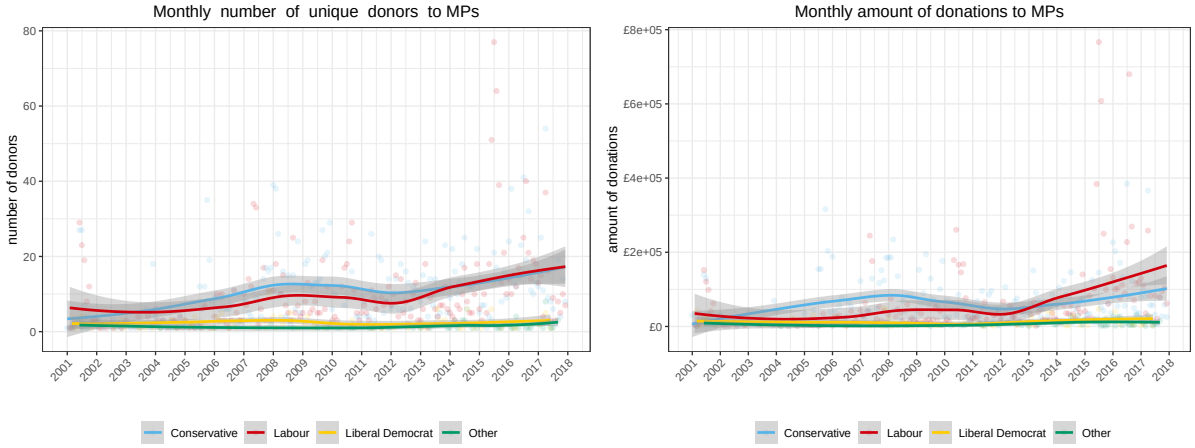
[35] https://www.electoralcommission.org.uk/

Table A.1: OLS models of spending limit by winning candidates in 2005-2017 UK PGE.

	Percentage of spending limit			
	(1)	(2)	(3)	(4)
Margin of victory (%)	-0.5329***	-0.5491***	-0.5307***	-0.5049***
	(0.0283)	(0.0279)	(0.0279)	(0.0282)
Size of electorate (,000)		0.2633***	0.2468***	0.27***
		(0.0424)	(0.0432)	(0.0463)
Constituency (ref: Borough)				
Burgh		-7.9484***	-1.0894	-2.9966
		(2.3583)	(2.4876)	(2.8268)
County		-5.0336***	-5.8272***	-6.4495***
		(0.807)	(0.864)	(0.9181)
Party (ref: Conservative)				
DUP			-1.3927	14.2272
			(3.3445)	(15.1854)
Green			15.0177	15.3858
			(11.2555)	(11.1581)
Independent			-21.1488*	-12.2592
			(8.7224)	(12.413)
Labour			-3.9337***	-4.9582***
			(0.9138)	(0.9531)
Liberal Democrats			9.1599***	7.3085***
			(1.7815)	(1.8273)
Other			10.098	15.8458
			(13.7787)	(15.5568)
Plaid Cymru			16.2439**	16.0072**
			(5.749)	(5.8521)
SDLP			3.7783	19.8952
			(6.5129)	(16.1705)
Sinn Féin			2.8073	18.2418
			(4.2881)	(15.4148)
SNP			-17.1228***	-15.8917***
			(2.1818)	(2.7185)
Speaker			20.7055*	20.0578*
			(9.7642)	(9.6791)
UKIP			14.5709	19.3599
			(19.4476)	(19.2897)
UUP			-1.7577	16.5303
			(11.2485)	(18.5521)
Country (ref: England)				
Northern Ireland				-15.9302
				(14.8277)
Scotland				1.3389
				(2.0653)
Wales				0.4155
				(1.8389)
General Election (ref: 2005)				
2010				-2.3493*
				(1.0828)
2015				-7.6318***
				(1.1319)
2017				-2.4303*
				(1.1188)
Constant	84.6243***	69.5292***	72.2761***	73.9441***
	(0.7258)	(3.0658)	(3.3111)	(3.5039)
Observations	2582	2582	2582	2582
R^2	0.12	0.15	0.19	0.21

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Figure A.1: Donation patterns in the UK



B Transparency Policy

As the exact transparency policy provisions might affect the quality and nature of collected data, some further contextual information can improve our understanding of the data. In particular, I present two primary sources: verbatim detailed instructions provided for department employees on how the transparency reports should be compiled, which gives an insight into how these variables are formed. These instructions are quoted below:

```
__Minister__
Include name and title of all ministers who have attended external meetings
including with Newspaper and other media proprietors, editors or senior executives.
NOTE:
1) Don't create separate documents for individual ministers.
2) Do include ministers who have nothing to record (nil return).

__Date__
Record month and year. Use the format: April 2014

__Name of organisation or individual__
Include any group, company, organisation or person not from or connected to governmen

Include all official, political or personal meetings with newspaper and other
media proprietors, editors and senior executives as follow:

Proprietors
Newspapers: Chair, owner
Broadcasters: Chairmen
```

```
Editors
Newspapers: The editor
Broadcasters: Editors (including political editors), channel controllers,
directors of programming, radio controllers

Senior Executives
Newspapers: CEOs
Broadcasters: Director Generals, CEOs

Meetings with individuals from media organisations operating below this level
should be included but shown as the name of the organisation only.
Political and personal meetings at this level should not be included.

NOTE:
1) If meeting was with multiple organisations, list them separately or use a collecti
2) Don't include meetings held in a party or constituency capacity.
Contact Cabinet Office if you are unsure about whether to record a meeting.

__Purpose of meeting__
Briefly describe topic or objective of meeting. Do not use 'general discussion'.
```

In addition to clerical instructions, figure B.2 shows section 8.14 of the Ministerial Code, which is specifically focussed on the meetings with external organisations. The latest version of the Ministerial Code is available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/672633/2018-01-08_MINISTERIAL_CODE_JANUARY_2018__FINAL___3_.pdf

Figure B.2: Section 8.14 of the Ministerial Code

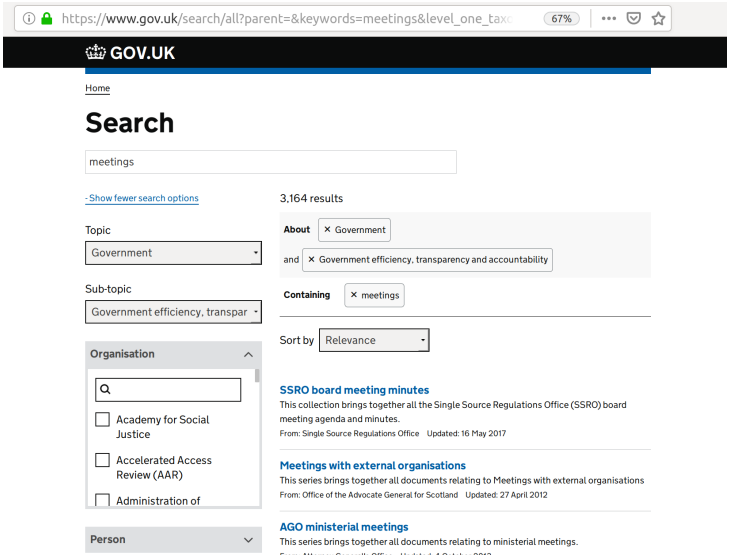
Meetings with external organisations

8.14 Ministers meet many people and organisations and consider a wide range of views as part of the formulation of Government policy. Meetings on official business should normally be arranged through Ministers' departments. A private secretary or official should be present for all discussions relating to Government business. If a Minister meets an external organisation or individual and finds themselves discussing official business without an official present – for example at a social occasion or on holiday – any significant content should be passed back to the department as soon as possible after the event. Departments will publish quarterly, details of Ministers' external meetings. Meetings with newspaper and other media proprietors, editors and senior executives will be published on a quarterly basis regardless of the purpose of the meeting.

C Data Collection

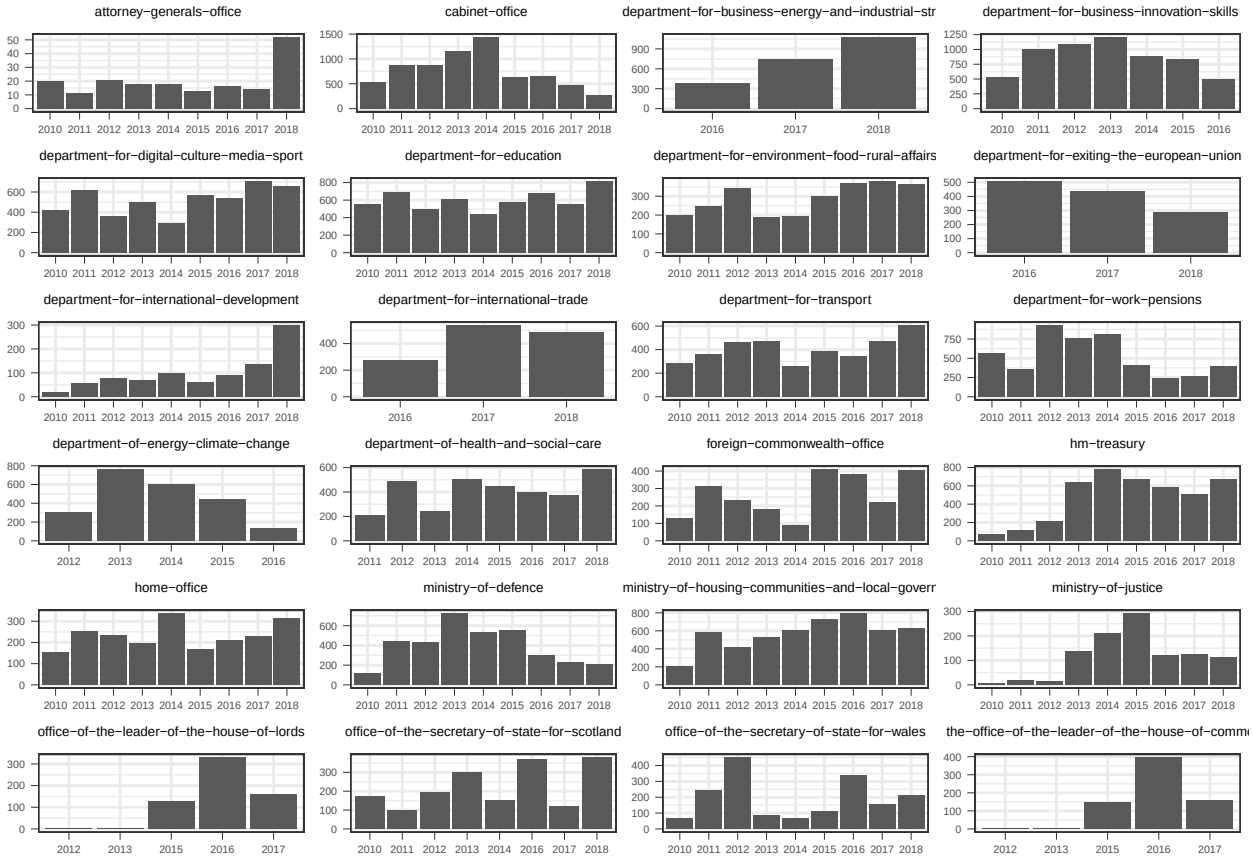
The transparency reports used in the analysis were obtained using the official government portal https://www.gov.uk/search/all? and its previous version https://www.gov.uk/government/publications. The interface of the core search engine is shown in figure C.3.

Figure C.3: Main page of the gov.UK search engine



A specially designed R package was used to systematically collect the data from the earlier version of the government publications website.

Figure D.4: Annual number of meetings by department



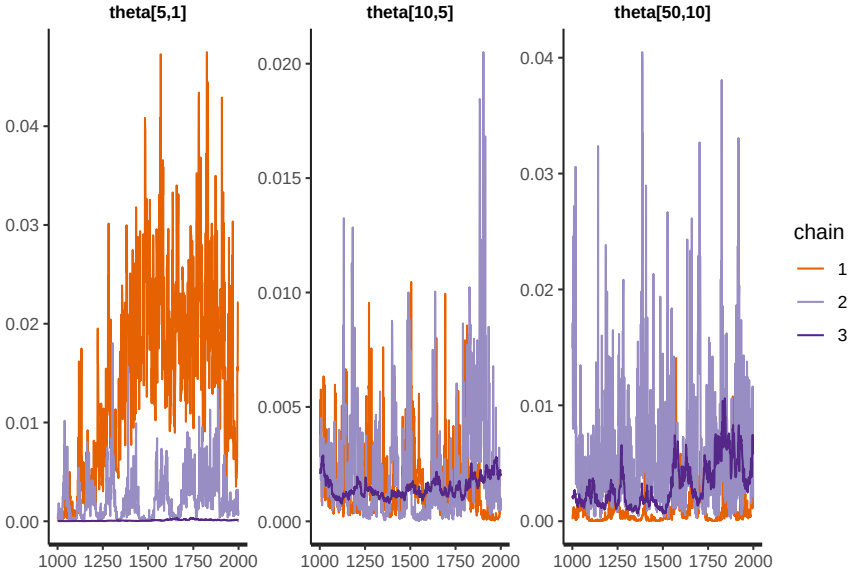
D Descriptive statistics

Figure D.4 shows the annual number of meetings held separately by each government agency.

E Dynamic Topic Model of Issue Salience

Figure E.5 shows traceplots of several document-level θ coefficients.

Figure E.5: Convergence diagnostics of document parameters



F Model of Economic Importance

Dynamic Factor Model

To diagnose the convergence of the dynamic factor model I use Gelman-Rubin R-hat measure and traceplots of β_j auto-regressive coefficients for three economic indicators: *assets*, *employees* and *revenue* from the measurement model:

$$y_{ijt} = \alpha + \beta_j \lambda_{it} + \epsilon_j$$

I took the logarithms of all indicators prior to fitting the model to ensure the normal distribution of the left-hand side variables. To estimate the model I use standard uniformative prior specification:

$$\delta \sim N(0, 10)$$

$$\xi_k \sim N(0, 10)$$

$$\theta \sim N(0, 10)$$

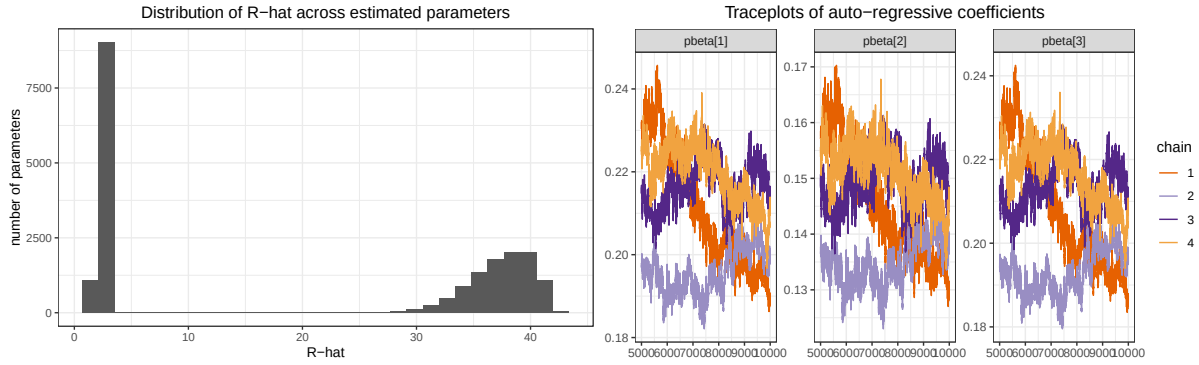
$$\gamma \sim N(0, 10)$$

$$\alpha \sim N(0, 10)$$

$$\beta_j \sim N(0, 10)$$

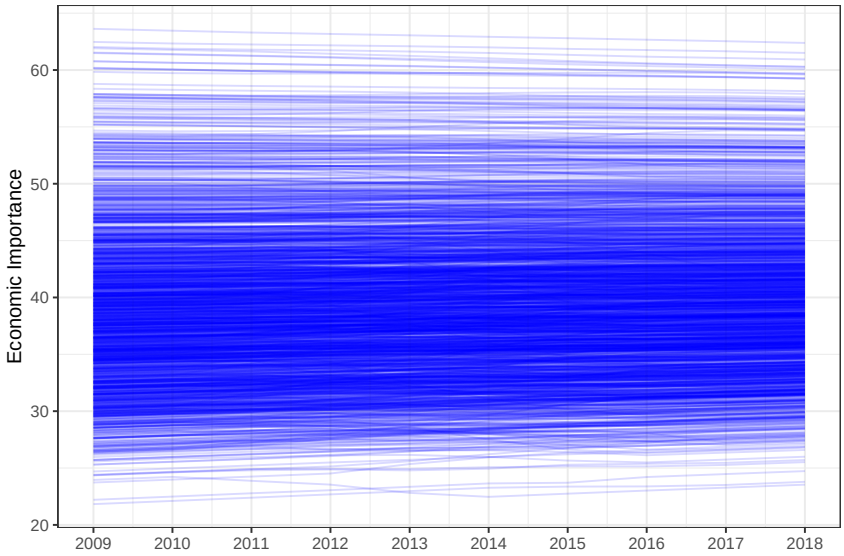
Figure F.6 shows the model convergence diagnostics. Of the rougly 20,000 estimated coefficients, the vast majority have low R-hats and the traceplots show moderately good mixture.

Figure F.6: Model convergence diagnostics



The dynamic part of the latent economic importance ensures the smoothness in any between-year transition. While most of the indicators can reasonably be assumed to change only slightly, to check whether variation of the estimated parameters remains constant over time, I plot the overall trajectories for the organizations included in the analysis. As Figure F.7 indicates, despite some large cross-year changes for a small number of organizations, between-organization variation remains constant over time.

Figure F.7: Trajectories of economic importance



Stan Code

The Stan code used for fitting dynamic factor analysis model to measure economic importance is provided below:

```
data {
  int<lower=0> NN; // number of non-missing observations,
  int<lower=0> NM; // number of missing observations
  int<lower=0> T; // length of time series
  int<lower=0> J; // number of measures
  int<lower=0> K; // number of latent trends/interest groups
  int<lower=0, upper=T+1> tt[NN + NM]; // time period id
  int<lower=0, upper=J> jj[NN + NM]; // measure id
  int<lower=0, upper=K> kk[NN + NM]; // interest group id
  int<lower=0, upper=NN + NM> nn[NN]; // index of non-missing observations
  int<lower=0, upper=NN + NM> mm[NM]; // index of missing observations
  real yn[NN]; //vectorized matrix of log-transformed observed values
}
transformed data {
  int<lower=0> N = NN + NM;
}
parameters {
  real lambda[K,T]; // matrix of latent economic importance
  real beta[J]; // vector of factor loadings, constrained to be positive
  real ym[NM]; // vector of missing observations
  real alpha; // measurement intercept
  real delta; // initial condition scale factor
  real ksi[K]; // initial condition fixed effect
  real theta; // latent intercept
```

```
  real gamma; // autoregressive coefficient
  // real<lower=0> sigma_lambda; // variance of latent factor
  // real<lower=0> sigma_y; // variance of measurement instrument
}
transformed parameters {
  // combine observed and missing economic indicators
  real y[N];
  y[nn] = yn;
  if (NM > 0) {
    y[mm] = ym;
  }
}
model {
  for (n in 1:N) {
    if (tt[n] > 1) {
      lambda[kk[n],tt[n]] ~ normal(theta + gamma * lambda[kk[n],tt[n]-1], 1);
      y[n] ~ normal(alpha + beta[jj[n]] * lambda[kk[n],tt[n]], 1);
    } else {
      // initial condition
      lambda[kk[n],tt[n]] ~ normal(delta * ksi[kk[n]], 1);
      // lambda[kk[n],tt[n]] ~ cauchy(0, 5);
    }
  }

  // Priors on coefficients
  delta ~ normal(0, 10);

  for (k in 1:K) {
```

```
    ksi[k] ~ normal(0, 10);
  }

  theta ~ normal(0, 10);

  gamma ~ normal(0, 10);

  alpha ~ normal(0, 10);

  for (j in 1:J) {
    beta[j] ~ normal(0, 10);
  }

  // Hyper-priors for variance parameters

  // sigma_lambda ~ cauchy(0,3);
  // sigma_y ~ cauchy(0,3);
}
generated quantities {
  // post-multiply latent factors and loadings to ensure positivity
  real plambda[K,T];
  real pbeta[J];
  for (k in 1:K) {
    for (t in 1:T) {
      if (lambda[k,t] < 0) {
        plambda[k,t] = -lambda[k,t];
      } else {
        plambda[k,t] = lambda[k,t];
```

```
      }
    }
  }
  for (j in 1:J) {
    if (beta[j] < 0) {
      pbeta[j] = -beta[j];
    } else {
      pbeta[j] = beta[j];
    }
  }
}
```

G Full Models of Access

Table G.2 shows the complete Poisson models from the main text, including exhaustive list of coefficients for individual sectors.

Table G.2: Poisson models of the number of interest group meetings.

	Meetings (12 months pre/post)			Meetings (24 months pre/post)		
	(1)	(2)	(3)	(4)	(5)	(6)
Post-Referendum	0.0056	-0.0593	0.0307	0.0367*	0.3044*	0.3111*
	(0.023)	(0.1742)	(0.1742)	(0.0169)	(0.1276)	(0.1275)
Economic Importance	0.0585***	0.0589***	0.0578***	0.0758***	0.0799***	0.0772***
	(0.0019)	(0.0026)	(0.0027)	(0.0014)	(0.0019)	(0.002)
Age	7e-04*	-3e-04	6e-04	7e-04**	-3e-04	5e-04
	(3e-04)	(5e-04)	(5e-04)	(2e-04)	(4e-04)	(4e-04)
Umbrella	0.4976***	0.4651***	0.4199***	0.6355***	0.5836***	0.5545***
	(0.0284)	(0.0398)	(0.0411)	(0.0207)	(0.0302)	(0.0314)
Post-Referendum * Importance		-3e-04	-0.002		-0.0079**	-0.0076**
		(0.0038)	(0.0038)		(0.0028)	(0.0028)
Post-Referendum * Age		0.0019**	0.0019**		0.002***	0.0016**
		(7e-04)	(7e-04)		(5e-04)	(5e-04)
Post-Referendum * Umbrella		0.0662	0.0364		0.0959*	0.0874*
		(0.0568)	(0.0578)		(0.0415)	(0.0423)
Sector (ref: Banking)						
Chemicals			-0.5***			-0.763***
			(0.1342)			(0.0978)
Construction			0.396***			0.326***
			(0.1123)			(0.0783)
Education/Health			0.2251**			0.0235
			(0.0864)			(0.0581)
Food/Beverages/Tobacco			-0.814***			-1.0744***
			(0.1898)			(0.145)
Gas/Water/Electricity			0.1881			0.1772*
			(0.1071)			(0.0716)
Hotels/Restaurants			-0.6677**			-0.7207***
			(0.2315)			(0.1356)
Insurance			-0.7292			-1.1121
			(0.5826)			(0.5797)
Machinery/Recycling			0.7289***			0.5928***
			(0.0863)			(0.0583)
Metal			-0.5501*			-0.7277***
			(0.2193)			(0.182)
Other services			0.3441***			0.2389***
			(0.0817)			(0.0541)
Post/Telecommunications			0.913***			0.8338***
			(0.0954)			(0.0653)
Primary sector			0.0722			-0.1502
			(0.1705)			(0.13)
Public administration/Defense			-0.0722			-0.0175
			(0.1588)			(0.1201)
Publishing			0.8318***			0.7159***
			(0.1122)			(0.0783)
Textiles			-0.1494			-0.6281
			(0.4154)			(0.3371)
Transport			-0.0413			-0.0581
			(0.0965)			(0.0657)
Wholesale/Retail trade			-0.2841**			-0.469***
			(0.1094)			(0.0763)
Wood			0.3251			-0.04
			(0.2707)			(0.2637)
Constant	-0.5661***	-0.5457***	-0.7934***	-1.0833***	-1.222***	-1.2763***
	(0.0869)	(0.1159)	(0.155)	(0.064)	(0.0885)	(0.1104)
Observations	894	894	870	1266	1266	1236

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

H Robustness checks

Placebo Referendum

Table H.3 shows the results of a placebo test, with June 2014 used as the Brexit date and the data restricted to meetings that occured 12 months before and after the threshold.

Table H.3: Poisson models using placebo Brexit (June 2014) as a cutoff.

	Meetings (12 months pre/post)					
	Poisson			Negative Binomial		
	(1)	(2)	(3)	(4)	(5)	(6)
Post-Referendum	-0.0966***	-0.1636	-0.1126	-0.0899	-0.0469	0.0042
	(0.0229)	(0.169)	(0.1681)	(0.055)	(0.3947)	(0.3874)
Economic Importance	0.0778***	0.0767***	0.0724***	0.0653***	0.0654***	0.0609***
	(0.0018)	(0.0024)	(0.0026)	(0.0046)	(0.0062)	(0.0064)
Age	-0.0011**	-4e-04	0	-3e-04	2e-04	4e-04
	(4e-04)	(5e-04)	(5e-04)	(0.001)	(0.0013)	(0.0013)
Umbrella	0.5914***	0.623***	0.6145***	0.4766***	0.4867***	0.447***
	(0.0308)	(0.0417)	(0.0425)	(0.0738)	(0.1009)	(0.1004)
Post-Referendum * Importance		0.0027	0.0013		-2e-04	-0.0018
		(0.0037)	(0.0036)		(0.0092)	(0.009)
Post-Referendum * Age		-0.0015*	-6e-04		-0.0012	-5e-04
		(7e-04)	(7e-04)		(0.0019)	(0.0019)
Post-Referendum * Umbrella		-0.0686	-0.0943		-0.0218	-0.0369
		(0.0618)	(0.0624)		(0.148)	(0.1451)
Sector (ref: Banking)						
Chemicals			-0.5739***			-0.5234
			(0.152)			(0.2935)
Construction			-0.0255			-0.0184
			(0.1079)			(0.2465)
Education/Health			0.1523*			0.0656
			(0.0708)			(0.1579)
Food/Beverages/Tobacco			-0.9048***			-0.9229**
			(0.1842)			(0.3328)
Gas/Water/Electricity			0.2455**			0.1882
			(0.0818)			(0.2029)
Hotels/Restaurants			-0.7364***			-0.7174*
			(0.1601)			(0.3082)
Insurance			-0.326			-0.4496
			(0.2952)			(0.5336)
Machinery/Recycling			0.7671***			0.4261*
			(0.0714)			(0.1919)
Metal			-1.054***			-0.9794*
			(0.296)			(0.4888)
Other services			0.1229			0.0559
			(0.0652)			(0.1488)
Post/Telecommunications			0.7749***			0.8699***
			(0.0816)			(0.2446)
Primary sector			-0.4828*			-0.5868
			(0.1936)			(0.3382)
Public administration/Defense			-0.3747*			-0.4063
			(0.1721)			(0.3244)
Publishing			1.0323***			0.723*
			(0.0973)			(0.295)
Textiles			-0.6824			-0.7798
			(0.4126)			(0.6733)
Transport			0.22**			0.0482
			(0.0801)			(0.2004)
Wholesale/Retail trade			-0.5845***			-0.5913**
			(0.1041)			(0.2141)
Wood			-0.9825*			-1.0443
			(0.4128)			(0.6736)
Constant	-1.3744***	-1.3491***	-1.3182***	-0.8283***	-0.8464**	-0.7098*
	(0.0839)	(0.1105)	(0.1371)	(0.1968)	(0.2629)	(0.3146)
Observations	973	973	955	973	973	955

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

I Software statement

The computational part of the study was conducted under Linux Ubuntu 18.04 using R version 3.6.0 (R Core Team 2019). I used the following R packages in my empirical analysis:

`dplyr` (Wickham et al. 2018),

`ggplot2` (Wickham 2016),

`kableExtra` (Zhu 2018),

`knitr` (Xie 2018),

`lubridate` (Grolemund and Wickham 2011),

`magrittr` (Bache and Wickham 2014),

`readr` (Wickham, Hester, and Francois 2017),

`rstan` (Stan Development Team 2018),

`stringi` (Gagolewski 2018),

`stringr` (Wickham 2018),

`tibble` (Müller and Wickham 2018),

`tidyr` (Wickham and Henry 2018), and

`quanteda` (Benoit et al. 2018).

Chapter 3

Gender, Justice and Deliberation: Women's Presence without Influence in Peace-making

ABSTRACT

Scholars have pinpointed that women's underrepresentation in peace-making results in gendered outcomes that do not address women's needs and interests. Despite recent increased representation at the negotiating table, women still have a limited influence on peace-making outcomes. We propose that differences in female and male speeches reflected in the gendered patterns in discourse during peace-making explain how women's influence is curtailed. We examine women's speaking behavior in transitional justice debates in the post-conflict Balkans. Applying multi-method quantitative text analysis to over half a million words in multiple languages, we analyze structural and thematic speech patterns. We find that men's domination of turn-taking and the absence of topics reflecting women's needs and interests lead to a gendered outcome; the sequences of men talking after men are longer than those of women talking after women, which restricts women's deliberative space and opportunities to develop and sustain arguments that reflect their concerns. We find no evidence that women's limited influence is driven by lower deliberative quality of their speeches. This study of gendered dynamics at the micro-level of discourse identifies a novel dimension of male domination during peace-making.

3.1 Introduction

Peace is much more than the cessation of violence. Scrutiny of the quality of peace has revealed that peace often fails women (Wallensteen, 2015, 45). The end of a conflict provides an opportunity to lay the foundations for gender-just peace, which transforms unequal gender relations providing for women's political, social, and economic agency (Björkdahl and Mannergren, 2013; Lake, 2018). However, post-conflict peace-making also introduces norms, structures, and power relations that disadvantage women. Some of these are an extension of gendered conflict dynamics, stemming from different experience of violence by men and women (Melander, 2016); others entail a reversal of women's wartime gains in political and social agency (Berry, 2018; Tripp, 2015; Kreft, 2019; Østby et al., 2019). An imperative to make peace work for women has motivated scholars and practitioners to tackle gender inequality during peace-making. Women's participation in peace processes matters; it is associated with longer and better peace (Melander, 2005; Gizelis, 2009; Demeritt et al., 2014; Krause et al., 2018). Transitional justice is integral to peace (Sharp, 2013). Thus, peace that works for women advances women's representation and rights, including the right to justice for wartime sexual and gender-based violence (SGBV) (Chinkin and Kaldor, 2013)[1].

Women's presence in peace efforts is critical for bringing about gender-just peace, because women's presence provides women "communicative advantages" (Mansbridge, 1999, 642). Women can insert their perspectives into the peace-making process, which paves the way for recognition of their needs (Brown and Aoláin, 2015, 147). Criticism of women's marginalization in peace-making has resulted in international and national efforts to include women in peace processes (Adjei, 2019). However, women's increased presence in these processes has had only a limited impact on their outcomes. Scholars have shown how provisions of peace agreements and mandates of transitional justice instruments overlook or marginalize women's needs, interests and entitlements (Bell and

[1] A discussion of SGBV against men or SGBV perpetrated by women is beyond the scope of this article (Schulz, 2020)

O'Rourke, 2010; Burnet, 2008; Haynes et al., 2011; Hogg, 2009; Sandole and Staroste, 2015; Borer, 2009). The failure to bring about gender-just peace in contexts where women are represented in the peace-making process poses a significant puzzle (Castillo Diaz and Tordjman, 2012; Paffenholz et al., 2016; UN Women, 2015).

We address this puzzle of women's representation without influence by examining the black-box between women's representation indicated by 'bodies at the table' and outcomes that reflect women's concerns. It is important to understand women's contributions to debates during peace-making. A gendered pattern of speaking behavior, that reveals differences in speech by men and women, has a "cumulative effect on power and influence" (Kathlene, 1994, 573). We study how discourse in a transitional justice process is gendered since recognition of women's justice needs is part of "gender-sensitive and gender-responsive" perspectives on peace (Davies and True, 2019; UN Women, 2015). Gender is a political, social, and cultural construction, which should not be conflated with the sex identity of women and men (Carver, 1996). The concept of gender makes visible how behaviors, norms, and discourses associated with the female and male sexes in institutional and informal process are implicated in the production and reproduction of inequality and oppression (Sandole and Staroste, 2015, 119-120; Krook and Mackay, 2011; Cameron, 1998, 4). Our analysis identifies how gender-based differences in discourse during peace-making account for a gendered outcome that does not reflect women's needs, entitlements and interests.

To explain the lack of women's influence despite their broadly equal representation, we propose and test three mechanisms operating in discourse during peace-making: deliberation, emboldending, and de-centering. The first mechanism concerns how women's quality of deliberation, focused on justification of arguments, compares with that of men's. It is premised on the feminist critique of democratic deliberation and institutions that male domination is perpetuated by prescribed modes of communication (Sanders, 1997; Mackay et al., 2010; Acker, 1990). Foregrounding the structure of discourse and focusing on turn-taking sequences, the second mechanism probes whether women are emboldened

to put forth their views in deliberative enclaves (Sunstein, 2007). It assumes that when women speak after each other in succession in mixed-sex debates, a deliberative space is created that is conducive to the articulation of their needs. De-centering is the third mechanism. Considering that women often de-center, i.e. they avoid discussing directly violence that they have suffered (Kashyap, 2009; Theidon, 2013; Porter, 2016), we explore whether thematic differences in speeches by men and women can explain a gendered outcome that marginalizes women's interests and concerns.

We find that, in conditions where women are broadly represented equally, gendered outcomes that disadvantage women result from men's domination of turn-taking and the absence of topics reflecting women's concerns and interests; in mixed-sex debates, the sequences of men talking after men are longer than those of women talking after women, which restricts women's deliberative space and opportunities to develop and sustain arguments that reflect their needs and entitlements. We find no evidence that women's limited influence is driven by lower deliberative quality of their speeches. This research shows that a micro-level of discourse during peace-making is a domain of male domination that has been overlooked by scholars and practitioners puzzled by the elusive influence of women who are present at the peace-making table.

We use a case of a civil society-led transitional justice process in the post-conflict Balkans, known by its acronym RECOM,[2] to scrutinize gender differences in discourse. From 2010 to 2011, the multi-ethnic initiative organized debates dedicated to designing the mandate of the regional fact-finding commission, which had emerged as a preferred transitional justice approach in previous rounds of consultations. These debates produced the commission's draft Statute[3]. Defining the commission's mandate, the document failed to respond to women's concerns, needs and interests, exposing the gender dimension as a weakness of the RECOM's process (Bonora, 2019). The draft Statute did not provide for women's equal inclusion in different facets of the commission's operation, nor did

[2]RECOM stands for the Regional Commission for Establishing the Facts about War Crimes and other Serious Human Rights Violations in former Yugoslavia from January 1991 to the end of December 2001.

[3]See Statut Koalicije za Rekom, 29 June 2011 at http://recom.link/wp-content/uploads/2011/06/Statut-Koalicije-za-REKOM-26.06.2011-SRB.pdf

it envisage appropriate procedures to facilitate recognition of women victims of SGBV (although SGBV was listed among abuses to be investigated). To advance gender justice, transitional justice instruments also need to include appropriate gender-responsive procedures (Swaine, 2018, 231-232), for example for staging women's testimonies about their experience of violence. Why did seemingly vocal contributions of women present in RECOM's debates not lead to the commission's mandate that responds to women's needs, concerns and interests? To answer this question, we interrogated whether the patterns in discourse are gendered by applying quantitative content analysis, which involves human coding and computer-assisted text analysis of a corpus of over half a million words comprised of the transcripts of RECOM's debates.

Our evidence drawn from the study of women's speaking behavior that reveals how women's influence is curtailed in a mixed-sex deliberative setting advances research about peace more broadly. First, it demonstrates theoretical benefits of the empirical study of processes that can help ensure gender-just "quality peace" (Waylen, 2014; Wallensteen, 2015), which has lagged behind the study of peace-making outcomes, such as peace-agreements, and their effects. Second, we sound a note of caution about crude measurement of women's participation in peace-making in the existing scholarship and practice (Paffenholz et al., 2016); it captures women's physical presence, describes their roles as signatories or negotiators, and specifies whether they take up senior roles (UN Women, 2015, 45), but neglects more refined measures such as how often they take the floor, how many arguments they make, and how long they speak relative to men. Third, we expose the untapped potential of quantitative analysis of discourse for the study of peace, that contributes to insights gained through qualitative study of discourse and its effects (Jennings, 2019). Quantifying and understanding the gendered patterns of discourse during peace-making can help us devise practical interventions that advance peace that works for women.

In the next part of this article, we review scholarly debates about gendered peace and justice, and outline a critique of existing approaches to women's representation and

limited influence in peace processes. We then present mechanisms in discourse that can explain why women's representation in peace-making does not translate into policies that promote gender equality. The article proceeds with a presentation of data and research design, followed by the results of the analysis of the gendered nature of discourse. The conclusion reflects on the contribution of this study to scholarship and policy.

3.2 Gendered Peace and Justice: (Re-)Assessing Women's Representation in Peace Processes

Inaugurating the Women, Peace and Security (WPS) Agenda, the 2000 UN Security Council Resolution (UNSCR) 1325 prompted critical rethinking by scholars and practitioners about how to bring about gender-just peace (Kirby and Shepherd, 2016, 252). Accounting for "gendered peace" (Pankhurst, 2008), scholars have pinpointed systematic underrepresentation of women in peace processes, despite a slow but steady trend of their greater inclusion following UNSCR 1325 (UN Women, 2015, 45). Only 9 percent of negotiators in 31 major peace processes between 1992 and 2011 were women (Castillo Diaz and Tordjman, 2012). Unsurprisingly, the outcomes of those processes were gendered, in that 16 percent of 585 major peace agreements in 102 peace processes from 1990 to 2010 had references to women and their concerns (Bell and O'Rourke, 2010; Ellerby, 2016). Our understanding of women's influence in processes that define mandates of transitional justice instruments is even more limited, although their proceedings and effects, for example those of the International Criminal Tribunal for the former Yugoslavia (ICTY), are also gendered (Gallagher et al., 2020; King et al., 2017)[4]. Analyzing the absence of women in peace processes and its consequences is important (McLeod, 2019), but we also need to better understand the limited influence of (the increasing number of) women who are present in these processes.

Looking beyond women's proportional representation, scholars have queried the *kind*

[4]Women were underrepresented even in trials involving SGBV charges at the ICTY (Sharratt, 2011).

of representation of women in peace processes. The inclusion of elite women in peace processes who are linked to elite men or clan leaders has resulted in a "veneer of female legitimacy" (Ní Aoláin, 2016) and underrepresentation of non-elite women's concerns. Others stress that women often remain silent during meetings, negotiations, and other events in peace-making process (Ellerby, 2016; Krause et al., 2018). Structural constraints provided an alternative explanation. Bell and O'Rourke (2010, 978) contend that incorporating women's concerns would make it harder to reach an agreement or might destabilize existing agreements in the context of power-sharing. Normative considerations also play a role. Local men in many conflict and post-conflict settings oppose women's emancipation and gender equality, which are often perceived to be externally imposed (Anderlini, 2007; Khodary, 2016).

The WPS Agenda has promoted institutionalization of gender equality norms in peace processes (Adjei, 2019). However, peace-making that now includes more women still produces outcomes that do not adequately reflect women's needs and concerns. The study of gender-just peace can benefit from engaging with scholarship on political representation and communication that examines the gap between women's descriptive and substantial representation. Descriptive representation refers to women's presence in political processes (e.g. national legislatures), while substantive representation captures their influence on policy (Pitkin, 1967). Mendelberg et al. (2014) observe that "even high descriptive representation does not consistently erase [women's] low substantive representation" in deliberative settings. Women's limited impact on policy outcomes is evident in both Western and non-Western countries, such as Rwanda (Devlin and Elgie, 2008). At the same time, women's greater representation may cause a backlash. Kathlene (1994) confirmed Yoder's "intrusiveness thesis" (Yoder, 1991), which holds that men react to women's increased presence in the legislative setting by themselves becoming more vocal[5]. However, underrepresentation can also motivate greater participation by women in political debates by incentivizing women to increase their visibility (Pearson and Dancey, 2011b, 910).

[5]See Karim et al. (2018).

In contrast to scholars of peace-making, scholars of political representation and communication have refined measures of women's participation in public debates. Beyond 'counting women,' they consider the proportion of women's speaking turns to men's, as well as the duration of their speeches (Karpowitz et al., 2012, 21). As Kathlene (1994, 564) points out, "men and women may take an equal number of turns, but men may talk longer than women in any given turn". Men talk more than women in mixed-sex groups and in different conversational contexts (Leaper and Ayres, 2007; Parthasarathy et al., 2019). The shorter length of women's speeches can offset any benefits to descriptive representation – even if there is gender equality in terms of the number of speaking turns. Male dominance in communication holds true even when family issues, which might be expected to stimulate women's participation, are discussed in legislative settings (Kathlene, 1994, 569). Furthermore, during a single speaking turn, speakers may present arguments about one or more policy points[6]. Men's dominance in terms of the number of policy points they address may amplify their impact on policy outcomes; alternatively, women's dominance over policy points may compensate for the fewer speaking turns they have.

These insights from the fields of political representation and communication reveal a need for a more robust assessment of women's representation in peace-making beyond the binaries: women's presence vs. women's absence or women's silence. When addressing the question of why women's representation does not translate into influence in peace-making, assessment of women's representation needs to consider both women's proportional presence and the number of a speaking turns, the duration of their turns, as well as the number of policy points they make. If, women's presence thus re-assessed is (broadly) at parity with men's, and if it fails to translate into influence on policy, we can turn to the analysis of speaking behavior to find out what hinders translation of descriptive representation into substantive representation of women's concerns in peace-making.

[6]Scholars of deliberative democracy take positions on policies as the unit of analysis rather than speaking turns (Steenbergen et al., 2003).

3.3 Presence without Influence in Peace-making: Mechanisms

We contend that understanding speaking behavior is integral to understanding processes that result in gendered peace-making outcomes. As Kathlene (1994, 573) points out, discourse analysis of political discussions can explain the gap between women's representation and women's influence on policy-outcomes (Dolan, 2006; Cowley, 2014; West, 2017; Blumenau, 2019). We propose three mechanisms operating in public discourse that can account for gendered outcomes in peace and justice processes: deliberation, emboldening, and de-centering.

3.3.1 Deliberation

Deliberation spotlights the content of speakers' contributions. Focused on how female and male speakers substantiate their views when they take the floor, scholars of democratic deliberation have studied the quality of speakers' arguments. A reason-giving requirement is at the core of democratic deliberation (Thompson, 2008): speakers provide reasons for their positions and respond to reasons offered by others in an exercise of deliberative reciprocity (Guttmann and Thomson, 1996). Deliberation also entails respect for interlocutors and openness to hearing their views. (Steiner et al., 2005, 22) point out that respect requires empathy: "[t]he capacity and the willingness to put oneself in the shoes of others and to consider a situation from their perspective". Such 'other-regarding' communication embodies the principle of reflexivity. Deliberators reflect on their positions, weighing them in the light of counterarguments (Bächtiger and Steiner, 2005, 156).

Deliberative communication plays an important role in the transition from conflict to peace. Deliberative virtues can help overcome mistrust and polarization in divided societies (Dryzek, 2005; O'Flynn, 2006; Steiner, 2012; Caluwaerts and Deschouwer, 2014). They can also promote justice-seeking, by considering the perspectives of the ethnic Other, but also those of women. Recognition of women's concerns depends on their being

an equal deliberative partner to men. Women's communication styles, including deliberation, can be understood from the prism of the difference/dominance debate (Cameron, 1998, 14-15). The former centers on suitability of deliberation as a mode of communication in terms of women's ways of speaking, while the latter highlights structural underpinnings of male dominance of communication styles.

Difference democrats have pointed out that "[s]ome citizens are better than others at articulating their arguments in rational, reasonable terms" (Sanders, 1997, 349). Recognition that the requirement for dispassionate argument in deliberation particularly disadvantages women has led to calls for valuing diverse models of communication in a democratic discussion, such as greeting, rhetoric, narratives, and story-telling (Sanders, 1997; Young, 2001). Directing attention to the gendered nature of institutions, scholars have posited that communication and language are implicated in the process of control (Mackay et al., 2010, 579-583). Male dominance in these gendered social structures is secured by legitimizing certain rules, norms, and behaviors. The absence of emotionality is prescribed in institutions (Acker, 1990, 151), which restricts the range of permissible forms of articulation of needs and interests and impacts women adversely.

Addressing the question of "gendered deliberation" (Grünenfelder and Bächtiger, 2007), emerging empirical research has not produced compelling evidence that the quality of deliberation in national parliaments differs between men and women (Bächtiger and Hangartner, 2010)[7]. Nonetheless, deliberation points to a possibility that women's substantive marginalization in peace-making may be driven by different quality of argumentation between men and women.

3.3.2 Emboldening

Scholars have highlighted gendered differences in the use of language between women and men "in terms of both what they say and how they say it" (Krook, 2010, 233). However,

[7]In fact, women are more able to meet some deliberative standards, such as respect for one's interlocutors, which facilitates deliberative exchange (Lord and Tamvaki, 2013; Pedrini, 2014; Gerber et al., 2018).

besides the quality of deliberation, a gendered pattern of speech also includes a structural dimension of public discourse: who takes the floor, when do they do it, and to what effect? Gender-specific features of language thus result from conversational interactions (Hannah and Murachver, 2007, 275). Who follows whom may also matter: what if men are more likely to speak in succession than women?

Research in political science, social psychology, and communication has shown that women are interrupted more frequently by men than men are by women in legislative and non-legislative settings, whilst specifying conditions under which interruptions occur (Mattei, 1998; Mendelberg et al., 2014), and whether they are hostile or not (Kathlene, 1994). A gendered pattern of interruptions produces gendered consequences. Women are less successful than men at taking and holding the floor (Grob et al., 1997, 293), and their influence in the group is undermined (including women's perception of their own efficacy) (Mendelberg et al., 2014, 29). The gendered pattern of interruptions underscores the importance of sustaining speaking opportunities in public debates. However, women do not necessarily have to be interrupted by men in order to be marginalized in mixed-sex debates. A gendered pattern of dominance may be sustained at the level of speaking sequences throughout the debate. A speech by a previous woman participant may embolden another woman to contribute.

Whether a woman speaker is more likely to be followed by a another woman or a man indicates whether women are speaking in succession, thereby creating a deliberative space or an enclave conducive to women asserting their perspectives. Recognizing that women's perspectives are often marginalized in public fora, Sunstein (2007, 277) has argued that "a special advantage of enclave deliberation is that it promotes the development of positions that would otherwise be invisible, silenced, or squelched in general debate." Deliberative enclaves can "protect" (Mansbridge, 1999, 63) the discourse of the disadvantaged and marginalized, ensuring greater equity and quality of deliberation (Karpowitz et al., 2012, 605). For these scholars, enclaves refer to separate marginalized groups given an opportunity to deliberate together. Alternatively, the structure of speaking turns can also

restrict this deliberative space. It can pave the way for dominance in debates if speaking sequences are gendered, and if men speak in longer sequences than women in mixed-sex debates. If this is the case, such a pattern across the whole debate can cumulatively limit women's substantive contributions and their influence on the outcome of the debate.

3.3.3 De-centering

Both scholars of political representation and transitional justice have found evidence that themes of contributions by men and women differ. Therefore, topics women address in public debates can also be an indicator of their influence, or of the lack of it. Women and men talk about different issues, and these thematically gendered patterns persist in a range of settings: formal and informal, public and private, and virtual and face-to-face communication, e.g. in national parliaments, on a campaign trail, or on social media (Carroll, 2008; Krook, 2010; Dabelko and Herrnson, 1997).

The gendered pattern of political speech diversifies policy and legislative agendas (Greene and O'Brien, 2016). At the same time, women speaking about women's issues enhances women's representation and influence (Pearson and Dancey, 2011a; Herrnson et al., 2003; Bratton and Haynie, 1999). This includes legislation on gender-based violence, as illustrated by the toughening of sentences in the Egyptian parliament for performing female genital mutilation (Abdelgawad and Hassan, 2019). However, in post-conflict settings, during proceedings in truth commissions and war crimes trials, women de-center, i.e. they are reluctant to talk about their own experience of conflict-related violence. Instead, when they talk about violence, women center their narrative on others: their husbands, partners, and children (Kashyap, 2009; Yarwood, 2013; Crosby and Lykesy, 2011; Theidon, 2013). This stands in contrast to women's public advocacy on women's issues, including SGBV. Studies of women's advocacy reveal the efficacy of frames, opportunity structures, and network dynamics (Berry, 2018), but this is of limited analytic utility for understanding how women's influence is limited in mixed-sex, face-to-face public debates.

Investigating the topics women and men address can also indicate to what extent peace-making outcomes, such as mandates of transitional justice instruments, are responsive to women's concerns. Considering that conflicts impact women differently than men, whether women talk about violence they suffered in debates addressing the criminal legacy, or they de-center, captures an important aspect of a likely broader pattern of thematic differences between women and men. This gendered thematic pattern can lead to a gendered outcome.

In sum, the biggest challenge for researchers of gender and language is to establish "why and where differences exist" (Hannah and Murachver, 2007, 275). These differences matter; differentiated speech reflects differences in power, status, and authority, which in turn determine speakers' influence on policy. Despite the growing scholarship focusing on women's speaking behavior, gender-based thematic and structural differences in speech patterns have not been studied together with the deliberative quality of women's contributions. The mechanisms we propose and test to account for why women's representation in a transitional justice process fails to produce a gender-responsive transitional justice instrument incorporates novel measures of the gendered nature of discourse: the sequential structure of turn-taking and the deliberative quality of speeches, alongside the thematic content of their speeches.

3.4 Research Design

To study women's representation without influence and to test the proposed mechanisms to explain gendered outcomes, we focus on the RECOM process in the post-conflict Balkans.

3.4.1 The Background

The RECOM grassroots civil society initiative advocates the creation of a regional fact-finding commission that would compile a list of all victims of the wars surrounding the dis-

solution of the former Yugoslavia [8]. With their narrow focus on perpetrators, externally-led efforts to promote justice in the Balkans through the operation of the ICTY from 1993 to 2017 had left a shared sense of elusive justice among victims on all sides of a series of conflicts through the 1990s and into the early 2000s in the region. The international court's narrow focus on perpetrators and not victims is inherent in the pursuit of retributive justice.[9] As a multi-ethnic, victim-centered transitional justice process, RECOM has provided a local restorative approach to the legacies of the Balkan wars. It was embodied by consultations with a wide range of stakeholders in 134 one- or two-day-long debates from 2006 to 2011. Like other human rights initiatives in the poor, post-conflict region, RECOM's activities were supported by foreign donations. Nonetheless, the agenda-setting for RECOM's meetings remained in the hands of local actors, who launched and drove this bottom-up transitional justice process (Rangelov and Teitel, 2014)[10].

3.4.2 Case Selection

Case selection refers to the event that the theory tries to explain as well as to selection of countries, both of which require attention (Gerring and Cojocaru, 2016, 408). With its lack of provisions that reflect women's concerns and ensure women's equal involvement in the commission's work, RECOM's draft Statute is a typical case of peace-making with a gendered outcome[11]. Further, as a transitional justice process in the Balkans, RECOM is a response to criminal legacy typical of civil wars fought along identity lines where sexual violence is a part of the overall repertoire of violence (Wood, 2014, 461)[12]. Nested in a larger body of the extant literature, a typical case contributes to theory development by producing arguments that can explain some, but not all, cases (Toshkov, 2016, 292). The value of case studies is in strictly "contingent generalizations that apply to the subclass of

[8] More details about the historical development of the initiative can be found in the Appendix.
[9] For a comprehensive assessment of the ICTY, see Orentlicher (2018).
[10] See Proces REKOM (2011).
[11] For example, see original mandates of the South African and Peruvian truth commissions (Borraine, 2000; Bueno-Hansen, 2015).
[12] For an overview of violence during the Bosnian war see Berry (2018, 116-129).

cases" similar to those that are studied (George and Bennett, 2005, 32). The patterns of discourse at the intersection of identity, gender, and wartime victimization can shed light on gendered peace-making outcomes after other intra-state conflicts (Allansson et al., 2017), and, specifically those fought along the ethnic identity axis. Because religion does not frame peace-building efforts in the Balkans[13], the findings are of limited applicability to understanding women's influence in the aftermath of intra-state conflicts in contexts where religious norms (which are not limited to a single religious group) shape women's participation in peace-making [14].

3.4.3 Data

The text data analyzed in this study consists of transcripts of 20 debates about the commission's draft Statute organized by the RECOM[15], comprising over 500,000 words. The transcripts of these debates are publicly available on the RECOM's website[16]. In this text corpus, we code the gender and role (discussant or moderator) of each speaker. The order of speaking was determined by moderators, who responded to participants' requests to take the floor. Participants themselves were drawn from broad sections of civil society in the region. Because of the consultative nature of the RECOM's process, the organizers' priority was to make debates diverse and inclusive along different identity axes (Bonora, 2019, 145): men and women, people from all ethnic groups involved in Balkan conflicts, from different constituencies (victims, veterans, human rights activists, and professionals, such as lawyers, journalists and teachers), and different generations. Their aim was to ensure a wide representation of different experiences of conflict and views on their appropriate redress by the regional commission, which would be codified

[13]Balkan conflicts cannot be classed as religious regardless of politicization of religious identities, see Harris and Baumann (2019).

[14]Religion can both restrict women's influence, for example when used to justify women's exclusion from political processes, and facilitate articulation of their concerns, as illustrated by religious-based activism for women's rights used in Libya's Noor Campaign, see UN Women (2015).

[15]We check for the possibility that the conditions under which the draft Statute was adopted differ from the conditions under which debates were held.

[16]See https://www.recom.link/sr/

in the draft Statute. Holding debates both in rural and urban locations in all countries of former Yugoslavia was an additional strategy to ensure the diversity of views.

3.4.4 Preprocessing

Preprocessing the text data involved a number of steps. We tagged each speaker's speaking turn (i.e. speeches), and within each speaking turn we code each *speech acts* i.e. *arguments* (*demands* in the DQI terms) about any given issue under discussion, e.g. location of the Commission's seat or selection of commissioners. This enables us to capture the gendered nature of discourse by analyzing not only who spoke and for how long, but also how many arguments they made (about the articles of the draft Statute) and how well substantiated those arguments were. Lastly, as the original sessions were held in multiple languages (Albanian, Bosnian, Croatian, Macedonian, Montenegrin, Serbian, Slovenian), the corpus was manually translated into Serbian[17]. We then applied a set of natural language processing tools developed for Balkan languages.

3.4.5 Methods

We combine quantitative content analysis, which relies on interpretative coding of text segments, to measure the quality of deliberation with a Discourse Quality Index (DQI), as well as computer-assisted quantitative text analysis to quantify the word frequencies in utterances at the speaker level. The combination of these two methods allows us to conduct a granular analysis of the content of speakers' contributions and of the frequency of participants' speeches and their speaking sequences, which could not be achieved by conducting only a computer-assisted text analysis of gendered speech patterns[18]. We fit Bayesian multi-level and structural topic models to estimate the effects of gender on speech behavior: the quality of deliberation, turn-taking, and thematic content. In what

[17]Details on the linguistic aspects of preprocessing that was carried out by one of the authors with some research assistance are provided in the Appendix.

[18]For example see Parthasarathy et al. (2019).

follows, we first re-assess women's representation in a transitional justice process, with measures that have previously not been used by scholars of peace-making.

3.5 Verifying the Puzzle of Women's Representation and Influence in Peace-Making

We have argued that a more robust measure of women's representation in peace-making is needed before we can claim that their representation results in gendered outcomes that marginalize their interests and concerns. Consequently, we distinguish three levels of women's representation: (1) physical presence, (2) representation in turn-taking, and (3) participation in argumentation. The first level captures the turnout rate of women and men in the debates about the Statute; the second level reflects the proportion of participants who actually spoke during the debate as opposed to those who remained silent; and the third level accounts for those speakers who made an argument about the provision(s) of the draft Statute when they spoke.

Table 3.1: **Average participation by men and women at different levels.**

	Men	**Women**
Presence $\frac{discussants}{participants}$	50.5%	38.1%
Speech $\frac{speech+arguments}{discussants}$	55.4%	37.4%
Argumentation $\frac{arguments}{speech+arguments}$	81.1%	74.2%

Note: The percentages for presence do not add up to 100% as moderators are excluded.

Table 3.1 presents average participation levels across 20 debates at different levels for discussants of both sexes. The number of participants varies between 18 and 70[19]. Of all the participants, 38% were female discussants, of whom 37% made at least one utterance (i.e. made a speech), and of those who spoke, 74% made a statement pertaining to at least one of the articles of the draft Statute. While it is impossible to entirely rule out underrepresentation of women driving a gendered outcome, other indicators of women's

[19]Further summary statistics on participants are available in the Appendix.

representation in speech, which have been overlooked in the existing literature, need to be considered. We proceed to measure the length of participation: instead of time it took for each individual to deliver a speech (Karpowitz et al., 2012), we use the number of words that each speech contained. To assess the association between a speaker's sex and speech length, measured by the logarithm of the number of word units (tokens), we use a random-intercept Bayesian multilevel model with consultations as level-2 units. Given the number of consultations included in our analysis, Bayesian estimation allows us to avoid potential problems with biased coefficient estimates and confidence interval coverage that can be encountered when the number of groups is small (Stegmueller, 2013). Table 3.2 shows the estimated coefficients for a speaker's sex and consultation-level predictors (ethnic diversity, level of debates, participating community, and whether translation was required, e.g. because of the ethnic composition of a debate). In addition to categorical descriptors of the consultations, we include the proportion of women present as another independent variable. Although the coefficient for female speakers is slightly negative ($\hat{\beta}_{female} = -0.112$), meaning that women's utterances, on average, tend to be shorter in terms of the number of words spoken, this relationship is not statistically significant.

Table 3.2: **Multi-level Linear Models of Speech Participation.**

	Log(words)	
	(1)	(2)
Sex (ref: Male)		
Female	-0.116 (-0.252, 0.031)	-0.112 (-0.25, 0.028)
Consultation-level covariates		✓
Groups	20	20
Observations	1472	1472

Note: 95% HPD intervals are shown in parentheses. Complete output is available in the Appendix.

Having checked women's representation at several levels, in addition to their physical presence, our analysis provides comprehensive assessment of their participation in the RECOM's debates; women made a vocal contribution both in terms of taking speaking

turns and presenting arguments. Although it is somewhat lower than men's, women's (under)representation does not provide a convincing explanation for the absence of provisions reflecting women's needs and interests in the draft Statute of the fact-finding commission. We therefore examine empirically three mechanisms: deliberation, emboldening, and de-centering to assess whether there are any gendered patterns in discourse that could lead to the gendered outcome.

3.6 Measuring the Quality of Deliberation

The first potential mechanism to account for a gendered outcome is the quality of deliberation. To test this empirically, one of the authors adapted the measurement instrument and constructed the Discourse Quality Index (DQI) for Transitional Justice, a variant of technique for analyzing the quality of deliberation (Steenbergen et al., 2003; Steiner, 2012). The DQI is a set of analytical constructs that jointly measure the quality of deliberation. Its construction responded to a need to supplement theorizing about deliberation with "empirical investigations of real-life deliberations" (Steiner et al., 2005, 43), in particular with application to parliamentary debates. The dimensions of the DQI are underpinned by Habermas' notion of "communicative action" (Habermas, 1984), which stipulates: "individuals give and criticize reasons for holding or rejecting particular validity claims, so that universally valid norms can be discovered through reason" (Steenbergen et al., 2003, 25). The Discourse Quality Index for Transitional Justice contains eight components: the first code captures the presence of (1) *interruptions*[20]. Habermas' *level of justification of demands* is denoted by (2) *justification rationality*, and *content of justification* is considered individually with reference to the (3) *common good of a community*, (4) *specific subgroup*, such as victims or young generations, or (5) *abstract principles*, such as peace, while *respect* is subdivided into two types: (6) *respect towards participants and their ar-*

[20]In the original work (Steenbergen et al., 2003) this component is labeled as *participation*, but here we use *interruption* to avoid confusion with a general term participation referring to a speaker's overall contribution to debates.

guments, and (7) *respect towards groups*, and, lastly, (8) *story-telling* captures whether participants use stories alongside rational arguments.

The principal unit of analysis in the DQI coding strategy is a *speech act*, defined as "the public discourse by a particular individual delivered at a particular point in a debate" (Steenbergen et al., 2003, 27). The relevance of a speaker's utterance for coding is determined by whether it contains a *demand*, i.e. "a proposal on what decision should or should not be made"(Steenbergen et al., 2003, 27). For example, a position on whether the commission's seat should be in Sarajevo illustrates a demand. The RECOM's text corpus consists of 1,211 speech acts uttered by discussants over 20 debates (excluding speech acts by moderators in line with the practice followed in the analysis of parliamentary debates, as well as in experimental studies). All speech acts were identified and manually coded according to the DQI for Transitional Justice codebook developed by one of the authors. Each of the 1,211 speech acts was coded twice, independently, along all dimensions of deliberation, by one of the authors and a trained coder[21].

Aggregation of multiple components of the DQI has received only limited attention in the literature. It is not uncommon to calculate a simple additive index by summing up the ordinal codes assigned to each speech act (Hangartner et al., 2007) or to consider each of them separately (Steiner et al., 2005). Other researchers have applied principal component analysis to combine the items of DQI (Caluwaerts and Deschouwer, 2014). Here we largely follow the approach of Gerber et al. (2018) and estimate a two-parameter Bayesian item-response theory (IRT) model to calculate an aggregate measure of the quality of deliberation. This modeling strategy is appealing for several reasons. Substantively, it assumes that the quality of deliberation is a unidimensional latent construct that is manifested through multiple items (DQI components), each of which has a difficulty and discrimination parameter. In the context of deliberation, the former can be viewed as how big of a political, social, or psychological challenge each of the items presents to a discussant. For example, while it could be a relatively easy task not to interrupt other

[21]The inter-coder reliability statistics are available in the Appendix. All of them indicate an acceptable level of reliability.

participants, in the context of a transitional justice process, substantiating one's arguments with references to the shared abstract principles can be far more challenging. At the same time, some, perhaps more achievable objectives, such as delivering an argument that overcomes an ethnic interest, can differentiate better between a lower and higher quality speech act. This idea is captured by the second parameter of the model: discrimination. Apart from theoretical appeal, IRT offers a number of statistical advantages. First, it allows one to model the observed variables, derived from hand-coded speech acts, as categorical variables[22], without making assumptions that they are measured on an interval scale and possess an additivity property that allows them to be meaningfully summed up. Second, other types of modeling strategies such as factor analysis would require another implausible assumption of normally distributed error terms[23].

[22]In our analysis we simplify the codes with more than two categories by dichotomizing them into binary variables, see the Appendix.

[23]While the ensuing analysis is focused on this aggregation approach, we also demonstrate its close correspondence to other strategies in the Appendix.

Figure 3.1: **Response functions of each DQI component.** Side panels show the examples of speech acts with high and low quality of deliberation.

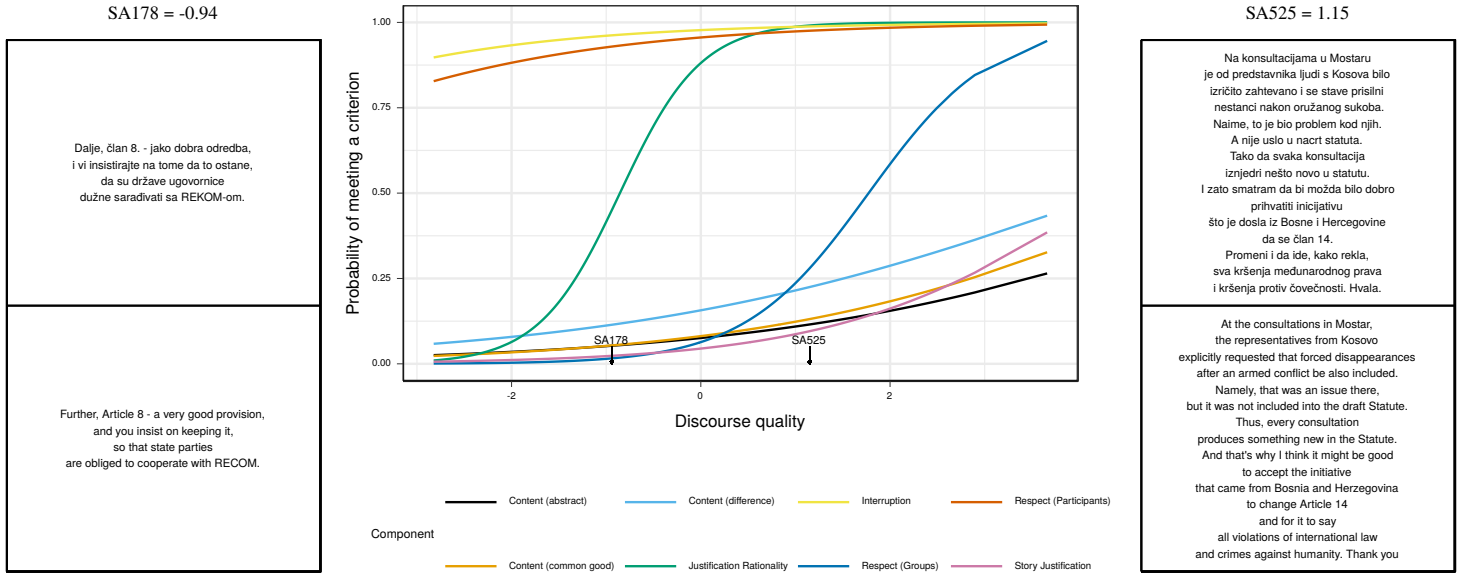


Figure 3.1 shows the response functions of each of the DQI components. Avoiding interruptions and showing respect towards other participants appear to be the easiest deliberation criteria to satisfy. These results are also a consequence of very few identified interruptions[24] and instances of open disrespect that we observe in the data. This is a noteworthy observation, given how demanding it is for people from different sides of an armed conflict to discuss the criminal legacy together. All components related to the content of justification as well as story-telling are the hardest principles of deliberation to meet in practice. The steep curves for justification rationality and respect towards other groups show that these two components can best discriminate between those speech acts the quality of which falls just below or just above their respective difficulty.

To estimate the association between gender and the quality of deliberation we fit multilevel model that includes both demand-level (model 2) and consultation-level (model 3) explanatory variables. Speech acts delivered by women tend to have slightly higher quality of deliberation ($\hat{\beta}_{female} = 0.031$), although this relationship is not significant. The empirical assessment of the quality of deliberation in a civil society context does not support arguments that deliberation as a mode of communication disadvantages women. As such, it is consistent with the findings from parliamentary debates (Bächtiger and Hangartner, 2010). In the case of a transitional justice process, these results suggest that the quality of arguments presented by women and men does not account for the content of the adopted draft Statute that does not respond to women's needs, interests and concerns.

3.7 Gendered Structure of Debates: Emboldening

The focus of the previous mechanism was speech, considered in isolation. Speeches, however, do not occur in isolation; they typically constitute part of a larger *in situ* or *ex situ* conversation. It is plausible that gender becomes an important determinant not of how well substantiated one's arguments are, but of whether the arguments are voiced in

[24]Only about 4% of speech acts contain some form of interruption.

Table 3.3: **Multi-level Linear Models of the Quality of Deliberation.**

	DQI		
	(1)	(2)	(3)
Sex (ref: Male)			
Female	0.026 (-0.054, 0.107)	0.031 (-0.052, 0.111)	0.031 (-0.047, 0.112)
Repeated Speaker (ref: No)			
Yes		-0.092 (-0.185, -0.002)	-0.087 (-0.182, 0.006)
Issue Polarization (ref: Low)			
Medium		0.126 (0.042, 0.208)	0.127 (0.047, 0.21)
High		0.341 (0.226, 0.457)	0.327 (0.214, 0.445)
Diversity (ref: Mono-ethnic)			
Dyadic			0.204 (-0.185, 0.579)
Multi-Ethnic			-0.016 (-0.385, 0.357)
Level (ref: Non-regional)			
Regional			0.085 (-0.318, 0.481)
Type (ref: General)			
Professionals			0.054 (-0.223, 0.331)
Victims			0.249 (-0.008, 0.501)
Translation (ref: No)			
Translation			-0.025 (-0.318, 0.279)
Intercept	0.025 (-0.072, 0.122)	-0.008 (-0.121, 0.108)	-0.147 (-0.471, 0.179)
Groups	20	20	20
Observations	1211	1211	1211

Note: 95% HPD intervals are shown in parentheses.

the first place. As the results of the DQI analysis indicate, there were very few direct interruptions in our corpus. Interruptions are a focus of a considerable body of literature as the result of them being a prominent feature of discourse that is also easy to measure (Mattei, 1998; Mendelberg et al., 2014; Kathlene, 1994). However, a debate can be structured in such a way that participants do not feel emboldened to speak in the first place. While a thorough analysis of this phenomenon would require looking at the underlying psychological processes, we are still able to study some observable implications of this mechanism from the transcripts of debates. Specifically, we look at the sequence in which men and women deliver a speech.

In the absence of any gendered dynamics, we would expect to find no differences in the number of speeches delivered in sequence by men and women. To test this mechanism we fit a Poisson multi-level model with the number of speeches in a row delivered by male and female discussants as the dependent variable. We adopt a hierarchical approach here in particular to control for the percentage of female discussants, which varies at the consultation-level[25]. While our approach is somewhat different from direct modeling of transition probabilities when treating speech sequences as Markov chains (Eggers and Spirling, 2014), we adopt similar underlying assumptions and exclude moderators from analysis[26]. As Table 3.4 demonstrates, contrary to our expectations, the gender of the speaker has a significant association with the number of speeches in sequence. The results show that sequences of speeches delivered by women are on average 40% shorter ($\hat{\beta}_{female} = -0.48$) than sequences delivered by men, while controlling for the percentage of female discussants and other consultation-level characteristics (model 2). Overall, the average length of women's sequences is 1.98 speeches, while that of men's is 3.22. This finding indicates that gendered dynamics of debates need not manifest itself in interruptions or other conspicuous demonstrations of power asymmetry. Rather, they can result

[25]The number of groups is 19 in this case rather than 20, as one consultation contained only men, which prevents us from modeling speech sequences there.

[26]As disentangling the mechanical effects of slight imbalance in participation ratios between genders from the genuine differences in the structure of debates is not straightforward, we provide further robustness checks and alternative modeling strategies in the Appendix.

Table 3.4: **Multi-level Poisson Models of the Number of Speeches Made by the Discussants of the Same Sex in Sequence.**

	Number of speeches in sequence	
	(1)	(2)
Gender (ref: Male)		
Female	-0.488 (-0.596, -0.384)	-0.488 (-0.592, -0.384)
% Female Discussants		0.005 (-0.022, 0.033)
Diversity (ref: Mono-ethnic)		
Dyadic		-0.105 (-1.11, 0.859)
Multi-Ethnic		0.605 (-0.378, 1.547)
Level (ref: Non-regional)		
Regional		-0.569 (-1.529, 0.359)
Type (ref: General)		
Professionals		0.022 (-0.64, 0.701)
Victims		0.094 (-0.759, 0.953)
Translation (ref: No)		
Yes		0.231 (-0.5, 0.901)
Intercept	1.269 (1.064, 1.471)	0.838 (-0.361, 2)
	0.165	0.214
log-posterior	-1309.095	-1314.809
Groups	19	19
Observations	548	548

Note: 95% HPD intervals are shown in parentheses.

from a subtler pattern of speech dominance that is sustained at the level of the sequence of speaking turns. The pattern of one woman speaking and emboldening another woman to speak, which is what we would expect if women were able to deliberate in enclaves, does not occur. This finding has substantive implications, because women's lines of argumentation cannot be given expression and be sustained in a mixed-sex setting. It can result in the absence of gender-responsive provisions in the draft Statute. In contrast,

when women do speak in succession, their experience of conflict, for example as bereaved mothers of killed recruits, is articulated as a demand for the recognition of this loss in the commission's definition of human losses[27].

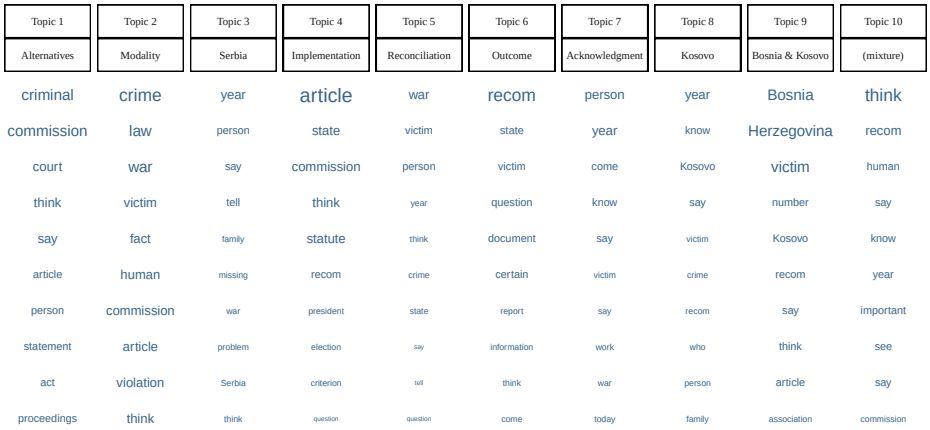
3.8 Thematic Differences: De-centering

After considering the quality and sequences of female and male contributions, we shift our focus to the thematic content of the speeches by men and women. Like other scholars (Terman, 2017), we approach the analysis guided by theoretical expectations. In this study, they are drawn from scholarly debates in the fields of political representation and transitional justice. The text-as-data approach (Grimmer and Stewart, 2013) offers an innovative way of studying transitional justice debates. Like Parthasarathy et al. (2019), we show that this approach is suitable for studying deliberation in civil society meetings. We use the text-as-data approach to augment the qualitative reading, manual coding, and statistical modeling of the structure of debates by applying structural topic models to our text corpus to estimate the differences in proportions of speaking time that women and men dedicate to different topics in their speeches. In order to prepare the corpus for analysis, we used a newly developed set of natural language processing tools developed as part of the Regional Language Development Initiative (ReLDI) for several Balkan languages, including the Serbian language that was used to standardize the multi-language text corpus (Ljubešić et al., 2016), removed stopwords and lemmatized the texts[28].

Structural topic models (Roberts et al., 2014) are an extension of classical topic models based on Latent Dirichlet Allocation, proposed by Blei et al. (2003). Apart from the estimation of topic proportions for each individual document, they allow one to incorporate meta information and estimate how additional covariates affect topic prevalence. Rather than fitting a model on individual utterances, we aggregate them at the speaker

[27]See Konsultacije sa udruženjima žrtava, Beograd, Srbija, 3 July 2010.
[28]More details on preprocessing are available in the Appendix.

Figure 3.2: **Top 10 Words by Topic.**



Note: English translation from Serbian original was made after fitting STM model. Size of terms is proportional to the probability of being generated by a given topic.

level. Through an iterative procedure we find that a 10-topic structural topic model produces the best balance between statistical fit and substantive interpretation[29]. As with other analyses presented above, we include only the speeches delivered by discussants and not those delivered by moderators.

Figure 3.2 displays the most prevalent terms in each topic. We leave topic 10 unlabeled as it largely represents a combination of the other nine topics. Figure 3.3 shows the estimated effect of gender on the proportion of different topics and indicates a gendered thematic pattern of discourse; women and men speak about different issues demonstrating different ways in which they approach the Statute deliberations. The topic "evaluation" indicates that women scrutinize the proposed articles of the draft Statute from the perspective of criminal justice, arguing that the Statute should maintain a distinction between the regional fact-finding commission (which is as a restorative transitional justice mechanism) and criminal justice, whilst seeking clarification of the relationship between the commission and local courts. They are concerned, for example, that the

[29]The Appendix contains additional information on topic diagnostics and alternative specifications. In addition to gender, we include country, level (regional/non-regional), and type of participating community (general/professional/victims) as covariates.

draft Statute might give the commission quasi-legal powers akin to those that a national *court* exercises in *criminal* proceedings[30], such as providing for a *criminal* sanction for non-appearance of individuals summoned to testify before the *commission* (instead of voluntary testimonies). Likewise, women are preoccupied with legal consequence for individuals alleged to have committed war crimes in the proceedings before the *commission*, and so on. Women are also associated with the topic dedicated to Kosovo[31]. Lastly, we find in the RECOM's case that women de-center, directing their contributions away from their own experience of violence. For example, the needs of SGBV victims are elided by a generic reference to all victims, while harm suffered by men in detention camps is singled out in the topic 'acknowledgement'[32]. By contrast, men focus more on practical issues involved in the operationalization of the fact-finding commission. This is captured by the topics: 'implementation' focused on the *articles* of the draft *Statute* and the *election* of commissioners including *criteria* for their *election*, 'reconciliation' as the rationale of the RECOM process focused on recognition of *victims* of *crimes* as *humans*, and 'outcome' reflected in their preoccupation with the *report* on all *victims* to be produced by *states* participating in the *RECOM* commission.

Demonstrating a lack of arguments rooted in consideration of gender-specific experiences of conflict, this exploratory analysis of themes addressed by female and male speakers reveals a broad pattern of thematic differences, which also includes de-centering in women's speeches as they do not discuss how their own experience of violence should be addressed. Both women and men contribute to shaping important aspects of the draft Statute, although women approach the task from a holistic perspective on how restorative justice should operate in post-conflict societies alongside retributive justice, whereas men focus on 'nuts and bolts' of the commission's operationalization[33]. This gendered thematic

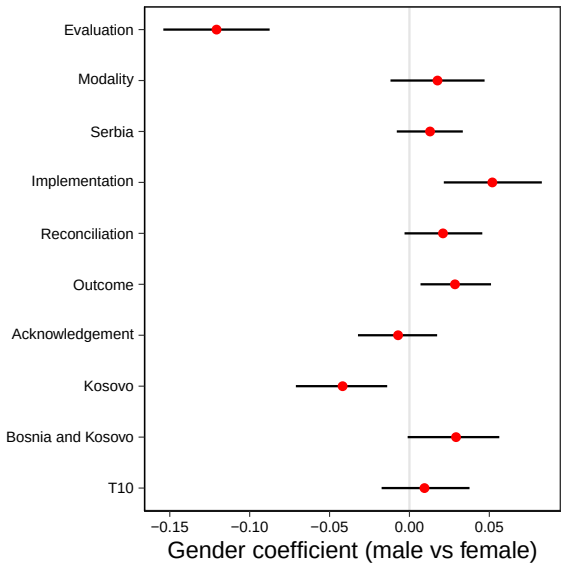
[30]Italics indicate words in topics. Further information on topic interpretation is in the Appendix, including speeches with highest scores on each topic.

[31]This, however, could be an artifact of the few women present at some of the meetings that happened in Kosovo.

[32]See the harmonic mean words frequency and exclusivity (FREX) which also aided our interpretation in the Appendix.

[33]We are grateful to an anonymous reviewer for suggesting this emphasis

Figure 3.3: **Topical Prevalence by Male and Female Discussants.**



Note: Estimates from structural topic model (STM) with 10 topics are shown.

pattern of speeches can be related to the lack of provisions for women's equal participation in the commission's work and the absence of procedures on how to address SGBV in the commission's operation illustrating the gendered character of the draft Statute that is not responsive to women's needs and interests.

3.9 Conclusions

Peace-making outcomes that are not responsive to women's concerns are at the heart of the reproduction of gender injustice and the persistent elusiveness of "quality peace" (Wallensteen, 2015) for women. Inequalities persist even after the Women Peace and Security Agenda spurred women's peace activism, women's demands for access to peace-making, and women's articulation of their particular needs (Shepherd, 2017). In the area of post-conflict transitional justice, women's advocacy has led to the codification of accountability for wartime sexual and gender-based violence in international law. These developments have in turn had an impact on public perceptions and policy responses to

this issue in post-conflict zones (Warren et al., 2017). The burning question now is: why, despite such "norm augmentation" in the post-Cold War period (Ní Aoláin, 2014, 625), do we still see "old specters of unseen hierarchies operating to the detriment of addressing harms experienced by women"?

This study of women's speaking behavior during peace-making departs from the well-trodden research agenda focused on women's representation and gendered outcomes in peace and justice processes. It provides a new perspective on gendered dynamics of peace-making by identifying a novel axis of male domination at the micro-level of public discourse. We know that fewer women than men are likely to be at the negotiating table, despite recent progress in narrowing the gap between men's and women's attendance. Nonetheless, an important part of the puzzle of women's limited substantive representation has been overlooked, given our still weak understanding of what happens when women engage in the exchange of arguments with men on the other side of the table.

Feminist scholars have noted that the "add women and stir" solution has done little to advance gender-just peace (Ní Aoláin, 2016). Going beyond the issue of representation, our research reveals that the patterns of men's domination during a public discussion are subtle but nonetheless consequential – even when they are not expressed in obvious forms such as interruptions. Women's relative underrepresentation in peace-making continues to be an issue that ought to be addressed. However, the focus merely on (numerical) underrepresentation neglects the question how women's voices matter in peace-making. As we have shown in this study of post-conflict justice debates, not only do women take the floor almost as often as men, but there is no substantial difference between men and women in terms of the quality of deliberation. While the content analysis reveals that men and women address different topics, we propose that the gendered structure of turn-taking is key to women's limited influence on peace outcomes. Women's speech is restricted during debates; women speak in shorter sequences than men, which shrinks women's deliberative space to develop their argumentation, including on those issues and topics that would better reflect a whole range of women's concerns and demands for

equality, even if they wish to remain silent about sexual and gender-based violence[34].

This lacuna in the scholarship that concerns a discursive dimension of a peace-making process is also linked to the issue of data (Anderlini, 2007) and methods used to study how peace-making is gendered. Efforts have centered on counting 'bodies at the table' or provisions of peace-agreements that refer to women, which have become standard indicators of women's inequality and gendered peace that disadvantages women (Paffenholz et al., 2016)[35]. Our empirical analysis of gendered speech patterns in a transitional justice process contributes to efforts to quantify gendered dynamics during peace-making and sheds new light on constraints on women's influence.

While this research furthers the study of how conflict-resolution and peace-making are gendered (David et al., 2018), it also provides new insights for scholars of political representation and communication interested in the study of male domination in political communication. Our investigation of women's speaking behavior in civil society debates confirms the value of increasing the number of comparisons by expanding the "sites" of political representation (Krook, 2010; Parthasarathy et al., 2019), which are usually restricted to institutional settings such as parliaments. We elucidate women's participation in a parallel non-state civil society process that also marginalizes women.

Exposing men's dominance at the level of turn-taking, this study provides another possible solution toward greater substantial gender equality for women in peace-making and in politics more generally, beyond the imperative of equal representation and access that have preoccupied scholars and practitioners. Observed at the micro-level of discourse, our findings point to the need for greater attention to the management of speaking turns during public debates. Extant research has pointed to the benefits of recognizing the marginalized and their views through enclave deliberation conceptualized and implemented as deliberation in separate group(s) made up of those who are disadvantaged and whose perspectives are sidelined in mainstream debates (Mansbridge, 1999;

[34]On silence as a site of power and agency, see Selimovic (2020).

[35]Notwithstanding this, compiling new datasets (Bell and Badanjak, 2019) and refining definitions of inclusion (Arthur, 2016) will provide more robust explanations in this vein.

Sunstein, 2007; Karpowitz et al., 2012). This research provides a novel insight that the benefits of enclave deliberation can also potentially be gained from turn-taking sequences nested in the debates in mixed-sex groups and fora. Ultimately, we recognize that the pattern we found might be only one aspect of a number of such micro-level instances of lopsidedness that can lead to gendered outcomes. Whether this is so remains to be established by future research aided by the inter-disciplinary study of the micro-level of discourse, which stands to reveal new insights into the nature and consequences of communicative interactions between women and men during post-conflict peace-making.

Bibliography

Abdelgawad, H. and M. Hassan (2019). Women in the Egyptian parliament: a different agenda? *Review of Economics and Political Science ahead-of-print*.

Acker, J. (1990). Hierarchies, jobs, bodies: A theory of gendered organizations' gender and society. *Gender & Society 4*(2), 139–158.

Adjei, M. (2019). Women's participation in peace processes: a review of literature. *Journal of Peace Education 16*(2), 133–154.

Allansson, M., E. Melander, and L. Themnér (2017). Organized violence, 1989–2016. *Journal of Peace Research 54*(4), 574–587.

Anderlini, S. N. (2007). *Women Building Peace: What They Do, Why It Matters*. Boulder, CO: Lynne Rienner.

Arthur, P. (2016). Notes from the Field: Global Indicators for Transitional Justice and Challenges in Measurement for Policy Actors. *Transitional Justice Review 1*(4), 9.

Bächtiger, A. and D. Hangartner (2010). When Deliberative Theory Meets Empirical Political Science: Theoretical and Methodological Challenges in Political Deliberation. *Political Studies 58*(4), 609–629.

Bächtiger, A. and J. Steiner (2005). Introduction. *Acta Politica 40*(2), 153–168.

Bell, C. and S. Badanjak (2019). Introducing PA-X: A new peace agreement database and dataset. *Journal of Peace Research 56*(3), 452–466.

Bell, C. and C. O'Rourke (2010). Peace Agreements or Pieces of Paper?: The Impact of UNSC Resolution 1325 on Peace Processes and their Agreements. *International and Comparative Law Quarterly 59*(4), 941–980.

Berry, M. E. (2018). *War, women, and power: From violence to mobilization in Rwanda and Bosnia-Herzegovina*. New York: Cambridge University Press.

Björkdahl, A. and J. Mannergren (2013). Advancing women agency in transitional justice.

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

Blumenau, J. (2019). Legislative Role Models: Female Ministers, Participation, and Influence in the UK House of Commons. *British Journal of Political Science* (Forthcoming).

Bonora, C. (2019). *New Critical Spaces in Transitional Justice: Gender, Art, and Memory*, Chapter The Question of Gender Inclusiveness of Bottom-up Strategies in Bosnia and Herzegovina. Bloomington, Indiana: Indiana University Press.

Borer, T. A. (2009). Gendered War and Gendered Peace: Truth Commissions and Post-conflict Gender Violence: Lessons from South Africa. *Violence Against Women 15*(10), 1169–1193.

Borraine, A. (2000). *A Country Unmasked: Inside South Africa's Truth and Reconciliation Commission*. Oxford: Oxford University Press.

Bratton, K. A. and K. L. Haynie (1999). Agenda setting and legislative success in state legislatures: The effects of gender and race. *The Journal of Politics 61*(3), 658–679.

Brown, K. and F. N. Aoláin (2015). Through the Looking Glass: Transitional Justice Futures through the Lens of Nationalism, Feminism and Transformative Change. *International Journal of Transitional Justice 9*(1), 127–149.

Bueno-Hansen, P. (2015). *Feminist and Human Rights Struggles in Peru*. Chicago: University of Illinois Press.

Burnet, J. E. (2008). Gender Balance and the Meanings of Women in Governance in Post-Genocide Rwanda. *African Affairs 107*(428), 361–386.

Caluwaerts, D. and K. Deschouwer (2014). Building bridges across political divides: Experiments on deliberative democracy in deeply divided Belgium. *European Political Science Review 6*(3), 427–450.

Cameron, D. (1998). *The Feminist Critique of Language: A Reader*, Chapter Introduction: Why is Language a Feminist Issue?, pp. 1–21. London and New York: Routledge.

Carroll, S. J. (2008). Committee Assignments: Discrimination or Choice? In B. Reingold (Ed.), *Legislating Women*, pp. 135–156. Boulder, CO: Lynne Rienner.

Carver, T. (1996). *Gender is not a Synonym for Women*. London: Lynne Rienner.

Castillo Diaz, P. and S. Tordjman (2012). Women's Participation in Peace Negotiations: Connections between Presence and Influence. Technical report, United Nations Entity for Gender Equality and Empowerment of Women (UN Women).

Chinkin, C. and M. Kaldor (2013). Gender and New Wars. *Journal of International Affairs 67*(1), 167–187.

Cowley, P. (2014). Descriptive Representation and Political Trust: A Quasi-natural Experiment Utilising Ignorance. *Journal of Legislative Studies 20*(4), 573–587.

Crosby, A. and M. B. Lykesy (2011). Mayan Women Survivors Speak: The Gendered Relations of Truth Telling in Postwar Guatemala. *International Journal of Transitional Justice 5*(3), 456–476.

Dabelko, K. L. C. and P. S. Herrnson (1997). Women's and Men's Campaigns for the U.S. House of Representatives. *Political Research Quarterly 50*(1), 121–135.

David, Y., N. Rosler, and I. Maoz (2018). Gender-empathic Constructions, Empathy, and Support for Compromise in Intractable Conflict. *Journal of Conflict Resolution 62*(8), 1727–1752.

Davies, S. E. and J. True (2019). *The Oxford Handbook of Women, Peace, and Security*, Chapter Women, Peace, and Security: A Transformative Agenda?, pp. 4–14. Oxford: Oxford University Press.

Demeritt, J. H., A. D. Nichols, and E. G. Kelly (2014). Female Participation and Civil War Relapse. *Civil Wars 16*(3), 346–368.

Devlin, C. and R. Elgie (2008). The Effect of Increased Women's Representation in Parliament: The Case of Rwanda. *Parliamentary Affairs 61*(2), 237–254.

Dolan, K. (2006). Symbolic Mobilization? The Impact of Candidate Sex in American Elections. *American Politics Research 34*(6), 687–704.

Dryzek, J. S. (2005). Deliberative Democracy in Divided Societies: Alternatives to Agonism and Analgesia. *Political Theory 33*(2), 218–242.

Eggers, A. C. and A. Spirling (2014). Ministerial Responsiveness in Westminster Systems: Institutional Choices and House of Commons Debate, 1832–1915. *American Journal of Political Science 58*(4), 873–887.

Ellerby, K. (2016). A Seat at the Table is not Enough: Understanding Women's Substantive Representation in Peace Processes. *Peacebuilding 4*(2), 136–150.

Gallagher, M. E., D. Prakash, and Z. Li (2020). Engendering justice: women and the prosecution of sexual violence in international criminal courts. *International Feminist Journal of Politics 22*(2), 227–249.

George, A. L. and A. Bennett (2005). *Case Studies and Theory Development in the Social Sciences*. Cambridge, Massachusetts: MIT Press.

Gerber, M., A. Bächtiger, S. Shikano, S. Reber, and S. Rohr (2018). Deliberative Abilities and Influence in a Transnational Deliberative Poll (EuroPolis). *British Journal of Political Science 48*(4), 1093–1118.

Gerring, J. and L. Cojocaru (2016). Selecting cases for intensive analysis: A diversity of goals and methods. *Sociological Methods & Research 45*(3), 392–423.

Gizelis, T.-I. (2009). Gender Empowerment and United Nations Peacebuilding. *Journal of Peace Research 46*(4), 505–523.

Greene, Z. and D. Z. O'Brien (2016). Diverse parties, diverse agendas? female politicians and the parliamentary party's role in platform formation. *European Journal of Political Research 55*(3), 435–453.

Grimmer, J. and B. M. Stewart (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*(3), 267–297.

Grob, L. M., R. A. Meyers, and R. Schuh (1997). Powerful/Powerless Languages in Group Interactions: Sex Differences or Similarities? *Communication Quarterly 45*(3), 282–303.

Grünenfelder, R. and A. Bächtiger (2007). Gendered Deliberation? How Men and Women Deliberate in Legislatures. In *European Consortium for Political Research Joint Sessions*.

Guttmann, A. and D. Thomson (1996). *Democracy and Disagreement: Why Moral Conflict Cannot Be Avoided in Politics and What Should Be Done About It*. Boston: Harvard University Press.

Habermas, J. (1984). *The Theory of Communicative Action. Vol. 1, Reason and the Rationalization of Society*. Cambridge: Polity Press.

Hangartner, D., A. Bächtiger, R. Grünenfelder, and M. R. Steenbergen (2007). Mixing Habermas with Bayes: Methodological and Theoretical Advances in the Study of Deliberation. *Swiss Political Science Review 13*(4), 607–644.

Hannah, A. and T. Murachver (2007). Gender Preferential Responses to Speech. *Journal of Language and Social Psychology 26*(3), 274–290.

Harris, E. and H. Baumann (2019). Identity and war: comparisons and connections between the Balkans and the Middle East. *East European Politics 35*(4), 401–414.

Haynes, D. F., F. Ní Aoláin, and N. Cahn (2011). Gendering Constitutional Design in Post-Conflict Societies. *William and Mary Journal of Women and the Law 17*, 509–545.

Herrnson, P. S., J. C. Lay, and A. K. Stokes (2003). Women Running "as Women": Candidate Gender, Campaign Issues, and Voter-Targeting Strategies. *Journal of Politics 65*(1), 244–255.

Hogg, C. L. (2009). Women's Political Representation in Post-Conflict Rwanda: A Politics of Inclusion or Exclusion? *Journal of International Women's Studies 11*(3), 34–55.

Jennings, K. M. (2019, 01). Conditional Protection? Sex, Gender, and Discourse in UN Peacekeeping. *International Studies Quarterly 63*(1), 30–42.

Karim, S., M. J. Gilligan, R. Blair, and K. Beardsley (2018). International gender balancing reforms in postconflict countries: Lab-in-the-field evidence from the liberian national police. *International Studies Quarterly 62*(3), 618–631.

Karpowitz, C. F., T. Mendelberg, and L. Shaker (2012). Gender Inequality in Deliberative Participation. *American Political Science Review 106*(3), 533–547.

Kashyap, R. (2009). Narrative and truth: a feminist critique of the South African Truth and Reconciliation Commission. *Contemporary Justice Review 12*(4), 449–467.

Kathlene, L. (1994). Power and Influence in State Legislative Policymaking: The Interaction of Gender and Position in Committee Hearing Debates. *American Political Science Review 88*(3), 560–576.

Khodary, Y. M. (2016). Women and Peace-Building in Iraq. *Peace Review 28*(4), 499–507.

King, K. L., J. D. Meernik, and E. G. Kelly (2017). Deborah's voice: The role of women in sexual assault cases at the international criminal tribunal for the former yugoslavia. *Social Science Quarterly 98*(2), 548–565.

Kirby, P. and L. J. Shepherd (2016). Reintroducing women, peace and security. *International Affairs 92*(2), 249–254.

Krause, J., W. Krause, and P. Bränfors (2018). Women's Participation in Peace Negotiations and the Durability of Peace. *International Interactions 44*(6), 985–1016.

Kreft, A.-K. (2019). Responding to sexual violence: Women's mobilization in war. *Journal of Peace Research 56*(2), 220–233.

Krook, M. L. (2010). Studying Political Representation: A Comparative-Gendered Approach. *Perspectives on Politics 8*(1), 233–240.

Krook, M. L. and F. Mackay (2011). *Introduction: Gender, Politics, and Institutions*, pp. 1–20. London: Palgrave Macmillan UK.

Lake, M. (2018). *Strong NGOs and Weak States: Gender Justice and Transnational Advocacy in the Democratic Republic of Congo and South Africa*. Cambridge: Cambridge University Press.

Leaper, C. and M. M. Ayres (2007). A Meta-Analytic Review of Gender Variations in Adult's' Language Use: Talkativeness, Affiliative Speech, and Assertive Speech. *Personality and Social Psychology Review 11*(4), 328–363.

Ljubešić, N., T. Erjavec, D. Fišer, T. Samardžić, M. Miličević, F. Klubička, and F. Petkovski (2016). Easily Accessible Language Technologies for Slovene, Croatian and Serbian. In *Proceedings of the Conference on Language Technologies & Digital Humanities*, pp. 120–124.

Lord, C. and D. Tamvaki (2013). The Politics of Justification? Applying the 'Discourse Quality Index' to the Study of the European Parliament. *European Political Science Review 5*(1), 27–54.

Mackay, F., M. Kenny, and L. Chappell (2010). New institutionalism through a gender lens: Towards a feminist institutionalism? *International Political Science Review 31*(5), 573–588.

Mansbridge, J. (1999). Should Blacks Represent Blacks and Women Represent Women? A Contingent "Yes". *The Journal of Politics 61*(3), 628–657.

Mattei, L. R. W. (1998). Gender and Power in American Legislative Discourse. *The Journal of Politics 60*(2), 440.

McLeod, L. (2019). Investigating "Missing" Women: Gender, Ghosts, and the Bosnian Peace Process. *International Studies Quarterly 63*(3), 668–679.

Melander, E. (2005). Gender Equality, Wealth, and Intrastate Armed Conflict. *International Studies Quarterly 49*(4), 695–714.

Melander, E. (2016). Gender and Civil Wars. In T. David Mason and S. McLaughlin Mitchell (Eds.), *What Do We Know About Civil Wars?*, pp. 197–214. Lanham: Rowman & Littlefield.

Mendelberg, T., C. F. Karpowitz, and J. B. Oliphant (2014). Gender Inequality in Deliberation: Unpacking the Black Box of Interaction. *Perspectives on Politics 12*(1), 18–44.

Ní Aoláin, F. (2014). Gendered Harms and their Interface with International Criminal Law: Norms, Challenges and Domestication Fionnuala Ni Aoláin. *International Feminist Journal of Politics 16*(4), 622–646.

Ní Aoláin, F. (2016). The Relationship of Political Settlement Analysis to Peacebuilding from a Feminist Perspective. *Peacebuilding 4*(2), 151–165.

O'Flynn, I. (2006). *Deliberative Democracy and Divided Societies*. Edinburgh: Edinburgh University Press.

Orentlicher, D. (2018). *Some Kind of Justice: The ICTY's Impact in Bosnia and Serbia*. Oxford: Oxford University Press.

Østby, G., M. Leiby, and R. Nordås (2019). The Legacy of Wartime Violence on Intimate-Partner Abuse: Microlevel Evidence from Peru, 1980–2009. *International Studies Quarterly 63*(1), 1–14.

Paffenholz, T., N. Ross, S. Dixon, A.-L. Schluchter, and J. True (2016). *Making women count-not just counting women: Assessing Women's Inclusion and Influence on Peace Negotiations*. Geneva: Inclusive Peace and Transition Initiative (The Graduate Institute of International and Development Studies) and UN Women.

Pankhurst, D. (2008). *Gendered Peace: Women's Struggles for Post-War Justice and Reconciliation*. New York: Routledge.

Parthasarathy, R., V. Rao, and N. Palaniswamy (2019). Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies. *American Political Science Review 113*(3), 623–640.

Pearson, K. and L. Dancey (2011a). Elevating Women's Voices in Congress. *Political Research Quarterly 64*(4), 910–923.

Pearson, K. and L. Dancey (2011b). Speaking for the Underrepresented in the House of Representatives: Voicing Women's Interests in a Partisan Era. *Politics & Gender 7*(04), 493–519.

Pedrini, S. (2014). Deliberative Capacity in the Political and Civic Sphere. *Swiss Political Science Review 20*(2), 263–286.

Pitkin, H. F. (1967). *The Concept of Representation*. Berkeley, CA: University of California Press.

Porter, E. (2016). Gendered Narratives: Stories and Silences in Transitional Justice. *Human Rights Review 17*(1), 35–50.

Proces REKOM (2011). *Konsultativni process o utvrdjivanju činjenica o ratnim zločinima i drugim teškim kršenjima ljudskih prava počinjenim na području nekadašnje SFRJ*. Beograd: Fond za humanitarno pravo.

Rangelov, I. and R. Teitel (2014). *Transitional Justice*, Chapter 19, pp. 338–352. Chichester: John Wiley & Sons, Ltd.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science 58*(4), 1064–1082.

Sanders, L. M. (1997). Against Deliberation. *Political Theory 25*(3), 347–376.

Sandole, D. J. D. and I. Staroste (2015). Making the Case for Systematic, Gender-Based Analysis in Sustainable Peace Building. *Conflict Resolution Quarterly 33*(2), 119–147.

Schulz, P. (2020). *Male Survivors of Wartime Sexual Violence: Perspectives from Northern Uganda*. Berkeley: University of California Press.

Selimovic, J. M. (2020). Gendered silences in post-conflict societies: a typology. *Peacebuilding 8*(1), 1–15.

Sharp, D. N. (2013). Interrogating the Peripheries: The Preoccupations of Fourth Generation Transitional Justice. *Harvard Human Rights Journal 26*, 149–178.

Sharratt, S. (2011). *Gender, Shame and Sexual Violence: The Voices of Witnesses and Court Members at the War Crimes Tribunals*. New York: Routledge.

Shepherd, L. J. (2017). *Building Peace: Feminist Perspectives*. London and New York: Routledge.

Steenbergen, M. R., A. Bächtiger, M. Spörndli, and J. Steiner (2003). Measuring Political Deliberation: A Discourse Quality Index. *Comparative European Politics 1*(1), 21–48.

Stegmueller, D. (2013). How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches. *American Journal of Political Science 57*(3), 748–761.

Steiner, J. (2012). *The Foundations of Deliberative Democracy: Empirical Research and Normative Implications*. Cambridge: Cambridge University Press.

Steiner, J., A. Bächtiger, M. Spörndli, and M. R. Steenbergen (2005). *Deliberative Politics in Action: Analysing Parliamentary Discourse*. Cambridge: Cambridge University Press.

Sunstein, C. R. (2007). Ideological Amplification. *Constellations 14*(2), 273–279.

Swaine, A. (2018). *Conflict-related violence against women: transforming transition*. Cambridge: Cambridge University Press.

Terman, R. (2017, 11). Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage. *International Studies Quarterly 61*(3), 489–502.

Theidon, K. (2013). *Intimate Enemies: Violence and Reconciliation in Peru*. Philadelphia: University of Pennsylvania Press.

Thompson, D. F. (2008). Deliberative Democratic Theory and Empirical Political Science. *Annual Review of Political Science 11*(1), 497–520.

Toshkov, D. (2016). *Research Design in Political Science*. Basingstoke: Palgrave Macmillan.

Tripp, A. M. (2015). *Women and Power in Postconflict Africa*. Cambridge Studies in Gender and Politics. Cambridge: Cambridge University Press.

UN Women (2015). A Global Study on the Implementation of United Nations Security Council resolution 1325. Technical report, UN Women.

Wallensteen, P. (2015). *Quality Peace: Peacebuilding, Victory & World Order*. Oxford: Oxford University Press.

Warren, R., A. Applebaum, B. Mawby, H. Fuhrman, R. Turkington, and A. Mayesha (2017). Inclusive Justice: How Women Shape Transitional Justice in Tunisia and Colombia. Technical report, Georgetown Institute for Women, Peace and Security.

Waylen, G. (2014). A Seat at the Table—Is it Enough? Gender, Multiparty Negotiations, and Institutional Design in South Africa and Northern Ireland. *Politics and Gender 10*(4), 495–523.

West, E. A. (2017). Descriptive Representation and Political Efficacy: Evidence from Obama and Clinton. *The Journal of Politics 79*(1), 351–355.

Wood, E. J. (2014). Conflict-related sexual violence and the policy implications of recent research. *International Review of the Red Cross 96*(894), 457–478.

Yarwood, L. (2013). Women and Indigenous Conflict. In L. Yarwood (Ed.), *Women and Transitional Justice: The Experience of Women as Participants*. Abingdon: Routledge.

Yoder, J. D. (1991). Rethinking Tokenism: Looking Beyond Numbers. *Gender & Society 5*(2), 178–192.

Young, I. M. (2001). Activist Challenges to Deliberative Democracy. *Political Theory 29*(5), 670–690.

Appendix

A The RECOM Initiative in the Balkans: The Background

The Coalition for RECOM, which stands for the Regional Commission for Establishing the Facts about All Victims of War Crimes and other Serious Human Rights Violations committed on the territory of the former Yugoslavia from January 1991 to 31st December 2001, is a network of non-governmental organizations, associations, and individuals who support a regional transitional justice process (Proces REKOM 2011). RECOM is a regional network, and as such is distinct from global transnational networks whose membership includes representatives of international institutions, international non-governmental organizations (NGOs) and/or national governments. The RECOM initiative embodies a regional rather than a commonly pursued state-centric approach to justice, either through trials or truth commissions. Also, in contrast to internationally-imposed instruments of post-conflict justice, the RECOM is a locally-driven initiative (Rangelov and Teitel 2014).

This initiative emerged as a response to a complex post-conflict legacy in successor states of former Yugoslavia, characterized by a cross-border nature of crimes. It is also a response to the limits of the international strategy of 'exogenous justice' pursued through the International Criminal Tribunal for the former Yugoslavia at The Hague (ICTY). Notably, with its focus on victims, the RECOM is a response to inability of trials that focus on the perpetrator and the punishment, either to acknowledge the suffering of the victims or to promote reconciliation.

Responding to this legacy, together with the Documenta, an NGO from Zagreb, Croatia, and the Investigative-Documentation Centre, an NGO from Sarajevo, Bosnia and Herzegovina, the Centre for the Humanitarian Law, from Belgrade, Serbia, initiated a regional approach to transitional justice in the Balkans. In May 2006, they launched a debate on the mechanisms for establishing and documenting facts of war crimes in former Yugoslavia. In May 2008, this initiative transformed into the Coalition for RECOM. From then on the consultative process focused on building a model of a regional fact-finding commission (Kandić 2007). The coalition has amassed significant membership of nearly 2,000 NGOs, associations, groups, victims, prominent individuals, veterans,

lawyers, artists, journalists, academics, and youth – from all areas of the former Yugoslavia. But, the RECOM coalition's reach was much wider owing to the consultative process under its auspices. The consultations were debates about how best to address the criminal legacy throughout the Balkans. They involved nearly 6,000 civil society members from all ethnic groups affected by the wars fought in the former Yugoslavia in the 1990s (Serbs, Croats, Muslims, Albanians, Montenegrins, Macedonians, Slovenians, including members of minority groups in the region). The consultations, held at the regional, national and local levels, unfolded in two stages.

In the first stage, the consultations were of a general nature, and produced an agreement on a regional approach to transitional justice (Humanitarian Law Center 2009). The second stage of consultations was focused on the proposed draft Statute. This document was compiled by the Working Group that was tasked to translate the ideas about a regional approach to transitional justice heard during the consultation process into a document with specific provisions. These proposals were then put up for the discussion before the broadest section of civil society stakeholders. Despite the legacy of violence and the diversity of views, the consultative process produced a cross-ethnic agreement on the Statute for the regional fact-finding commission.

Table A.1 shows the summary of the 20 consultations where the composition of the draft Statute was discussed. It shows that the meetings covered a broad range of geographical locations, included a diverse part of the public, both general, as well as those directly affected by the conflict.

B The Draft Statute: The Process and Data

The deliberation on the draft Statute of the regional fact-finding commission comprised the last stage of the consultative process, and lasted from May 2010 to March 2011[36]. The draft Statute spelled out the commission's mandate. It contained the provisions that

[36]Statut: Predlog Regionalne komisije za utvrdjivanje činjenica o ratnim zločinima i drugim teškim kršenjima ljudskih prava na području nekadašnje SFRJ, 26 March 2011. Paper copy on file with one of the authors.

Table A.1: Summary of RECOM Statute consultation meetings

ID	Date	Country	Place	Level	Community
1	2010-05-29	Montenegro	Podgorica	regional	general
2	2010-05-29	Bosnia	Tuzla	non-regional	victims
3	2010-06-01	Croatia	Zagreb	non-regional	general
4	2010-06-05	Bosnia	Banja Luka	regional	general
5	2010-07-03	Serbia	Beograd	non-regional	victims
6	2010-07-13	Croatia	Osijek	non-regional	general
7	2010-07-14	Croatia	Vukovar	non-regional	general
8	2010-09-02	Croatia	Knin	non-regional	general
9	2010-09-10	Slovenia	Ljubljana	regional	professional
10	2010-09-15	Kosovo	Priština/Prishtinë	non-regional	victims
11	2010-09-18	Bosnia	Sarajevo	regional	victims
12	2010-10-22	Croatia	Pakrac	non-regional	general
13	2010-06-11	Croatia	Zagreb	regional	professional
14	2010-12-04	Serbia	Beograd	regional	professional
15	2010-08-28	Bosnia	Mostar	regional	general
16	2010-12-17	Kosovo	Priština/Prishtinë	regional	victims
17	2010-12-17	Croatia	Zagreb	regional	general
18	2010-12-18	Macedonia	Skoplje	regional	general
19	2011-01-23	Serbia	Beograd	regional	professional
20	2011-01-29	Bosnia	Sarajevo	non-regional	general

would regulate all aspects of the commission's work, including: the remit, the seat, official languages, the procedures for establishing the Commission, such as appointment of commissioners, modalities of the commission's operation, such as summoning of witnesses, type of hearings, relationship with the judiciary, and the commission's report. The Working Group comprising a multi-ethnic team of legal experts from the former Yugoslavia, drafted their initial proposal based on the consultations held prior to the deliberations on the draft Statute and on the analysis of statutes of other national truth and reconciliation commissions in other post-conflict cases globally, while taking into account the laws of all former Yugoslav countries. The consultations were held with a wide range of stakeholders such as survivors and family members of victims, human rights activists, journalists, teachers, veterans, lawyers and representatives of the youth groups.

Each consultation was a one- or two-day long session. During the consultations the participants had an opportunity to hear and consider various proposals on each article of the draft Statute, and express their views. The proceedings were transcribed verbatim in their entirety, and have been publicly available on the RECOM's website. Keeping a meticulous record of the consultative process had a two-fold purpose: documenting and tracking the diversity of opinions before settling on the final version of the draft Statute, as well as ensuring that this local transitional justice process is transparent and open to scrutiny. Each consultative debate was dedicated to the same issue areas that corresponded with the headings in the draft Statute as outlined above.

The draft Statute was adopted at the Assembly of the RECOM Coalition on 26 March 2011. The Assembly is one of the RECOM's governing bodies (alongside the Secretariat), and is comprised of the members for the Coalition. Without a hard and fast rule on the membership of the Assembly, the Coalition considered members to be active participants of the Coalition and of the consultation process, but made sure to maintain ethnic representation of all groups involved in the conflicts of Yugoslavia's dissolution. In practice, the draft Statute was created by the participants of the consultative process as it reflected their views presented during the debates[37]. For the purpose of our analysis, the act of the adoption of the draft Statute cannot be interpreted as being a result of qualitatively different dynamics than those that characterized the debates.

The draft Statute subsequently underwent minor amendments at the RECOM's Assembly meeting, on 14 November 2014[38]. Notably, in the revised version, the reference to the representation of one female commissioner at least from each prospective state backing the Commission was removed[39]. These amendments were a part of the institutionalization process of the RECOM initiative. It consisted of appointing advocates representing

[37]See (Proces REKOM 2011).

[38]See Izmene Statuta REKOM, 14.11.2014. at http://recom.link/sr/izmene-statuta-rekom-28-oktobar-2014-2/

[39]See Article 24, Criteria for Selection of Commissioners, Izmene Statuta Regionalne komisije za utvrdjivanje činjenica o ratnim zločinima i drugim teškim kršenjima ljudskih prava na području bivše SFRJ na osnovu predloga Izaslanika predsednika/Predsedništva BiH za REKOM, 28 October 2014, available at http://recom.link/wp-content/uploads/2014/11/SR-Izmene-Statuta-FINAL-12.11.2014-ff.pdf

the RECOM Coalition, and their engagement with state authorities in former Yugoslav states, resulting in the changes to the draft Statute.

The final document - the draft Statute - is a major achievement as it represents a consensual outcome following deliberation that included representatives of different ethnic groups involved in the Balkan conflicts of the 1990s and the early 2000s. However, from the perspective of gender inclusiveness, the draft Statute represents a typical case of a gender-insensitive outcome of a transitional-justice process. The draft Statute includes the reference to 'rape and other grave forms of sexual abuse' in the definition of a war crime (Koalicija za REKOM 2011), but falls short of including provisions that would ensure women's equality in many facets of founding and running the commission; nor does it contain arrangements appropriate for addressing war-time sexual violence, despite the fact that public hearings of victims represent a lynch pin of this restorative transitional justice process. Consequently, the consideration of the gender dimension in the Statute has been considered a weakness of the RECOM process (Bonora 2019).

C Participants

Table C.2 shows the number of participants and moderators of both genders at each of the consultations. The range is, roughly, between 20 and 70 with only consultation with the former political prisoners in Kosovo (#10), without any women women discussants. Overall, while the number of women discussants tends to be lower, this difference is not stark.

Full speech participation models are given in Table C.3.

D Robustness checks of turn-taking effects

As described in the article we find significant negative effect of gender on the number of speeches delivered in a sequence. Given that the number of direct interruptions is very small, as our manual coding of transcripts has shown, the mechanism driving this

Table C.2: Summary of consultation participants

		Moderators		Discussants	
ID	**Participants**	**Women**	**Men**	**Women**	**Men**
1	49	2	1	26	20
2	32	1	2	9	20
3	34	2	1	20	11
4	39	5	0	18	16
5	41	1	2	11	27
6	18	2	1	7	8
7	27	1	2	16	8
8	26	1	1	10	14
9	47	4	2	24	17
10	47	2	1	0	44
11	70	3	4	22	41
12	51	2	2	22	25
13	54	2	1	24	27
14	50	4	2	17	27
15	45	5	6	14	20
16	60	3	2	21	34
17	58	2	2	30	24
18	44	2	3	6	33
19	27	3	3	11	10
20	27	2	3	10	12

result could be higher probability of men being followed by men, rather than women being followed by women. Here we provide an alternative modelling of the effect, complimentary to the Poisson specification found in the article. Table D.4 shows the models with the binary outcome of whether the speaker's gender at each turn alternates, given the gender of a previous speaker. As these are essentially lagged models, in all cases we also disregard the very first speaker in each of the 20 consultations.

E Corpus Summary

The text corpus consists of 20 debates that were held in the languages spoken in the Balkans. These include Slovenian, Macedonian, Albanian, Serbian, Croatian, Bosnian and Montenegrin languages. The multi-language nature of the corpus presents a particular

Table C.3: Multi-level models of speech participation with 95% HPD intervals in parentheses

	Log(words)	
	(1)	(2)
Sex (ref: Male)		
Female	-0.116 (-0.252, 0.031)	-0.112 (-0.25, 0.028)
% Female Discussants		0 (-0.033, 0.033)
Diversity (ref: Mono-ethnic)		
Dyadic		-0.204 (-1.813, 1.451)
Multi-Ethnic		0.176 (-1.363, 1.703)
Level (ref: Non-regional)		
Regional		-0.207 (-1.656, 1.328)
Type (ref: General)		
Professionals		-0.068 (-1.149, 0.988)
Victims		-0.075 (-1.257, 1.092)
Translation (ref: No)		
Yes		0.123 (-1.017, 1.27)
Intercept	4.917 (4.624, 5.198)	4.951 (3.33, 6.611)
σ_y	1.273 (1.23, 1.319)	1.274 (1.228, 1.323)
σ_α	0.386 (0.178, 0.767)	0.625 (0.244, 1.377)
log-posterior	-2481.234	-2488.032
Groups	20	20
Observations	1472	1472

challenge for quantitative text analysis. It raises the question of availability of tools for translation, should researchers not be familiar with the language(s) of the data. This issue can be resolved in a straightforward way if all documents (assuming that each document represents a debate) are in the same language, or at least each document individually is in one language. The RECOM debate presented a double challenge. Some

Table D.4: Multi-level logistic models of changes in speaker's gender with 95% HPD intervals in parentheses

	Speaker's gender alternates	
	(1)	(2)
Previous speaker (ref: Male)		
Female	0.141 (0.088, 0.194)	0.134 (0.081, 0.186)
% Female Discussants		0.007 (0.002, 0.011)
Diversity (ref: Mono-ethnic)		
Dyadic		-0.106 (-0.346, 0.122)
Multi-Ethnic		-0.162 (-0.397, 0.044)
Level (ref: Non-regional)		
Regional		0.119 (-0.099, 0.349)
Type (ref: General)		
Professionals		0.022 (-0.132, 0.176)
Victims		0.083 (-0.078, 0.233)
Translation (ref: No)		
Yes		-0.082 (-0.245, 0.079)
Intercept	0.313 (0.249, 0.378)	0.104 (-0.097, 0.33)
$\hat{\sigma}_y$	0.468 (0.451, 0.485)	0.468 (0.451, 0.485)
$\hat{\sigma}_\alpha$	0.015 (0.005, 0.034)	0.007 (0, 0.025)
log-posterior	-992.762	-1000.535
Groups	20	20
Observations	1452	1452

debates were held entirely in a single language, i.e. one debate in Serbian and one in Albanian. However, most debates were transcribed in multiple languages, as speakers from different ethnic groups joined in the discussions. From the practical perspective of multi-ethnic deliberation this did not present a problem as Serbian, Croatian, Bosnian and Montenegrin languages, spoken by most participants in the debates, are mutually

intelligible. This, however, is not the case for Albanian, Macedonian and Slovenian that had to be translated. However, from the point of view of quantitative text analysis, the differences even in the languages that are mutually intelligible and do not require translation are both lexical (e.g. Serbs use the word 'mleko' and Croats 'mlijeko' for milk; or 'hleb' and 'kruh' for bread) and grammatical imply that these languages de facto have to be treated as different languages. These differences derive from their historical development as variants of the Slavic language. Following the violent break-up of former Yugoslavia, languages were subject to nationalization. Language was used to assert the identity of newly-independent nations. Linguistic engineering also included banishing words commonly used by different nations (Bugarski 2009). The result was a greater 'cultural and linguistic separation' (Kuhiwczak 1999).

The descriptive summary statistics of the consultation corpus are shown in table G.2. While there is considerable variation in the length of consultations, most of them contain enough data for the application of quantitative text analysis in general and structural topic models in particular.

F Intercoder reliability

We assess intercoder reliability by calculating raw percentage of agreement, Cohen's κ (Cohen 1960) and Krippendorff's α (Krippendorff 2004) for each individual component of the DQI. All speech acts were coded by two independent coders (once by one of the authors and separately by a research assistant after extensive training). The estimated reliability indices are shown in Table F.6. The α-agreement ranges from ~65% to 90%, which is not dissimilar from the previous application of comparable coding schemes in the literature (Gerber et al. 2016) and, given the overall complexity of the task, represents an acceptable level of agreement.

Table E.5: Summary of the corpus of consultation transcripts

ID	Tokens	Types	Sentences	Utterances	Speech Acts
1	15568	9195	645	63	27
2	24214	13214	939	66	31
3	33085	19914	1622	342	103
4	29466	17850	990	184	88
5	65862	35305	2642	252	61
6	18903	10267	841	46	67
7	20389	11379	792	99	27
8	19616	9846	800	54	18
9	22806	12885	913	111	12
10	41317	22781	1851	220	31
11	23302	13383	1060	99	38
12	20694	11641	952	57	34
13	37390	22618	1584	145	142
14	26158	15416	1146	67	101
15	29986	16380	1269	72	66
16	20052	11716	987	130	40
17	22057	12515	601	124	68
18	27166	14020	1216	148	28
19	59164	38553	2952	714	194
20	20737	11380	917	138	35
Total	577932	330258	24719	3131	1211

G Measuring Quality of Deliberation

To aggregate the components we fit a standard two-parameter IRT model of the following form:

$$logit^{-1}(P(x_{ij} = 1|\gamma_j, \alpha_i, \beta_j)) = \gamma_j(\alpha_i + \beta_j)$$

In other words, we are interested in the probability of speech act i satisfying the deliberation component j (being coded as 1), given the quality of a speech act α_i and the difficulty, β_j, and discrimination, γ_j of a given component. We fit the model in Stan (Carpenter et al. 2017) by running three Markov chains with 10'000 iterations from randomly generated starting values. We use the standard uninformative prior specification for the

Table F.6: Inter-coder reliability for DQI coding

Components	**Agreement**	**Cohen's κ**	**Krippendorff's α**
Interruption	98.76	0.821	0.821
Justification Rationality	78.78	0.652	0.652
Content (common good)	92.49	0.671	0.671
Content (difference)	88.93	0.681	0.681
Content (abstract)	95.21	0.710	0.710
Respect (participants)	93.48	0.678	0.678
Respect (groups)	90.67	0.661	0.661
Story Justification	98.84	0.907	0.907

model:

$$\alpha \sim N(0, 1)$$

$$\beta \sim N(\mu_\beta, \sigma_\beta)$$

$$\mu_\beta \sim Cauchy(0, 5)$$

$$\sigma_\beta \sim Cauchy(0, 5)$$

$$\gamma \sim LN(0, \sigma_\gamma)$$

$$\sigma_\gamma \sim Cauchy(0, 5)$$

Figure G.1 shows convergence diagnostics for γ parameters. In the interest of space traceplots for other parameters are omitted, but are available upon request.

Table G.7 shows how the levels of the original categorical or ordinal variables assigned to each speech act were dichotomised into binary items for aggregation into single composite score.

To evaluate and compare the model we also aggregate DQI component by running principal component analysis and extracting the first principal component, doing factor analysis and using factor scores on the first factor, as well as calculating summative index. Figure G.2 shows correlations between the aggregated scores the quality of deliberation estimated through different methods.

Table G.7: Aggregation of DQI categories

Component	Aggregation Code	Original Label
Interruption	0	interruption
	1	normal participation
Justification Rationality	0	no justification
		inferior
	1	qualified
		sophisticated
Content (common good)	0	neutral (no reference)
	1	ethnic group
		my country
		my region/multie-ethnic
Content (difference)	0	no reference
	1	reference
Content (abstract)	0	no reference
	1	reference
Respect (participants)	0	negative (disrespectful, foul language)
		no reference
	1	neutral reference
		positive (explicitly respectful)
Respect (groups)	0	other groups denigrated
		not mentioned
	1	neutral (mentioned but not denigrated)
		explicit respect
Story Justification	0	no story
	1	unrelated story
		related story (sole justification)
		related story (reinforces rational justification)

Figure G.1: Convergence diagnostic of DQI aggregation

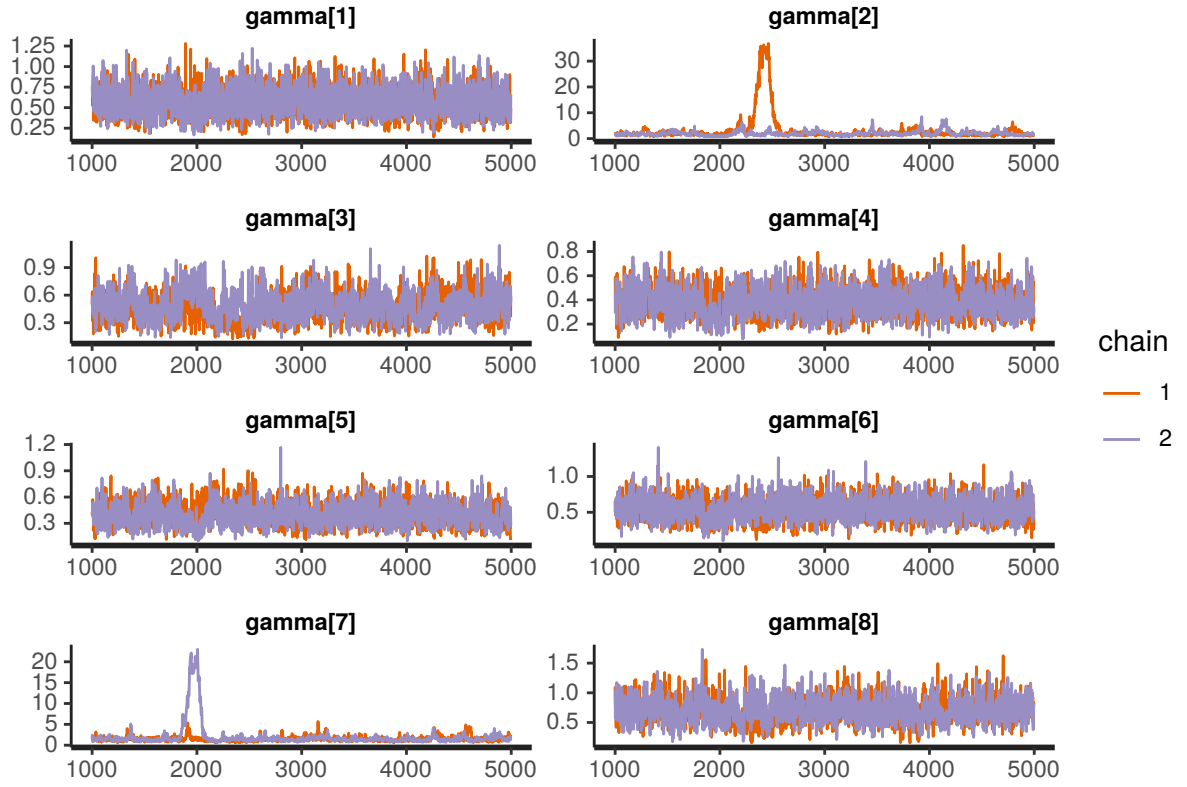
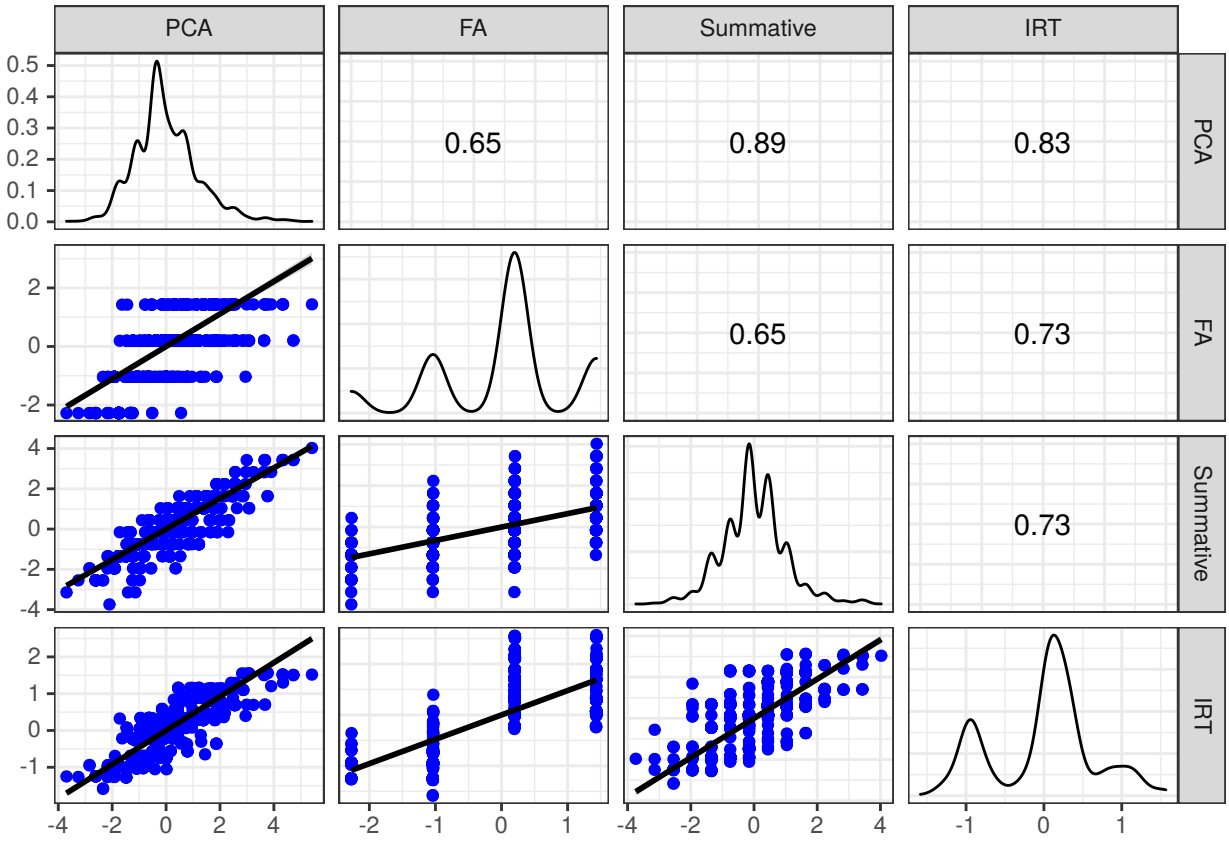


Figure G.2: Comparison of DQI aggregation methods



H Topic Models

Despite recent work on the importance of preprocessing decisions for quantitative text analysis in general (Denny and Spirling 2018) and topic models in particular (Schofield and Mimno 2016,@Schofield2017), this literature looks only at data in the English language. Some experimentation with the corpus has shown that the use of a stopwords list (as there is no off-the-shelf stopwords list in Serbian, it has been compiled by one the authors) and lemmatization largely resulted in more interpretable estimates without changing the conclusions substantively. A more advanced form of stemming, *lemmatization* involves standardizing words into linguistically meaningful *lemmas* rather than truncating words to their technically convenient, but often uninformative *stems*. This approach is more computationally involved due to required part-of-speech resolution prior to lemmatization. However, it is necessary for Serbian which is a heavily conjugated language in comparison to English. To prepare the corpus for the analysis, we further removed punctuation and converted all tokens into lower case. Numbers were retained as a considerable part of the discussion revolves around specific provisions and articles of the Statute, which are labelled with numbers. In addition, we removed all the word types that occurred fewer than 10 times across all consultations. For model fitting, we aggregate all speeches at the speaker-level as individual documents. As most discussants participated in only one consultation, here we use the characteristics of their first consultation

To determine the number of topics we compared held-out likelihood, semantic coherence and residuals presented in figure H.3, as well substantive interpretability of models with 5 to 50 topics. The model with 10 topics yielded the best trade-off between different criteria. This model offers the highest held-out likelihood, estimated on portion of the words that were held-out during model training, as well as semantic coherence or pointwise mutual information. In other words, the extent to which the words that more probable under the a topic co-occur within the same document. Furthermore, it provides the largest improvement in residuals (how far the sample dispersion is from 1, the closer the better). Table H.8 further provides an output from a range of algorithms used to

generate sets of words describing each topic.

As a robustness test to show the main results, namely, between-gender differences in topics evaluation and implementation, are not sensitive to the number of topics we present an alternative specification with 5 topics. Figures H.4 and H.5 show the translated words with highest probabilities for each topic and the estimated effects of gender on topic prevalence. We have also fitted a model with all topic prevalence covariates, apart form gender, excluded. It did not affect the substantive findings and we omit it in the interest of space. These results are available upon request.

Figure H.3: Diagnostics of STM fit

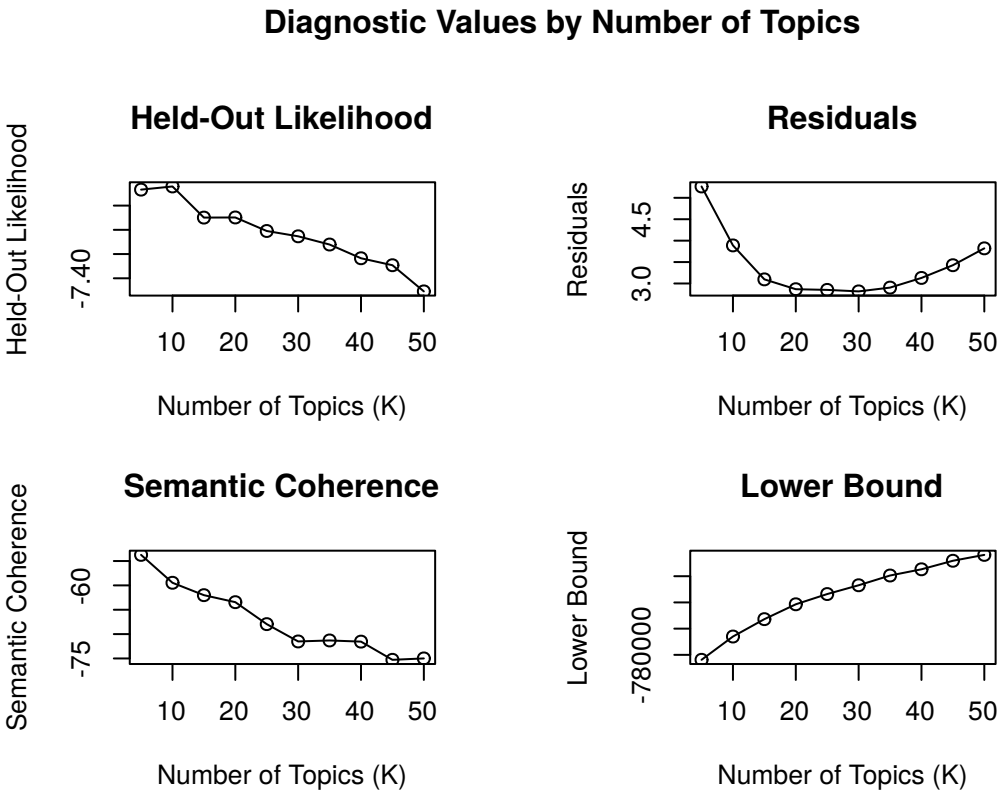


Figure H.4: Distribution over words in 5 topics

Figure H.5: Topical Prevalence by Male and Female Discussants for 5 Topics

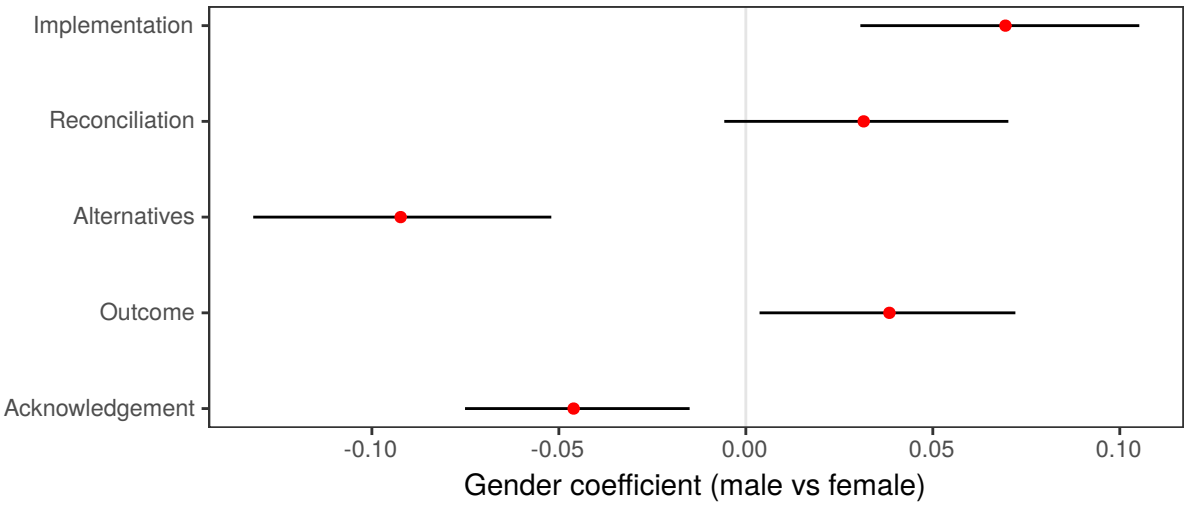


Table H.8: Topic Labelling

Topic	Labelling Algorithm	Words
1	prob	krivični, komisija, sud, misliti, kazati, član, lice
	frex	krivični, učinilac, sumnja, tužilac, izjava, davanje, priznanje
	lift	47, hitan, neispunjavanje, nepouzdan, obezbeđenje, 1.09, 2009
	score	krivični, sumnja, učinilac, postupak, tužilac, kazna, izjava
2	prob	zločin, pravo, ratni, žrtva, činjenica, ljudski, komisija
	frex	čovečnost, kršenje, alternativa, oružan, popis, težak, utvrđivanje
	lift	amirov, dal, izvorište, konsultacioni, konzultacija, ograničavajući, osuđenik
	score	kršenje, alternativa, čovečnost, zločin, oružan, definicija, popis
3	prob	godina, čovek, reći, kazati, porodica, nestali, rat
	frex	zrenjanin, zatvor, državljanstvo, krst, naprimer, crven, selo
	lift	feniks, izgoreti, plav, 88, 93, bala, bljesak
	score	zrenjanin, euforija, mentalitet, 213, metak, krst, naprimer
4	prob	član, država, komisija, misliti, statut, rekom, predsednik
	frex	panel, članica, selekcioni, izbor, sposobnost, kriterijum, donositi
	lift	aneks, blisko, dvotrećinski, funkcioniranje, haotičan, isključenje, izmjena
	score	panel, selekcioni, sposobnost, kvorum, osobina, dvotrećinski, članica
5	prob	rat, žrtva, čovek, godina, misliti, zločin, država
	frex	veteran, prošlost, budućnost, mlad, aleksić, suočavanje, otmica
	lift	06, 1.000.000, 1.059, 114, 16.500, 1968, 1971
	score	aleksić, veteran, nacionalizam, škola, pančev, neprijateljski, brigada
6	prob	rekom, država, žrtva, pitanje, dokument, određen, izveštaj
	frex	tajan, dokument, tribunal, preporuka, haški, aspekt, informacija
	lift	blajburg, delomičan, izbalansirati, jednosmeran, kompetentan, kulturno, kupovati
	score	dokument, izveštaj, tajan, preporuka, institucionalan, dobrovoljnost, ombudsman
7	prob	čovek, godina, doći, znati, reći, žrtva, kazati
	frex	vukovar, grad, avionski, logor, ispričati, ti, suditi
	lift	auto, bristol, kafa, kajati, obnova, otkaz, porušiti
	score	avionski, prijedor, grahovo, doboj, delikt, trifunović, verbalan
8	prob	godina, znati, kosovo, reći, žrtva, zločin, rekom
	frex	euleks, opraštati, posmrtni, oprostiti, 1999, ostatak, albanija
	lift	1244, baletić, dick, kidnaper, marti, milijana, neprimenjiv
	score	euleks, oproštaj, žaljenje, opraštati, oteti, metohija, 1998
9	prob	bosna, hercegovina, žrtva, broj, kosovo, rekom, govoriti
	frex	hercegovina, bosna, broj, gora, dubrava, opcija, sarajevo
	lift	divjak, komisionar, munira, opkoliti, reprezentativan, sejdić, tlo
	score	dubrava, hercegovina, srbin, kosovo, intenzitet, bošnjak, vučitrn
10	prob	misliti, rekom, čovek, reći, znati, godina, važan
	frex	zanimati, 5, kontekst, struka, važan, vremenski, odeljenje
	lift	1913, dačić, destruktivan, duhovan, gordan, indirektno, izvaditi
	score	sprema, 5, trn, odeljenje, obrazovni, okupiti, indirektno

I Software statement

The analysis was run under Linux Ubuntu 18.04 using R version 3.5.1 (R Core Team 2018). We relied on the following R packages in our empirical analysis:

`bayesplot` (Gabry and Mahr 2018),

`dplyr` (Wickham et al. 2018),

`GGally` (Schloerke et al. 2018),

`ggplot2` (Wickham 2016),

`irr` (Gamer et al. 2012),

`kableExtra` (Zhu 2018),

`knitr` (Xie 2018),

`lme4` (Bates et al. 2015),

`ltm` (Rizopoulos 2006),

`lubridate` (Grolemund and Wickham 2011),

`magrittr` (Bache and Wickham 2014),

`pander` (Daróczi and Tsegelskyi 2018),

`processx` (Csárdi and Chang 2018),

`readr` (Wickham, Hester, and Francois 2017),

`rstan` (Stan Development Team 2018),

`rstanarm` (Stan Development Team 2016),

`stm` (Roberts, Stewart, and Tingley 2018),

`stringi` (Gagolewski 2018),

`stringr` (Wickham 2018),

`tibble` (Müller and Wickham 2018),

`tidyr` (Wickham and Henry 2018), and

`quanteda` (Benoit et al. 2018).

Chapter 4

Record Linkage with Text: Merging Data Sets When Information is Limited

ABSTRACT

The recent years have seen the emergence of new, more scalable ways to link information about different entities across multiple data sources. However, merging data sets when the number of variables used for record linkage is restricted remains challenging. In this paper I consider the case when the information is limited to a single multi-token text string. This situation often occurs when researchers work with organization names, user accounts or any other short labels. Using Lobbying Disclosure Act data I illustrate substantive implications that the choice of record linkage approach can have in empirical research. I review the existing approaches and consider three types of noise that can typically be encountered in this scenario: character-level, word-level or a combination of both. Furthermore, I conduct a simulation study showing the sensitivity of the existing approaches to the presence of errors occurring at different levels. The results suggest that the optimal choice of a record linkage approach depends on contextual knowledge about the most likely type of noise, as well as stress the need to conduct sensitivity analysis using different record linkage approaches.

4.1 Introduction

Political science research is increasingly relying on more than one source of data (Brady, 2019). This brings exciting new opportunities for empirically testing old and new theories, that previously were inaccessible to scholars due to limitations of disparate datasets. However, this advancement comes at a cost. While in some cases it is possible to unambiguously link multiple datasets with few to no errors, this task becomes far more challenging in other circumstances. For example, in a cross-country analysis, it is usually feasible to enumerate all labels or codes that can be possibly used to refer to the same state. In many other cases this task is far from straightforward. Merging together datasets that contain individuals (Enamorado et al., 2019), organisation names (Bonica, 2014; Kim, 2017) or event records (Donnay et al., 2019) requires more elaborate design of linkage procedure. Some of the adopted approaches have a long lineage, when they have been used for decades in census and survey research (Newcombe et al., 1959; Newcombe and Kennedy, 1962), other, more recently developed (Sadinle, 2017), are yet to be tested on the types of data common in political science.

One particular task, not infrequently encountered in applied research, is the merging of multiple data sets when the only variable that they have in common is the text field containing the name of organization, geographic area or article name. In this paper I argue that this condition poses a distinct problem that cannot be adequately addressed neither by the methods developed within the emerging literature on text matching (Roberts et al., 2018; Mozer et al., 2019) due to the short length of the available text, nor by the more established approaches used for linking individuals (Fellegi and Sunter, 1969; Enamorado and Imai, 2020), due to exclusive usage of information contained in the name. The complications arise due to the unobserved nature of the noise found in the real-life data. More specifically, I consider three types of noise: (1) *character-level* corruption, (2) *word-level* corruption and (3) *combination* of the first two. Character-level noise is likely to occur as a result of errors introduced by the optical-character recognition (OCR), e.g.

when the scanned archival materials are processed with the specialized software, or as a typo made by a human operator manually entering the data. Word-level distortion can be encountered when there are multiple accepted names that are used to refer to the same geographic or corporate entities (e.g. the same parliamentary constituency can be referred to as "Cities of London and Westminster" or "Cities of London & Westminster"). In many scenarios, however, a researcher is likely to encounter the third type, some combination of character and word-level noise. In this paper I use a simulation study and a real-world dataset to evaluate the performance of different record linkage approaches to problems where information for matching is limited to text. First, to make the results of the study language- and dataset-agnostic and isolate the effects of different types of noise, I use simulated data that is generated to reflect the features of real text labels and then introduce randomly introduce noise of a known type. For additional tests I further use a dataset of organization names in the UK to assess the performance on real data. I argue that while none of the current approaches can be taken as a "silver bullet", empirical researchers can leverage their background knowledge about data-generating process to account for the most likely type of text distortion. At the same time, when core substantive results critically rely on a merged dataset it is important to conduct sensitivity analysis, showing the consistency of findings across different approaches to record linkage.

The purpose of this paper is three-fold. First, I review the currently available approaches for dealing with textual labels as the only available identifier. I conduct a case study, using LDA disclosure and PPP loans data to illustrate the substantive implications that the choice of a record linkage approach can have in practice. Second, I conceptualise and introduce three different types of noise that can be observed in such labels. And, third, I evaluate currently existing methods for record linkage based on these labels using a common metric, which makes them more comparable with each other and elicits different trade-offs that this task faces[1]. To assess their performance I use simulated and real

[1]While there are existing review articles on name matching (Cohen et al., 2003), they do not cover more recent developments in this area and the comparison is done on data sets that are considerably

datasets that are more akin to those that are encountered in political science literature.

The rest of the paper is organised as follows. In the following section I present the historical overview of the development of record linkage. Then I review the currently available approaches and conduct an illustrative case study. After that I outline a framework of analysing labels as textual data. In the fifth section. I evaluate these methods on the simulated and real data sets. And, lastly, I show their performance on real-world data.

4.2 Background

With its origins in public health and epidemiological research (Dunn, 1946), record linkage[2] has been advocated as an alternative to expensive large-scale longitudinal studies (Jutte et al., 2011). The first key insight was that the odds of observing an agreement pattern in a pair of records carries different amount of information depending on how rarely the values in those records occur in the population (Newcombe et al., 1959; Newcombe and Kennedy, 1962). In other words, the two records that contain the same first name "Catherine" are much less likely to be a true match than those with "Stavroula". Conveniently, the probabilities of observing different values for first names, last names, etc could be calculated from census data. Furthermore, as noted by the authors, the same logic could applied beyond names. For instance, the place of birth also has different discriminating power depending on the population of a given location and how common its name is.

The statistical foundation of most contemporary implementations of record linkage algorithms was laid in the seminal work of Fellegi and Sunter (1969). They proposed separation of record pairs into matches, non-matches and potential matches given the vector comparing individual fields for this pair (agreement pattern). While the assumption

different from those used by political scientists.

[2]These two sections are meant to provide a brief overview of record linkage as applicable to textual labels. For more thorough introduction to broader record linkage literature see Christen (2012).

that fields are independent from each other is likely to be violated in reality this often does not substantially affect the results. Another critical step of using Fellegi-Sunter framework is the calculation of posterior probabilities of a record pair being in each match class, given the agreement pattern. In practice, this is usually done either by relying on a "gold standard" dataset or using Expectation-Maximization (EM) algorithm for estimating unknown parameters (Winkler, 2000; Enamorado et al., 2019). The specific details of how agreement pattern is to be estimated varies across the tasks. While for the date of birth this could be a simple difference between the two dates in days or years, in case of names this would typically be the string distance expressed in the number of characters that need to be changed in order to convert one name into another.

So far most of the literature has been concerned with the task of matching records of individuals[3], who provide a range of characteristics for comparison, such as names, age, gender, address, etc. While this is hardly surprising given the amount of data collected by censuses and large-scale surveys on an annual basis, this leaves the question of how well the currently available approaches perform on tasks, where there the information is restricted to only a single variable containing textual data. In the following section I will review the key approaches.

4.3 Existing Approaches

Currently, there are four main approaches available to applied researchers for linking observations across multiple data sets. Table 4.1 summarizes them and their key features and drawbacks.

The first possible solution is to simply merge two data sets treating the label field as a unique identifier and linking rows that contain exactly the same names. Although this approach is very crude, depending on the label structure and the nature of noise in the data, it can offer high precision with low recall. More specifically, its performance depends

[3]However, see Cohen et al. (2003).

on whether the *uniqueness* assumption is met. While in many instances geographic units, organization names and newspaper titles can be assumed to be unique or almost unique in the population, this assumption appears highly implausible for individual names, surnames, gender and even addresses. Thus, one would not expect a similar number of false positives when applying simple merge to labels. At the same time this approach critically depends on labels being universal across different sources and limited noise from data entry, such as optical character recognition or clerical input.

Table 4.1: **Summary of Record Linkage Approaches.**

Method	Nature	Features	Example
Simple Merge	Deterministic	Unique labels assumed	SQL joins
String Distance	Similarity score	+/- Character-level comparison - Arbitrary threshold	Levenshtein Jaro-Winkler
Text Similarity	Similarity score	+/- Word-level comparison - Arbitrary threshold	cosine
Probabilistic Linkage	Probabilistic	+ Estimated probabilities +/- Character-level comparison	RecordLinkage fastLink

To allow more flexibility, instead of simple merge, a researcher might choose to calculate edit distances (Navarro, 2001) between all possible pairs and then choose some cutoff when a pair is considered to be a match. Edit distance is usually defined as some function that takes two strings as input and produces a scalar[4] that in its basic form counts the number of changes required to convert one string into another. One of the earliest and most elementary distance measures, Levenshtein distance (Levenshtein, 1965), simply counts insertions, deletions or substitutions of individual characters as possible edit operations for this transformation. This basic approach can be extended by incorporating other types of edit operations, such as transposition (Damerau, 1964) or accounting for the lengths of compared strings (Winkler, 1990). The two most important aspects of any record linkage procedures based exclusively on string distance measures are (1)

[4]Although in practice, string comparison function can take more than two strings and produce a vector or matrix of distances, for simplicity of exposition, I will focus my discussion here on the case of two input strings.

character-focused nature of comparisons made and (2) *arbitrariness* of cutoff threshold for matches. The focus on individual characters comprising a text string can be beneficial if the expected errors occur at this level. The examples could be texts that result from OCR and incorporate typos made by clerical error during data entry. However, the downside of this approach is that it completely ignores the structure of the label data. In cases, when alterations happen at the level of individual tokens (e.g. using "&" instead of "and" in names of geographic areas) it can result in a large number of false negatives. Another caveat of relying on string distance measures for merging data sets is the need to decide on a cutoff point when two records are considered to be a match. While this permits a higher degree of flexibility than a simple merge[5], this comes at a cost of the need to balance precision and recall. While setting the threshold too high can result in missing many true matches, making it too low, on the contrary, will result in many false positives.

A different way to consider short text labels is to think of them not as sequences of characters, but as collections of individual tokens that can change their position within the label. This approach in text-as-data literature (Grimmer and Stewart, 2013) is referred to as "bag-of-words". By disregarding the position of words in labels, the principle source of information on potential matches is the presence and the number of identical words in a candidate record pair. Some measure of similarity can then be applied to calculate the distance between the two representations of labels in the vector format. One of such measures, cosine similarity, between two labels i and j could be calculated as:

$$cos(\boldsymbol{l}_i, \boldsymbol{l}_j) = \frac{\sum_{w=1}^{W} l_{iw} l_{jw}}{\sqrt{\sum_{w=1}^{W} l_{iw}^2} \sqrt{\sum_{w=1}^{W} l_{jw}^2}}$$

where W is the total length of vector representations or, to put it differently, the number of unique words occurring in both labels. As opposed to string distance comparisons, text similarity measures offer a more flexible approach to deal with the *word-level*

[5]Effectively, simple merge can be considered a special case of string distance-based record linkage, when the allowed distance is taken to be 0.

variation in the labels. The same corporate entity, can be referred to as "Heathrow Airport", "Heathrow Airport Ltd" or, simply, "Heathrow". However, this approach does not solve the problem with arbitrary threshold, which has to be selected as a decision rule for separating candidate pairs into matches and non-matches.

While the statistical apparatus for probabilistic record linkage was developed some half a century ago (Fellegi and Sunter, 1969), accessible open-source implementations have appeared only recently (Sariyar and Borg, 2010; Enamorado et al., 2019). In a nutshell, probabilistic record linkage is a mixture model where the specific agreement pattern observed is a function of the underlying latent membership in the discrete class of true matches or true non-matches. The key quantity of interest can, thus, be defined as:

$$\theta_{ij} = \frac{\prod_{k=1}^{K} P(\gamma_{ij}|M_{ij}=1)}{\prod_{k=1}^{K} P(\gamma_{ij}|M_{ij}=0)}$$

where θ is the ratio between products of probabilities of observing values γ_{ij} in a k-component agreement pattern, given that labels i and j represent a genuine match: $M_{ij} = 1$. The crucial assumption that makes computation tractable is the independence between components k. In practice, the computation of this quantity of interest requires knowing whether i and j represent a true match, something which is rarely known. This can be learnt from a "gold standard" training dataset or using an EM algorithm (Winkler, 2000; Sariyar and Borg, 2010; Enamorado et al., 2019). While probabilistic record linkage allows for modelling the parameters of interest, instead of relying on arbitrary cutoffs, the need to have discrete components for comparison appears problematic for variable-length multi-token strings. Indeed, if two labels "Heathrow Airport" and "Heathrow" are segmented into individual tokens, then they would have different number of components. However, treating them as a single field for comparison necessitates the calculation of string distance between the two, which, in the presence of additional words, is likely to be overestimated.

4.4 Labels as Text

Political scientists successfully applied a range of computational approaches to textual data (Grimmer and Stewart, 2013) to estimate ideological positions (Laver et al., 2003; Slapin and Proksch, 2008; Diermeier et al., 2012), explain internal working of censorship apparatus (King et al., 2013), assess the impact of franchise extension on speech behavior in parliament (Spirling, 2015), explore the agenda-setting power of social media (Barberá et al., 2019) and predict the onset of political violence (Mueller and Rauh, 2018). All these applications use texts that can safely be assumed to contain natural language. Be it party manifestos, floor speeches in parliament, social media posts or newspaper articles, they were all written for the eventual perusing by other humans. Even when applying different 'bag-of-words' models, which scramble sentences and render them meaningless for a human reader, one can still be certain that the distributional properties of text remain intact. However, it is uncertain to what extent would a collection of short strings that contain names of different entities meet the same criteria. Before I proceed to assessing the performance of different methods, it is important to show that this type of textual data indeed behaves statistically similar to a more common source.

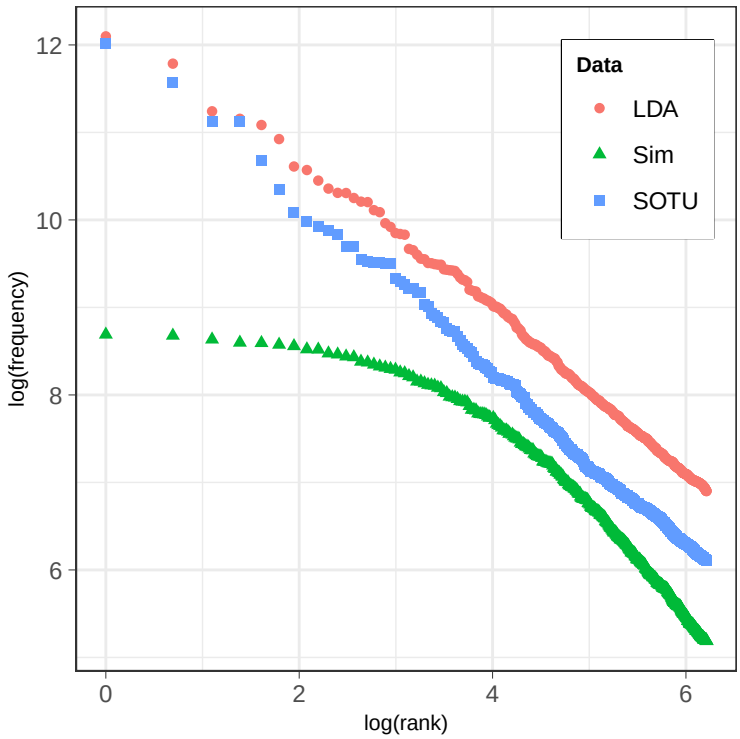
The universal statistical property of natural languages, most extensively documented by and named after Zipf (1935) provides one such test. It states that word frequency and its rank in the frequencies table have an inverse power relation. In other words, the second most frequent word will occur approximately half the number of the most frequent, the third half the number of the second and so on. More formally, the frequency of a word w with rank r can be expressed as:

$$F(w_r) = \frac{C}{r^{\alpha}} \tag{4.1}$$

where C is a normalization constant (for English language C is often taken as ≈ 0.1) and α is value of the exponent characterizing the distribution (in Zipf's original formulation $\alpha \approx 2$).

To make the distributions easier to compare in graphical form across different corpora,

Figure 4.1: **Log-log Token Plot.** Token frequency for 500 most frequent words plotted against token rank on a log-log scale for 3 data sources.



Note: Client names from data released under the Lobbying Disclosure Act and compiled by OpenSecrets (LDA), simulated dataset of 200,000 short labels (Sim), 239 State of the Union addresses in the US from 1790 to 2016 (SOTU) are used.

I plot the logarithm of frequency and rank instead of the raw numbers. Figure 4.1 shows Zipf's distribution for 3 different sources. The first one is a more canonical corpus of State of the Union addresses by the US president to Congress, extensively studied before (Rule et al., 2015; Benoit et al., 2019). The second source is more pertinent to the record linkage problem. Here the figure shows the frequency and rank of the 500 most frequent tokens used in names of organizations reported as lobbyists' client under the Lobbying Disclosure Act, often used in interest group studies (Ansolabehere et al., 2002; Kim, 2017; You, 2017). And the third one is an artificial dataset generated for the simulation study below, which was designed to exhibit all the properties of a typical collection of entity labels in text form without incorporating the specificities of any particular data set. As can be seen from the figure, the distributions for addresses and labels appear very much

alike with both almost following a straight diagonal. Although coarse, the comparison of Zipf's distributions shows that entity label data is not dissimilar to other textual corpora and its statistical properties should not pose a problem for the application methods for text analysis.

4.5 Case Study: Lobbying and PPP Loans

To illustrate the implications of choosing one record linkage approach over another it is helpful to consider some real-world datasets and approximate substantive analysis that a researcher might wish to carry out on them. One such dataset could be the already noted Lobbying Disclosure Act, which provides extensive details on lobbying activities in the US. Starting from 1996, all lobbyists have to file semi-annual reports, documenting all clients and income received from or expenditure on the in-house lobbying activities. This data has been used extensively in the lobbying literature (Ansolabehere et al., 2002; Kim, 2017; You, 2017) and at the same time it exists in unstandardized form, a not atypical situation in political science research. While in principle researchers can compile the disclosure reports themselves from the official government source[6] (Goldstein and You, 2017), it is not less common to use pre-existing compilations (You, 2017). In what follows I will be relying on the dataset compiled by the Center for Responsible Politics[7].

In the wake of the ongoing coronavirus (COVID-19) pandemic and the accompanying it economic crisis, the US Congress adopted a number of measures to alleviate the shock under the broad CARES (Coronavirus Aid, Relief, and Economic Security) Act. One of those measures was the Paycheck Protection Program (PPP), a $669-billion business loan program administered by the US Small Business administration. In essence the program allowed small business owners to apply for loans, backed by the government, to cover the payment of salaries to payroll employees. These loans can then be partially or fully forgiven, if a business owner retains the jobs. The public release of the data on companies

[6] https://lobbyingdisclosure.house.gov/
[7] https://www.opensecrets.org/

that received such loans since the start of the program has drawn criticism on the grounds that many lobbyists were among the businesses that received a stimulus package[8].

Given the controversy, surrounding the distribution of forgivable loans, a researcher might be interested in the relationships between corporate lobbying and getting a loan. Unfortunately, the most interesting question, whether an organization that was engaged in lobbying in the past few years was more likely to receive a loan than the one that did not, is impossible to test empirically. As the released data contains only successful borrowers and not all applicants, we cannot address this question directly. However, we can still ask a number of substantively interesting questions with this data. For example, one of the declared purposes of the PPP loans was job retention. Conceivably, one would expect the organizations that promised to keep more workplaces intact were more likely to be granted a loan. It is then possible to assess whether lobbying played any moderating role in this relationship. In other words, did organizations involved in lobbying managed to be among the receivers of loans. To conduct this analysis a researcher would need to see what organizations were listed in the PPP receivers data and compare them with those in the LDA data. As the number of entities listed in both datasets is in the tens of thousands, rendering manual matching infeasible, this requires some automated record linkage approach. Here I illustrate how the choice of a record linkage approach can have severe implications for inference and dramatically affect substantive conclusions drawn by a researcher.

To test the effects of record linkage approaches on inference, I downloaded the data released by the Small Business Administration[9] listing loan receivers across all US states and some territories (e.g. Guam and Puerto Rico), who got loans above $150,000. In total this dataset contains data on 661,218 organizations. For substantive and computational reasons, as well as to increase the chances of successful matches, I further restricted this dataset to those that received loans in excess of $1M, which reduced it to 82,708

[8]See, for instance, https://www.politico.com/newsletters/politico-influence/2020/07/07/which-lobbying-and-public-affairs-firms-got-ppp-loans-789005.

[9]https://www.sba.gov/funding-programs/loans/coronavirus-relief-options/paycheck-protection-program

organizations. Also, rather than trying to find these organization among all those that reported lobbying activity, I focus only on those that filed reports between 2017 and 2019. There are approximately 14,000 organizations that reported lobbying activity for this period[10].

[10]This number is approximate and not exact as the original datasets required deduplication, a subtask often encountered in record linkage. The precise number depends on how deduplication is done. In its raw form the dataset contains 13,931 unique strings for the given time period.

Table 4.2: **Analysis of Lobbying in PPP Loans.** Reported number of retained jobs is the dependent variable.

	Simple		Levenshtein Distance			Cosine Similarity		
	raw	cleaned	< 0.1	< 0.2	< 0.25	> 0.9	> 0.75	> 0.7
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lobbied (2017-19)	10.1698	10.1912	9.5699	5.2395	7.5744**	13.4242	3.1079	4.0893
	(9.7377)	(5.4342)	(8.1205)	(4.2899)	(2.853)	(9.5333)	(5.0294)	(3.6977)
Organization Type (Non-Profit)	45.0087***	44.9735***	44.9988***	44.9836***	44.815***	44.9651***	45.0441***	44.9298***
	(1.4612)	(1.4586)	(1.4607)	(1.4614)	(1.461)	(1.4611)	(1.4637)	(1.468)
Organization Type (Other)	0.228	0.2506	0.2211	0.2225	0.1977	0.2188	0.2367	0.2265
	(2.1905)	(2.1903)	(2.1905)	(2.1905)	(2.1904)	(2.1905)	(2.1905)	(2.1905)
State FE	✓	✓	✓	✓	✓	✓	✓	✓
Constant	137.1935***	137.1378***	137.2024***	137.2546***	137.0349***	137.1629***	137.1796***	137.1099***
	(9.3714)	(9.3711)	(9.3712)	(9.3709)	(9.371)	(9.3713)	(9.3727)	(9.3723)
Observations	77480	77480	77480	77480	77480	77480	77480	77480
R^2	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Note: OLS model estimates are shown. Reference category for organization type is business. Different datasets compiled using 3 record linkage approaches are used.

Table 4.3: **Analysis of Lobbying Expenditure in PPP Loans.** Reported number of retained jobs is the dependent variable.

	Simple		Levenshtein Distance			Cosine Similarity		
	raw	cleaned	< 0.1	< 0.2	< 0.25	> 0.9	> 0.75	> 0.7
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log(Lobbying Expenditure)	-2.6753	0.3985	-1.365	-2.4764*	-1.5021*	-1.4613	-1.4259	-0.7065
	(3.2759)	(1.5663)	(2.6208)	(1.0814)	(0.6688)	(3.1396)	(1.2504)	(0.8423)
Organization Type (Non-Profit)	31.1771	40.3861**	38.5782	57.1709***	61.7728***	19.715	33.8171**	40.3934***
	(30.5018)	(13.9689)	(21.8814)	(10.1629)	(7.1018)	(29.1699)	(11.2916)	(8.1442)
Organization Type (Other)	-11.5341	8.1971	23.1395	-4.7989	12.3039	-34.426	21.8922	12.9273
	(56.781)	(35.5595)	(43.7266)	(22.0556)	(14.5207)	(53.8893)	(26.5939)	(19.5993)
State FE	✓	✓	✓	✓	✓	✓	✓	✓
Constant	128.6863	131.5971	120.7588	127.4832	130.4711**	121.3411	115.455*	121.7887**
	(96.6657)	(77.4518)	(98.2976)	(92.5959)	(49.4061)	(96.7813)	(50.0778)	(43.7028)
Observations	162	521	233	843	1931	169	615	1152
R^2	0.3	0.13	0.25	0.13	0.1	0.29	0.11	0.09

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Note: OLS model estimates are shown. Reference category for organization type is business. Different subsets of the data compiled using 3 record linkage approaches are used.

The three record linkage approaches that I consider in this example are (1) simple merge, using raw and some basically cleaned data, (2) Levenshtein distance and (3) cosine similarity with multiple cut-offs. More specifically, in the case of cleaned simple merge, I remove all common words in corporate names (such as *llp*, *llc*, *corp*, etc.) and trim any extraneous whitespaces. In the case of Levenshtein distance I calculate the string distance between all pairs of names and then for each entry in the PPP dataset I choose the one from the LDA dataset containing the closest organization name. As absolute string distance can be sensitive to the length of the original string, I further normalize it by the number of characters in the names of loan receivers. Put differently, thus normalized Levenshtein distance of 0.1 indicates that 10% of characters in the original name have to undergo change to be converted into the name from the right-hand side dataset of lobbyists. i use three different cut-offs of 0.1, 0.2 and 0.25, naturally, resulting in progressively more successful matches. And, lastly, prior to calculating cosine similarity between all pairs of candidate names, I remove all punctuation marks. Here I use the cut-offs of 0.9, 0.75 and 0.7. The specific cut-offs in each case were selected to be both broadly equivalent[11] and provide a comparable number of successful matches.

After merging PPP dataset with the data on lobbying, I estimate two baseline OLS models, focussing on two principal explanatory variables: (1) *lobbied*, a binary indicator of whether the organization lobbied the government between 2017 and 2019 and (2) *log(lobbying expenditure)*, a continuous explanatory variable combining amount spend on both in-house and consultant lobbying activities. Tables 4.2 and 4.3 show the results of fitting these models on datasets created using different record linkage approaches. Comparing between the two models, it appears that record linkage had a larger impact on the second set of models. While relaxing some matching assumptions, such increasing the permissible normalized Levenshtein distance to 0.25 did result in the higher number of matches in the first set, which in turn made the coefficient for lobbying significant

[11]Cosine similarity of 0.9 indicates that word vectors representing the names of two organizations are very similar as does Levenstein distance of 0.1. Note the opposite direction of comparison, the lower the distance and the higher the similarity the closer are the matches.

($\hat{\beta}_{lobbied} = 7.57$, $p < 0.01$ in model 5), the sign and magnitude of coefficients is consistent across different record linkage approaches. This cannot be said about the relationship between lobbying expenditure and the number of retained jobs. Recall that our substantive hypothesis here is whether lobbyists were more successful in securing loans, while keeping fewer jobs. While the first set of models reported in Table 4.2 provide scant support for this, higher lobbying expenditure appears to be consistently associated with lower number of retained jobs, controlling for organizational type and location. However, in a more conservative approach of simple merging after some pre-cleaning we see the coefficient switching its sign, albeit it is not statistically significant. Overall, this relationship appears to be negative as per our hypothesis and consistent with the criticism of more political commentators.

Having shown with this short case study the importance and substantive implications that the choice of a record linkage approach can have for empirical analysis in political science research, the next logical step is to see what performance can be achieved in principle.

4.6 Simulation Study

A Basic Setup

The basic idea behind a simulation study on record linkage is to create artificial data sets the appear as similar as possible to real-world data, add several types of statistical noise that imitates different ways in which the information contained in one of them could be distorted and apply currently available methods to evaluate their performance against known ground truth about true matches. Correspondingly, I create simulated data sets that incorporate the features of real label data without having to rely on the particularities of any specific source. Thus, this simulation study, first, allows to assess the performance more precisely as the true matches are available and do not have to be generated by hand-coding the data. Second, it allows to isolate the effects of the different

types of noise and their combinations that could be present in real data. Third, it allows to test not only matching performance, but also computational efficiency and scalability of different approaches, as the size of data can be defined extraneously. Fourth, generating entirely artificial data makes the results more generalizable and potentially applicable to a wider range of domains than just organization names, that I use for evaluation in the following section. And, fifth, while some of the hyper-parameters used to generate the data are empirical quantities valid for English, the fact that the resultant tokens are not genuine English words, makes the results less linguistically restricted than a random draw from some pre-defined vocabulary.

B Data Generating Process

Modelling the data generating process for textual data is not unproblematic. Language is a complex socio-cultural phenomenon that displays a lot of variation across localities, classes and individual authors. Even short names of companies, geographical units or literary works can contain tropes that would be extremely hard or virtually impossible to identify automatically. For example, the company name 'YPlan' contains a pun which requires some phonetic knowledge that is not necessarily to a parsing system. Fortunately, the task for which the simulated text is used in this article is much simpler than the extraction of hidden meanings from labels.

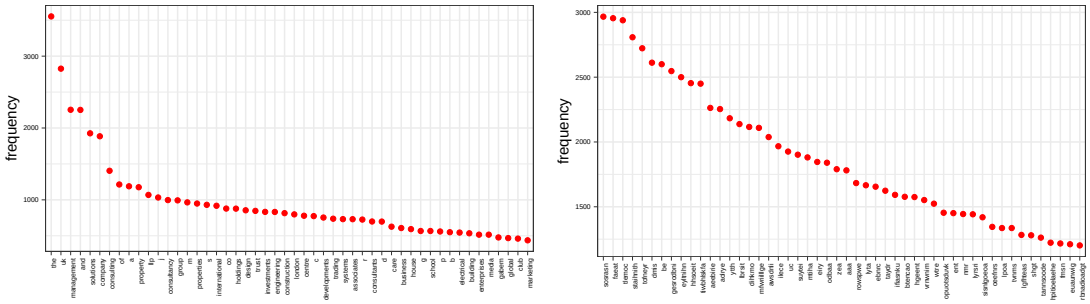
The simplest and often used approach to simulate textual data, common in computational linguistics (Li, 1992), is to define a set of characters (e.g. lowercase English letters), that includes whitespace, and generate a random text as a sequence of characters drawn from a multinomial distribution:

$$t \sim Multinomial(n, k, \boldsymbol{\pi}) \tag{4.2}$$

where n is the length of text t and k is the number of entries in the alphabet and

$(\pi_1, ..., \pi_k)$ are probabilities associated with each letter[12]. The whitespaces randomly occurring throughout the text are then treated as word boundaries. While being very intuitive and simple to implement, this approach, however, generates data that is statistically and qualitatively different from real data. Although the in-built inverse relationship between word length and word frequency accurately describes it in the limit, it often does not hold on real data of even relatively large sizes. As figure 4.2a shows, the lengths of most frequent words used in organization names vary considerably, from single-character articles 'a' to such words as 'international'.

Figure 4.2: **Frequency Distribution of Tokens.** 50 most frequent tokens are shown.



(a) Bureau van Dijk (b) Simulated

Note: A random sample of 100,000 organizations and 100,000 simulated labels are used. 3 most common tokens (*limited*, *ltd*, *services*) are excluded from the BvD figure.

In order to make the simulated data more realistic, I use the data-generating process that directly incorporates Zipf-Mandelbrot distribution. In addition to that I use some empirical regularities for English language, that make the tokes appear more natural and, oftentimes, not entirely implausible.

The data generating process is divided into two parts: vocabulary generation and text generation. In the first step I generate tokens as random iid draws from categorical distribution:

$$w_i \sim Multinomial(\mu_i, k, \boldsymbol{\pi}) \tag{4.3}$$

[12]As an extension of Zipf's law, Mandelbrot (1954) provided mathematical foundation for this distribution and showed that even without any human intention such text generating process would exhibit a very long right tail.

where a word i contains μ_i characters that come from alphabet of length k, each letter of which has $(\pi_1, ..., \pi_k)$ probability of occurring. The alphabet here is defined as all letters in English language, but without including the whitespace character. Although it is possible to set hyper-parameters to randomly generate $(\pi_1, .., \pi_k)$, to make the final result more akin to English-language labels, I use empirical relative frequencies of different letter in real words[13]. In turn, μ_i is generated from a Poisson distribution:

$$\mu_i \sim Poisson(\lambda) \tag{4.4}$$

where λ is taken to be 6, as the average length of words in English (Rothschild, 1986).

In the second part, after generating the vocabulary of a fixed pre-defined size, I simulate labels as a random sample of tokens from this population. To draw this sample I use Zipf-Mandelbrot distribution as follows:

$$t_j \sim Zipf - Mandelbrot(\tau_j, C, \alpha) \tag{4.5}$$

where the label j contains τ_j tokens. C and α are the scaling parameters that were described in the equation 4.1. Zipf-Mandelbrot distribution, which is sometimes described as the discrete version of the Pareto distribution (Adamic, 2011), provides a very useful approach to simulating textual data. The key advantage of using Zipf-Mandebrot distribution over discrete probability distributions, such as Poisson or negative binomial, is that, first, it incorporates power law relationship between frequency and rank, which characterises real-world texts. And, second, it models higher and more realistic type-to-token ratio across the generated labels, than even an overdispersed distribution such as negative binomial would allow without explicitly modelling its first moment as a function of other covariates. The length of labels is set to be between 1 and 5 tokens and is

[13]The letter frequencies were compiled by Peter Norvig and are available here: http://norvig.com/mayzner.html.

randomly drawn from the categorical distribution:

$$\tau_j \sim Categorical(l, \boldsymbol{\pi}) \tag{4.6}$$

with l, the maximum number of tokens being equal to 5 and $(\pi_1, ..., \pi_l)$ having uniform distribution over permissible label lengths.

Figure 4.2b plots the distribution of 100 most frequent tokens occurring in the simulated data set of 100'000 labels. While the curvature the line appears smoother than in real data on the left, it it s important to note the absence of 2 outliers in the simulated data. The log-log plot in figure 4.2 shows that apart from the dozen most frequently observed tokens, the distributions for all three data sets are closely aligned.

C Noise Simulation

After generating the population of text labels, each unique label is assigned an id[14]. I draw 2 random samples of 10,000 each from this population. The dataset A is kept intact, while the labels in the dataset B are distorted at a rate of 0.3[15]. Specifically, for each label j in dataset B a draw from a binomial distribution with a probability of corruption of 0.3 is used to decide whether the label remains unaltered. If not, given the pre-specified type of noise, one of the following scenarios can occur. For character-level distortion replacement, deletion or insertion of a number of characters[16] is chosen at random. In the case of word-level distortion one word selected at random is either removed or its position is changed. To simulate the combination of the two types of noise I, first, apply the procedure at the word-level and then the new label gets distorted at the character-

[14]The assumption of unique identifier based on labels is based on the expectation that in the population the names are genuinely unique. Some real-world deviations from this assumption should not be critical. However, if this assumption can be expected to be severely violated, such as with the personal names, the implications of this study might be less applicable.

[15]Introducing the noise into one and not both datasets provides an optimistic estimate of the performance of record linkage methods. This scenario, however, is not unrealistic as researchers often treat the names in one curated dataset as a golden standard and would like the other datasets to conform to this standard.

[16]The number of characters affected is picked at random and can be at most 3 or the maximum length of the label.

level. In total I create three versions of the distorted dataset B, each incorporating a different type of noise structure. While the text labels are affected, unique identifiers assigned at the earlier stage provide a way to systematically assess the performance of the record linkage approaches.

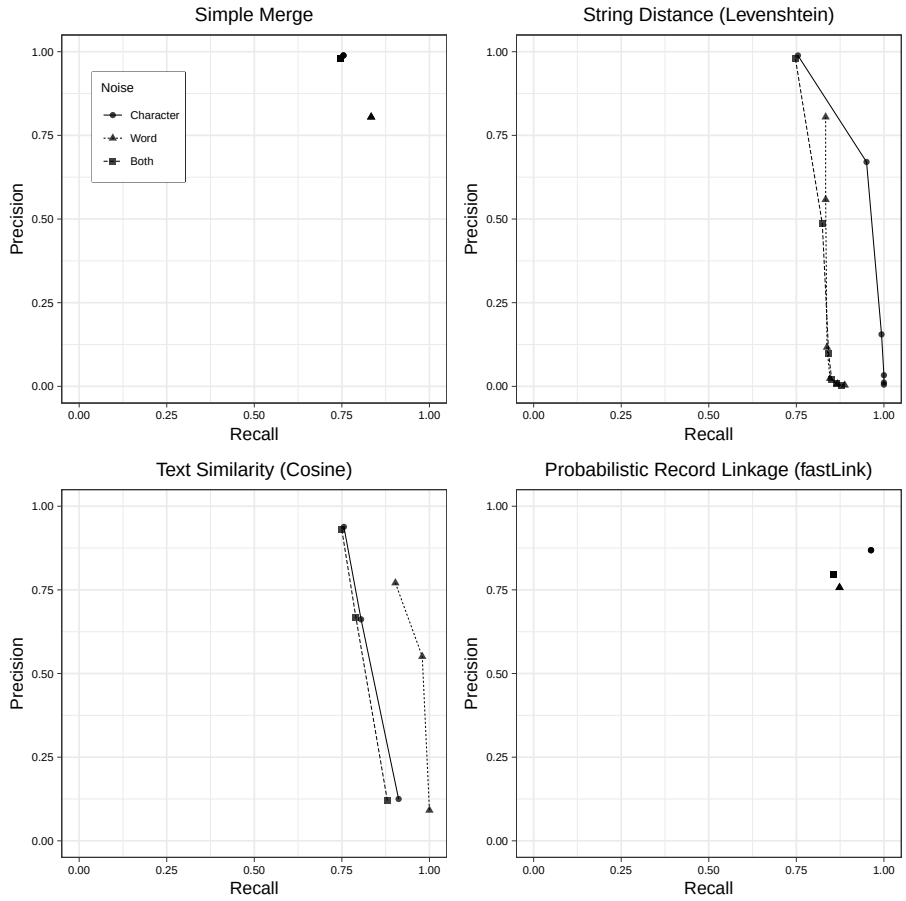
D Simulation Results

Figure 4.3 shows the main results of applying record linkage approaches on the simulated dataset. For methods which require the setting of arbitrary threshold for matches the sensitivity of these choices is shown with lines. Overall, the the two random samples contain 1,172 true matches (roughly 10% of each of the two datasets), textual labels that were originally identical across them. Restricting the total size of the datasets allows to limit the number of true non-matches, a metric which tends to inflate the estimates of record linkage performance. As matches often constitute a minority of all records in a given dataset, achieving high accuracy is easy in cases where data shows a large degree of variation and comparison pairs tend to be dissimilar.

Each panel shows the performance of record linkage approach according to two measures: precision and recall. As they do not include true negatives, these measures are more robust to class imbalance which is frequently encountered in matching problems. Recall that precision in this context is the number of true matches divided by the total number of matches suggested by record linkage approach: $\frac{True\ Matches}{True\ Matches + False\ Matches}$. Recall also captures the proportion of true matches, but out of all the matches present in the dataset: $\frac{True\ Matches}{True\ Matches + False\ Non\text{-}Matches}$.

Among all the noise types that can be encountered in textual labels, the one that includes a combination of character- and word-level corruption presents the most challenging problem for all record linkage approaches. The highest precision for this problem is achieved by simple merge (0.98), by only at the cost of losing more than 25% of true matches. The percentage of unmatched records is a function of the distortion rate, which was set at 0.3 for the the purposes of this simulation. In general, simply merging the two

Figure 4.3: **Performance Comparison of Record Linkage Approaches on Simulated Dataset.** Precision and recall varying by the type of noise introduced and different thresholds for match.



Note: The cutoffs used are 1-6 for Levenshtein, 0.9-0.7 for cosine similarity.

datasets is equivalent to matching all the uncorrupted records and dropping any records that contain any noise. However, in real-world settings the distortion rate can almost never be known directly. Thus, it is impossible to estimate either the proportion of the data that becomes missing due to failed matching or what kind of missingness that is[17]. In contrast to stable performance of strictly deterministic merge, the arbitrary choice of thresholds in string distance and text similarity measures illustrates the inherent trade-off between precision and recall. In the extreme, almost perfect recall could be achieved

[17]One can easily imagine a case, in which organizations with foreign-sounding names, get misspelled more frequently and, thus, are dropped out of the matched dataset at significantly higher rates than organizations that contain only common English words.

by lowering the cutoff, but this comes at a price of vastly inflating the number of false matches.

More importantly, the two approaches show the trade-off between adapting record linkage to dealing with character-level or word-level noise. While any string distance measure, be it Levenshtein (1965), Jaro (1995) or other[18], are adapted to matching records with character-level discrepancies, text similarity measures are better suited to tasks where entire words from the label get omitted or substituted. Probabilistic record linkage alleviates some of these problems, but having string distance at the core of creating agreement patterns, it also shows poor performance on word-level problems with only 0.76 precision and 0.87 recall.

Overall, the simulation study provides a useful baseline, language-agnostic framework for evaluating the performance of different record linkage approaches. While none of the methods is well-suited to tackle every type of problem, the trade-offs between precision and recall and character-level and word-level type of noise are important to bear in mind when designing record linkage step of the analysis.

4.7 Real Data Evaluation

It is possible that despite all the precautions, the simulated dataset differs substantially from any real-world data with labels. To check for this eventuality I conduct further performance evaluations using Bureau van Dijk dataset containing organization names in the UK. While in some sense this test is more restrictive as it is by construction limited to English language[19], it incorporates all the real-world-data caveats that might closer resemble the tasks faced by empirical researchers.

The setup of this experiment closely follows that of the simulation study above with

[18] An important exception to this is Monge and Elkan (1996) metric which includes a secondary function at the token-level.

[19] Despite the names of UK-registered organizations being largely English-based, it is worth noting that a small but, non-negligible, number of small businesses and local branches of global companies use English transliteration of foreign names derived from other languages.

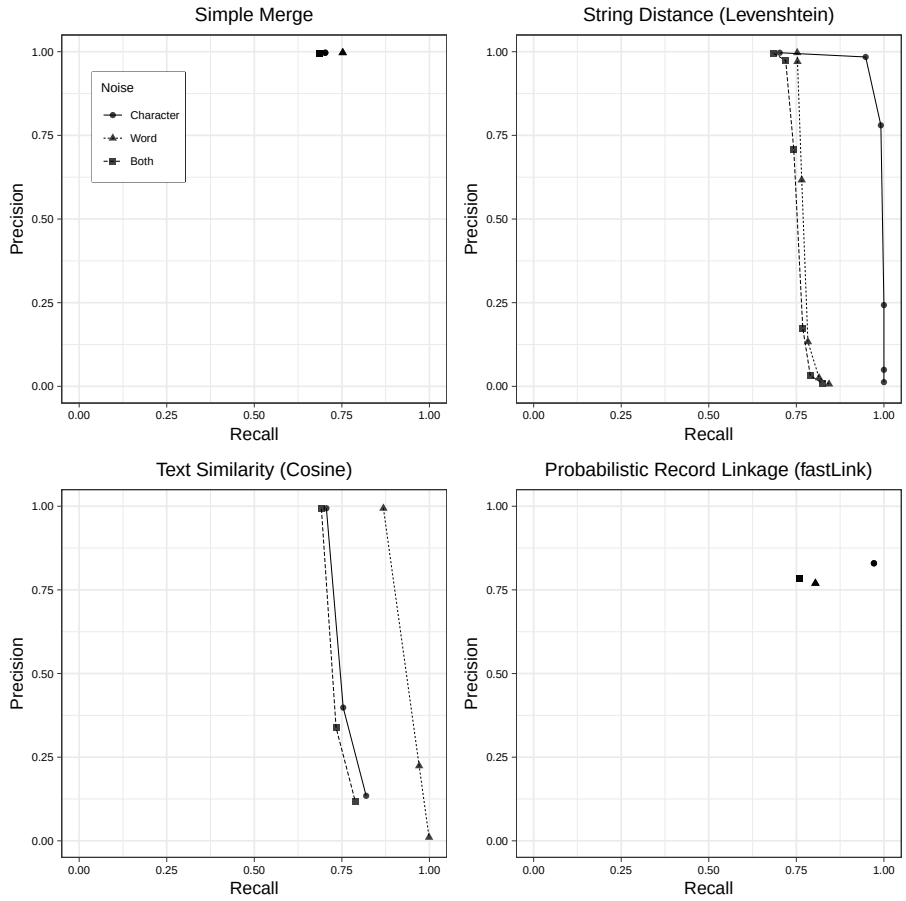
the important difference that instead of being generated, the text labels are drawn at random from the complete list of organization names. The entire labels list contains approximately 12 million records. For computational efficiency and to ensure the presence of non-negligible number of true matches, first, I scale this dataset down to 100,000 randomly drawn entries. It is important to note that in addition to using the original names, I also retain the original identifiers supplied by the data provider. While organizations are legally obliged to have distinct names, data-entry errors and higher-level aggregation of data from multiple sources do not guarantee the absence of identical names with different identifiers[20]. After scaling-down the original dataset, I draw two independent random samples of 10,000 records. By chance they contain 921 true matches, or, roughly, 10% of organization names are present in both datasets A and B. As with the simulated datasets, dataset A is retained in its original form[21], while the labels in the dataset B are distorted at the rate of 0.3.

Figure 4.4 presents the results of applying 4 record linkage approaches to matching uncorrupted records in the first dataset to partially distorted organization names in the second dataset. As before, three different noise types and the variation resulting from different choice of threshold for matching with string distance and text similarity are shown. First thing to note is that, overall, the results are largely consistent with the performance evaluation on the simulated dataset. The corrupted labels that contain both types of noise are the hardest to find a correct match. The string distance and text similarity measures exhibit the same trade-off between adjusting for character- and word-level noise. One of noticeable differences between the simulated and real data is higher ceiling for precision achieved by all of the approaches. The reason for that is the higher diversity of word types present in real data. Indeed, the 100,000 sample of BvD data includes 65,000 unique features, while the simulated dataset of the same size contains only 15,744. While this might appear as a poor simulation attempt, the variability of

[20]This provides a further extension to the simulation study above where the labels were treated as unique by design.

[21]The caveat of potentially over-estimating the performance of record linkage approaches due to distorting only one dataset applies here as well.

Figure 4.4: **Performance Comparison of Record Linkage Approaches on Dataset of Companies Names.** Precision and recall varying by the type of noise introduced and different thresholds for match.



Note: The cutoffs used are 1-6 for Levenshtein, 0.9-0.7 for cosine similarity.

word types in textual label data often surpasses that of usual corpora. For instance, SOTU corpus spanning over two centuries of language changes contains less than half of the number of word types than BvD (33,288). Thus, small changes in labels render the probability of falsely matching the records higher in more sparse datasets independent of the type of error introduced.

4.8 Conclusions

In this paper I review the four main approaches to addressing record linkage problems when information is limited to textual labels. A simulation study using artificially constructed data and an analogous study using real-world data illustrate the sensitivity of the results to the choice of an overall method as well as a specific threshold for matching. The results illustrate two important trade-offs: emphasis on precision as opposed to recall and adapting for character-level as opposed to word-level noise in the data. While the former is a more typical problem for automated methods of classification, the latter is a more problem-specific and could be explicitly acknowledged and accounted for in empirical research.

Although the distortion rate can rarely be known in real-world data analysis, the most likely type of noise is something that could be inferred from the details of data collection process or the circumstances of the original data generation. For example, manual input by humans or optical-character recognition are more prone to typos, misidentified letters or other character-level problems. On the contrary, discrepancies in data-collection standards across agencies, states or government departments can be a likely cause for encountering different ways of referring to the same entity, be it electoral ward or company name. In those cases using approaches that are more suitable for dealing with word-level noise can result in a matched dataset of a higher quality.

Bibliography

Adamic, L. A. (2011). Unzipping Zipf's law. *Nature 474*, 164–165.

Ansolabehere, S., J. M. Snyder, and M. Tripathi (2002). Are PAC Contributions and Lobbying Linked? New Evidence from the 1995 Lobby Disclosure Act. *Business and Politics 4*(2), 131–156.

Barberá, P., A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, and J. A. Tucker (2019). Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review 113*(4), 883–901.

Benoit, K., K. Munger, and A. Spirling (2019). Measuring and Explaining Political Sophistication Through Textual Complexity. *American Journal of Political Science 63*(2), 491–508.

Bonica, A. (2014). Mapping the Ideological Marketplace. *American Journal of Political Science 58*(2), 367–386.

Brady, H. E. (2019). The Challenge of Big Data and Data Science. *Annual Review of Political Science 22*(1), 297–323.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer.

Cohen, W. W., S. E. Fienberg, P. D. Ravikumar, and S. E. Fienberg (2003). A Comparison

of String Distance Metrics for Name-Matching Tasks. In *Proceedings of IJCAI-03 Workshop on Information Integration on the Web*, pp. 73–78.

Damerau, F. J. (1964). A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM 7*(3), 171–176.

Diermeier, D., J.-F. Godbout, B. Yu, and S. Kaufmann (2012). Language and Ideology in Congress. *British Journal of Political Science 42*(01), 31–55.

Donnay, K., E. T. Dunford, E. C. McGrath, D. Backer, and D. E. Cunningham (2019). Integrating Conflict Event Data. *Journal of Conflict Resolution 63*(5), 1337–1364.

Dunn, H. L. (1946). Record Linkage. *American Journal of Public Health 36*(12), 1412–1416.

Enamorado, T., B. Fifield, and K. Imai (2019). Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records. *American Political Science Review 113*(2), 353–371.

Enamorado, T. and K. Imai (2020, 01). Validating Self-Reported Turnout by Linking Public Opinion Surveys with Administrative Records. *Public Opinion Quarterly 83*(4), 723–748.

Fellegi, I. P. and A. B. Sunter (1969). A Theory for Record Linkage. *Journal of the American Statistical Association 64*(328), 1183–1210.

Goldstein, R. and H. Y. You (2017). Cities as Lobbyists. *American Journal of Political Science 61*(4), 864–876.

Grimmer, J. and B. M. Stewart (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*(3), 267–297.

Jaro, M. A. (1995). Probabilistic Linkage of Large Public Health Data Files. *Statistics in Medicine 14*(5-7), 491–498.

Jutte, D. P., L. L. Roos, and M. D. Brownell (2011). Administrative Record Linkage as a Tool for Public Health Research. *Annual Review of Public Health 32*(1), 91–108.

Kim, I. S. (2017). Political Cleavages within Industry: Firm-level Lobbying for Trade Liberalization. *American Political Science Review 111*(1), 1–20.

King, G., J. Pan, and M. Roberts (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review 107*(917), 326–343.

Laver, M., K. Benoit, and J. Garry (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review 97*(2), 311–331.

Levenshtein, V. I. (1965). Dvoichnyi kody s ispravleniem vypadenyi, vstavok i zameshenyi simvolov [Binary codes capable of correcting deletions, insertions and reversals of symbols]. *Doklady Akademii Nauk SSSR 163*(4), 845–848.

Li, W. (1992). Random Texts Exhibit Zipfs-Law-Like Word Frequency Distribution. *IEEE Transactions on Information Theory 38*(6), 1842–1845.

Mandelbrot, B. (1954). Structure Formelle des Textes et Communication [Formal Structure of Texts and Communication]. *Word 10*(1), 1–27.

Monge, A. E. and C. P. Elkan (1996). The field matching problem: Algorithms and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 267–270.

Mozer, R., L. Miratrix, A. R. Kaufman, and L. J. Anastasopoulos (2019). Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality. *Political Analysis* (Forthcoming).

Mueller, H. and C. Rauh (2018). Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review 112*(2), 358–375.

Navarro, G. (2001). A Guided Tour to Approximate String Matching. *ACM computing surveys (CSUR) 33*(1), 31–88.

Newcombe, H. B. and J. M. Kennedy (1962). Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information. *Communications of the ACM 5*(11), 563–566.

Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959). Automatic linkage of vital records. *Science 130*(3381), 954–959.

Roberts, M. E., B. M. Stewart, and R. A. Nielsen (2018). Adjusting for Confounding with Text Matching. *Working Paper*.

Rothschild, L. (1986). The Distribution of English Dictionary Word Lengths. *Journal of Statistical Planning and Inference 14*, 311–322.

Rule, A., J.-P. Cointet, and P. S. Bearman (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790-2014. *Proceedings of the National Academy of Sciences 112*(35), 10837–10844.

Sadinle, M. (2017). Bayesian Estimation of Bipartite Matchings for Record Linkage. *Journal of the American Statistical Association 112*(518), 600–612.

Sariyar, M. and A. Borg (2010). The RecordLinkage Package: Detecting Errors in Data. *R Journal 2*(2), 61–67.

Slapin, J. B. and S.-O. Proksch (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science 52*(3), 705–722.

Spirling, A. (2015). Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832-1915. *The Journal of Politics 78*(1), 235–248.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research, American Statistical Association*, pp. 354–359.

Winkler, W. E. (2000). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. Technical Report RR2000/05, Bureau of the Census Statistical Research Division.

You, H. Y. (2017). Ex Post Lobbying. *The Journal of Politics 79*(4), 1162–1176.

Zipf, G. K. (1935). *The Psychobiology of Language*. Houghton-Mifflin.

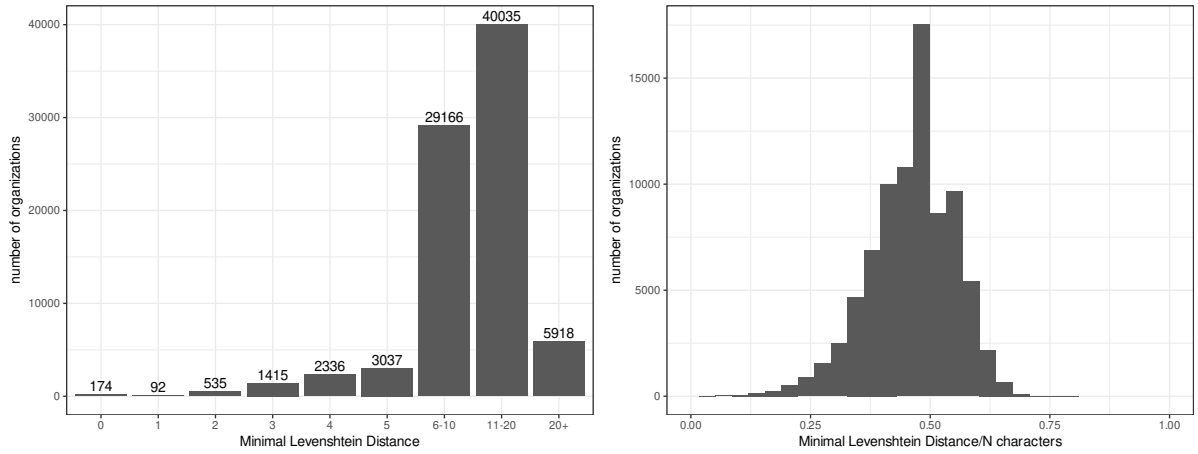
Appendix

A Case Study

Distribution of String Distances

Figure A.1 shows the distribution of Levenshtein distances between Paycheck Protection Program (PPP) loans data and released data under the Lobbying Disclosuer Act (LDA), compiled by OpenSecrets. As the comparisons of raw distances is sensitive to the length of the string, the right-hand Figure further shows the relative Levenstein distances, calculated by dividing the number of characters in the PPP data by the minimal Levenshtein distance to any record in the LDA data.

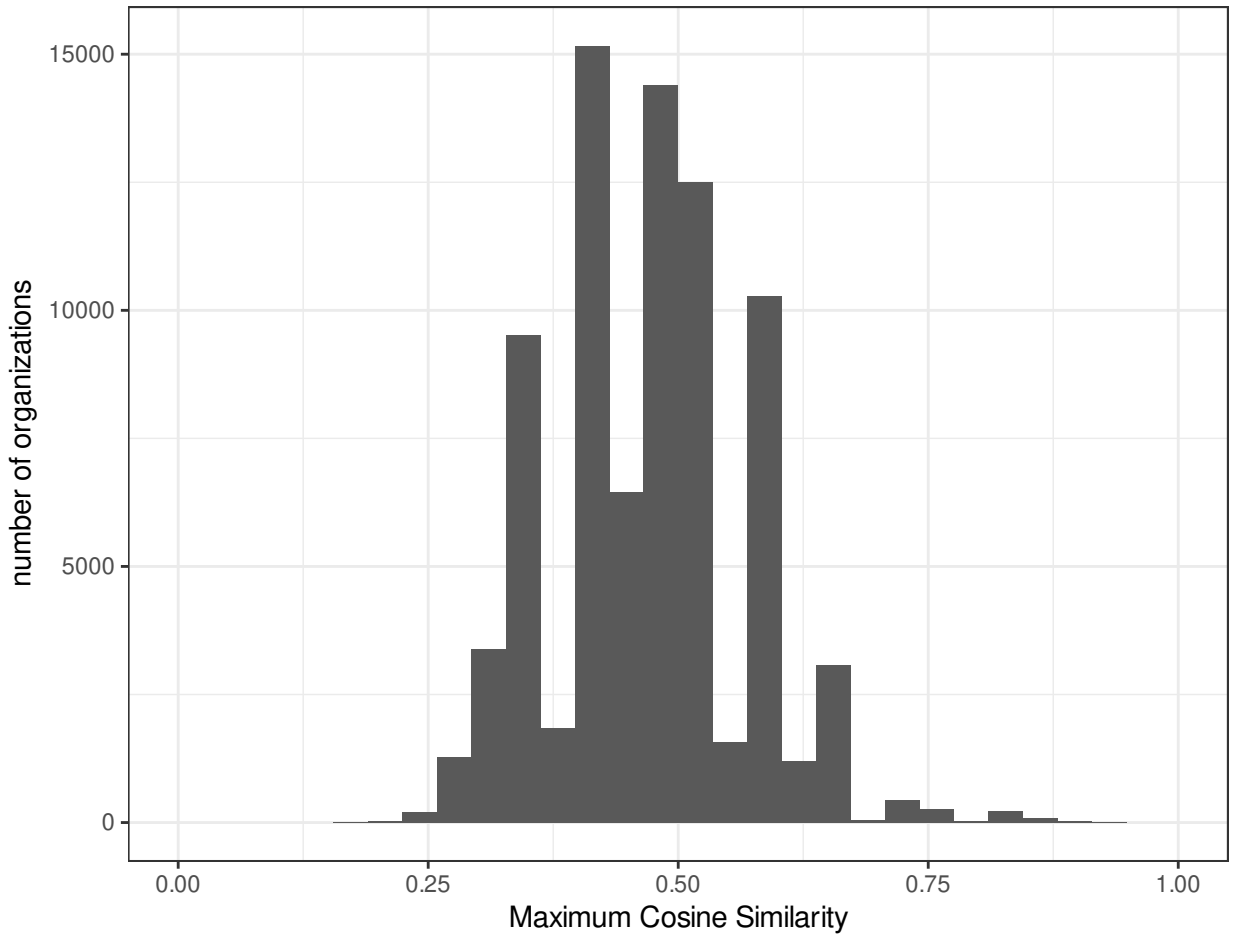
Figure A.1: Distribution of Levenshtein Distances when comparing PPP-LDA datasets



Distribution of Cosine Similarities

Figure A.2 illustrates the distribution of cosine similarities between Paycheck Protection Program (PPP) loans data and released data under the Lobbying Disclosuer Act (LDA), compiled by OpenSecrets.

Figure A.2: Distribution of Cosine Similarities when comparing PPP-LDA datasets



B Software Statement

The analysis was run under Linux Ubuntu 20.04 using R version 4.0.0 (R Core Team 2020). I relied on the following R packages in my empirical analysis:

`data.table` (Dowle and Srinivasan 2019),

`dplyr` (Wickham et al. 2018),

`ggplot2` (Wickham 2016),

`kableExtra` (Zhu 2018),

`knitr` (Xie 2018),

`lubridate` (Grolemund and Wickham 2011),

`magrittr` (Bache and Wickham 2014),

`readr` (Wickham, Hester, and Francois 2017),

`stringdist` (Van der Loo 2014),

`stringi` (Gagolewski 2018),

`stringr` (Wickham 2018),

`tibble` (Müller and Wickham 2018),

`tidyr` (Wickham and Henry 2018),

`quanteda` (Benoit et al. 2018), and

`zipfR` (Evert and Baroni 2007).