Interest Groups & Advocacy
https://doi.org/10.1057/s41309-021-00121-4

ORIGINAL ARTICLE

Computational approaches to mapping interest group representation: a test and discussion of different methods

Ellis Aizenberg[1] · **Anne Skorkjær Binderkrantz**[2]

Accepted: 13 April 2021

Abstract

Studying patterns of interest representation in politics is a central concern of scholars working on interest groups and lobbying. However, systematic empirical analysis of interest group representation entails a large amount of coding and is potentially prone to error. This letter addresses the potential of two computational methods in enabling large-scale analyses of interest group representation. We discuss the trade-offs associated with each method and empirically compare a manual, a query-based, and an off-the-shelf supervised machine learning approach to identify interest groups in a sample of 3000 news stories. Our results demonstrate the potential of automated methods, especially when used in combination.

Keywords Interest group representation · Media access · Computational methods

Introduction[1]

Democracies function best when different ideas are voiced in public and political debates (Dahl 1998; Danielian and Page 1994). Assessing diversity and bias in the interest system is therefore of major concern to scholars working on interest groups and lobbying (Lowery et al. 2015; Schattschneider 1975; Schlozman et al. 2012). In effect, extensive resources have been devoted to map the representation of organized

[1] This paper has benefited from the comments from several colleagues including Moritz Müller, members of the Comparative Politics section at Aarhus University and the participants in the panel on: "Methodical Innovation in the Study of Interest Groups" at the ECPR General Conference 2019. The authors would also like to thank the student assistants involved in the coding process: Max Scheijen, Anne Sofie Siggaard and Kathrine Skjoldborg.

✉ Ellis Aizenberg
E.Aizenberg@uva.nl

Anne Skorkjær Binderkrantz
asb@ps.au.dk

[1] University of Amsterdam, Amsterdam, Netherlands

[2] Aarhus University, Aarhus, Denmark

Published online: 20 May 2021

interests in different political arenas such as the news media, parliament, or administrative boards and committees (for examples, see Berkhout et al. 2018; Binderkrantz et al. 2015, 2017; De Bruycker and Beyers 2015; Schlozman et al. 2012; Thrall 2006). These studies rely on substantial manual coding, and even the most ambitious research projects face limitations regarding, for example, the range of different interest groups included, the time span covered, or the number of countries compared.

This letter addresses the potential of computer-assisted methods to identify organized interests in politics. In recent years, computational methods have increasingly been used to identify and classify political actors such as interest groups and political parties in different sources (for examples, see Aizenberg and Hanegraaff 2020; Aizenberg and Müller 2020; Fraussen et al. 2018; Garlick and Cluverius 2020). This is potentially a less costly and more reliable alternative to manual coding. It may also open up new avenues for addressing research questions that require large-scale comparisons over time or across different political systems. However, computational methods also come with their challenges (see Bunea et al. 2017; Grimmer and Stewart 2013; Klüver 2015 for overview and discussions) and may not be a viable alternative to manual coding if relevant actors are not reliably identified. A crucial question is therefore: *What is the potential of computational approaches in large-scale mapping of interest group representation?*

To address this question, we compare the use of manual coding to two prominent alternatives relying on computational methods: (1) the usage of search queries where a list of group names is used as the point of departure, and (2) an off-the-shelf generic supervised machine learning approach (named entity recognition) where no prior list of groups is required. We discuss the trade-offs associated with different approaches and empirically compare their ability to identify interest groups in a dataset of 3000 news stories from a UK newspaper. We find that our manual coding results in the highest number of correctly identified group appearances. Yet, it is notable that a combination of a query-based method and a named entity recognition (NER) approach is as effective as a manual approach in identifying group appearances. Moreover, a comparison of substantive results based on each of the three methods shows relatively similar findings.

While this letter focuses on mapping organized interest in the news media, the results are relevant for all scholars interested in the representation of political actors. For interest group scholars, the use of NER is particularly promising in settings where no prior register of relevant groups is available. This approach may also be helpful for those interested in the representation of, for example, business firms, individual citizens, or loose social movements. When scholars are interested in the representation of a predefined set of actors—political parties, members of parliament, or those appearing in a group register—automated methods involving search queries may be particularly helpful in answering research questions that require large-scale mapping of actor appearances in different sources.

The pros and cons of manual and computational mapping

A first challenge in empirical studies of interest group representation is mapping the group population in a given political system. In democratic societies, interest groups are numerous and the interest group population fluctuates as new groups mobilize and others cease to exist. The task of identifying the relevant entities differs markedly across different political systems depending on the pre-availability and quality of lists of organized interests (Berkhout et al. 2018; Braun 2012; Grant et al. 2012). A second challenge concerns identifying groups correctly in the selected sources. This section discusses these two challenges, with Table 1 summing up the discussion. For each step, the traditional manual approach is compared to two main computer-assisted alternatives—an approach relying on query-based searches for groups and a NER approach where groups are identified through a pretrained algorithm in the sources at hand.

An advantage of manual coding is that the list of relevant groups can be established or verified as part of the coding process. For manual coders, finding groups in the selected sources is the most time demanding part of the research process as the number of relevant sources will typically be high. Even careful coders may not always identify all relevant groups in the sources. On the other hand, the advantage of the manual approach is that coders will usually be able to find groups despite different spellings, the use of abbreviations, or inaccuracies in group names.

In contrast, for automated methods relying on identifying groups through searches in relevant sources, access to a list of relevant groups is paramount. Identification of groups through queries is also highly dependent on the quality of the list of groups used as a point of departure. Given the existence of an encompassing list, the method is highly reliable in finding a group appearing under the names specified on the list. It will, on the other hand, be more difficult to find groups where names differ markedly or where abbreviations are used. Also, the search may register instances where the name of a group appears in the text by coincidence—for example, when group names are similar to phrases commonly used in texts.

Finally, when employing NER, it is not necessary to use a predefined list since entities are identified from texts through an algorithm that is pretrained. The technique identifies references to entities such as locations, persons, numeric expressions such as time and date, and referral to organizations (Nadeau and Sekine 2007). It recognizes full names, abbreviations, and sometimes even identifies the correct organization in case of a typo. A downside of using the technique in this context is that most NER algorithms are supervised methods. This is an area of considerable innovation, and several specific implementations are available. We relied on a generic off-the-shelf system that not only identifies organized interests but also other types of organizations such as music bands and non-organizational entities such as persons. One can automatically filter out the organizations, but in this instance, a human coder has to decide whether it fits with the category that the researcher intends to measure. A generic system was chosen as such a tool is accessible for social scientists without much prior knowledge about computational methods.

Table 1 Comparison of three approaches to coding group representation

	Pros and cons in identifying relevant groups	Pros and cons in identifying groups in sources
Manual coding	Can be based on prior sources or part of the coding process	High validity due to human reading of texts
	High validity since human coders evaluate which groups to include	Issues of reliability because coders may miss group appearances
	Issues of reliability because group inclusion is a coding decision	Highly time-consuming
Query-based	List of relevant groups is a necessary prerequisite	High reliability in linking groups from list to groups in sources
	Possible concerns about the inclusiveness of group list	Difficulties in identifying groups appearing under different names and abbreviations than defined in list
		Less time-consuming compared to manual coding
Named entity recognition	Entities are inductively extracted from text	High reliability in extracting appearances of entities from text
	Validity has to be verified—task of the researcher after extracting entities	Less time-consuming compared to manual coding but more time-consuming compared to query-based method due to manual validation of results

Empirical comparison of approaches to identifying groups

This section presents an empirical comparison of the three approaches to identifying organized interests in politics. We focus on media appearances because this arena is likely to be more challenging to map than, for example, group representation in parliamentary hearings or government consultations. First, in news stories, groups may appear in different contexts. Second, interest groups appear alongside many other actors such as corporations or government entities, posing challenges of distinguishing between these actors. Third, there is a high likelihood that reporters will sometimes misspell a group name, use different names for the same group, or refer to groups with their abbreviation rather than their full name. These aspects are likely to complicate the identification of groups by automated methods.

We randomly selected 3000 articles from a dataset of all articles published in the UK newspaper The Guardian from 1 July 2017 to 30 June 2018. In the last years, the Brexit debate has been a major focal point in the British newspapers. To minimize this effect, we chose a period well after the Brexit referendum in June 2016 but also before the original deadline for finalizing negotiations in March 2019. This article set was saved in a NoSQL database. Prior to the random selection of articles, we excluded articles that were registered as concerning sports, culture, debate, or related to foreign affairs.

Description of three approaches to identifying groups

Our aim is to map the presence of interest groups in these articles. Consistent with most of the literature, we define interest groups as associations of members or other types of supporters that work to obtain political influence but do not seek political election. The coding does not include interest groups organized at the subnational level, at the international level, or from foreign countries. For groups that operate at the national as well as the international level (e.g. Greenpeace or Amnesty International), these are included if the article addresses UK political issues.

Our first strategy to identify groups is manual and similar to what has been done in previous research (Binderkrantz 2012; Binderkrantz et al. 2017; Danielian and Page 1994; Dimitrova and Strömbäck 2009; Tiffen et al. 2013). For this, coders were instructed to read the full text of all articles to identify interest group appearances. For articles with groups, coders registered the number and names of groups appearing. Each group is, thus, registered once for every article in which it appeared. The two student coders were trained on a separate dataset of 200 randomly selected articles. Hereafter, each of the 3000 articles was coded by one of the coders. Finally, 200 random articles were selected for a reliability coding. The Krippendorff's Alpha for this coding was 0.98 ($N = 200$), which is very high. The manual coding task took about 200 h.

Our second approach to identify groups within the newspaper articles was query-based—that is we use a query to search for a list of groups (Aizenberg and Müller 2020; Fraussen et al. 2018). The list of interest groups produced with the manual coding endeavor was used as input for the query. While it is, of course, an unlikely

scenario to have prior access to a list of groups appearing in the sources of interest, this allows us to test the extent of challenges that may arise even in this situation. Also, it serves as a test of whether the manual coders missed groups that appeared in several articles. With the help of the *amcatr* package in R, it is possible to communicate through an API with the NoSQL database AmCAT in which the documents of interests were stored (Van Atteveldt 2008). The query took 5 min to run. After the query, we removed more than 250 instances where group names corresponded to common terms such as 'agenda', 'scope', 'become', or 'motivation'. This demonstrates a first challenge in the use of automated searches, namely that some groups have chosen just a single word as their name. Setting up the query, running it, and filtering took about 10 h. The initial filtering of the results was carried out by a student assistant.

The third approach to identify organized interests in the dataset is a natural language processing technique called named entity recognition. For this paper, the analyzeEntities[2] method was employed, which inspects a text corpus for entities and subsequently provides information about these entities. An entity is a phrase in a text that can be called a 'known entity', such as a location, an event, or an organization. The API used associates' general information such as a link to the Wikipedia page and the saliency of a mention. The employed algorithm concerns a Google function that is implemented via the Google R package, which allows one to run this function on a Google server and retrieve results to your own computer for further analysis. The algorithm identified 74,985 entities ranging from people, locations, events, and organizations in both 'proper' and 'common' form. The former means that the actual name of a known entity is included in the article, while the latter means that there is a referral to this entity without the use of the actual name. For example, when an NGO is identified, 'Unicef UK' is the proper form and the common form is returned as 'the charity' or 'interest group'. When filtering for organizations in proper mentions, the search resulted in 9890 mentions. Subsequently, student assistants went through the list and excluded all actors that did not correspond with our definition of interest groups. This task took about 30 h.

Comparison of results

After conducting the three searches, we merged all results and matched group appearances. We then checked all discrepancies to determine whether the appearances were correctly identified or not. The results can be seen as three different samples of the true set of groups appearing in the 3000 news stories. In Table 2, we therefore compare the groups identified by each method to our best approximation of the true population of appearances: the total set of groups identified across all three methods.

[2] See for documentation and source: https://cloud.google.com/natural-language/docs/analyzing-entities.

Table 2 Correctly vs. incorrectly identified interest group appearances

	Correctly identified appearances	Share correct	Incorrectly identified appearances	Share incorrect[a]	Correctly identified unique groups	Share correct	F1 score[b]
Manual approach	578	0.86	0	0.00	357	0.91	0.92
Query-based approach	485	0.72	35	0.07	274	0.70	0.81
Named entity recognition	351	0.52	10	0.03	210	0.54	0.68
Across all methods	***674***	**= 100**	***44***		***392***	**= 100**	

[a] 'Share incorrect' is based on the number of correctly identified appearances with the same method

[b] F1 scores are calculated as the harmonic mean of precision and recall = 2 * (precision*recall)/(precision + recall) where precision is true positives/(true positives + false positives), and recall is true positives/(true positives + false negatives)

The manual approach was clearly superior to the two other when it comes to identifying the correct group appearances. The student coders found 91 per cent of all groups and 86 per cent of all appearances identified, and there were no incorrectly identified appearances. Still, this is surprisingly low in light of the very high intercoder reliability. While it is hard to ascertain why human coders miss groups, many of the groups missed were mentioned peripherally in the news stories.

The query-based search found 70 per cent of all groups and 72 per cent of all appearances. Some groups were missed because they were not included in the query as they were identified by the NER approach only. This illustrates an obvious limitation of a query-based method: it heavily depends on the list of groups used as point of departure. In addition, in about 50 cases, the group appeared under a slightly different name (often a shorter version of the group name) in the news story than on the list of groups used in the search. In about 30 cases, the reporter used an abbreviation of the group name that was not included in the search. Spelling errors caused other missed appearances, and yet others were related to punctuation in the news stories (e.g. missing spacing between words). In future endeavors, such errors may be reduced by systematically including abbreviations and short versions of group names. It may also be possible to allow for small deviations in names, but this comes at the risk of including more false positives in search results. Most of the 35 group appearances incorrectly identified by the query-based approach were instances where a group name was identified but referred to something else. This was the case for appearances such as 'abortion rights', 'positive action', 'vote leave', and 'powerful women'. A few additional cases were groups that operate under similar names in a domestic and international context such as Friends of the Earth.

With the employed off-the-shelf NER approach, we were able to find 54 per cent of all groups and 52 per cent of all group appearances. The algorithm mostly struggled to identify names consisting of multiple words. A reason for this could be that the training data contained mostly entities with shorter names. Another explanation might be that longer names are rarer in general. The algorithm might, therefore, attach lower probabilities to multi-word expressions because these then tend not to be entities more often. It is encouraging that more than half of all relevant hits were found by a generic supervised machine learning method that is trained to detect entities within general data such as Wikipedia, for example. An interesting outline for future research would be to apply a classifier pretrained on data such as political news that is closer to our material of interest or to use/build a hand-trained algorithm that detects organized interests specifically. It is also encouraging that a very low number of 'false positives' were present in the final list of group appearances based on this NER approach. After manually filtering the results of the NER approach, there were 10 irrelevant appearances in the dataset. These were mainly international groups sharing their name with UK groups.

Figure 1 illustrates the overlap in appearances found by each method. While the NER approach had the highest share of missed appearances, it is interesting to note that it also identified appearances found by neither of the other two methods. There is much overlap in the appearances found by the manual and query-based search, but it is evident that a combination of the two methods will be more effective than using one of them. Likewise, it is notable that a combination of the two automated

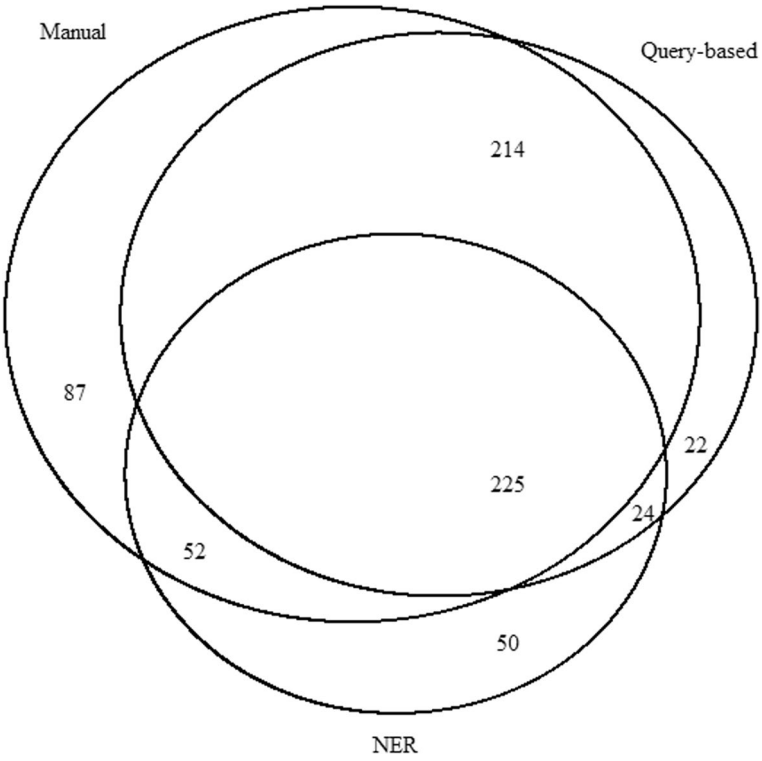

Fig. 1 Venn diagram of overlap in appearances identified ($N = 674$)

methods will significantly reduce the number of missed appearances by using only one method—in fact, only 13 per cent of group appearances were missed by both these methods.

For researchers, the ultimate choice of method depends on the ability to arrive at accurate conclusions with respect to the research question asked. Here, we compare one substantive aspect of the findings based on each method: the level of diversity in interest group media appearances. To gauge this, scholars often rely on Shannon's H as a measure of diversity, and Fig. 2 therefore displays the normalized version of Shannon's H for each of our 'samples'. In Panel A, this is calculated based on the distribution of attention across all groups in each sample. In Panel B, we have aggregated appearances for seven distinct group types commonly distinguished in the literature.[3] As can be seen from the figure, the conclusions with respect to the level of diversity in interest group media appearances are remarkably similar, and all confidence intervals overlap. While this is a highly aggregate level of analysis, it

[3] The coding of group types is based on the INTERARENA coding scheme (see Binderkrantz et al. 2020).

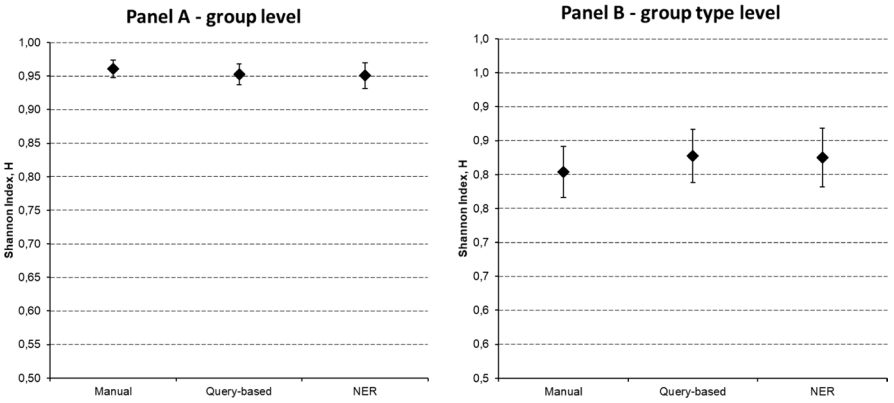

Fig. 2 Comparison of diversity across groups and group types (Shannon's H, normalized with 95% confidence intervals)

indicates that at least for some research questions, each of the three methods used to identify groups will result in reliable findings.

Conclusion

Firm conclusions about the nature of interest group representation in politics require large-scale empirical mappings of groups as they are represented in the news media, participate in public consultations, or get access to parliamentary hearings. To help researchers in these endeavors, automated methods seem promising. Such methods can, however, come with important pitfalls that may impede the results arrived at. This letter has provided a systematic comparison of manual and automated methods in order to identify the promises and challenges of automated methods in mapping interest group representation.

The first available automated method is a query-based search for groups in relevant text material. Here, we simply took the list of groups generated by our manual coding as the point of departure. This allowed us to zero in on issues associated with the query itself rather than the more trivial (but obviously important) fact that queries are only as good as the list used as a point of departure. The comparison of the manual and query-based method showed that the lion's share of groups identified in manual coding were also found in the query. However, a surprisingly high number of appearances were missed by the query because reporters used abbreviations, parts of group names, or simply misspelled the names of groups. With the second automated method—an off-the-shelf named entity recognition (NER) approach—it is possible to identify political actors in documents even in the absence of any prior list of groups. This method was able to identify more than half of the total number of appearances. It is also interesting to note that a combination of queries and entity recognition was just as efficient as manual coding.

Overall, there is reason for optimism with respect to the potential of computational methods in this context. While there are pitfalls involved, it is possible to identify strategies to remedy most of these. It should, however, be noted that even the use of automated methods rely on a significant amount of manual work. For example, prior to using a search-based method, it is necessary to make decisions on the inclusion of abbreviations and short names for groups. Importantly, the time needed to prepare the queries and post-coding of results will be subject to increasing returns to scale—for example, ten times as many articles will not require anywhere near ten times as much work in pre-coding and post-coding. This stands in contrast to a manual approach where little traction is gained when more empirical material is included in the coding. The automated methods are, therefore, particularly promising when researchers ask bold questions that require large-scale empirical studies over time, across countries, or the inclusion of various empirical sources.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aizenberg, E., and M. Hanegraaff. 2020. Time is of the essence: A longitudinal study on business presence in political news in the United Kingdom and the Netherlands. *The International Journal of Press/Politics* 25: 281–300.

Aizenberg, E., and M. Müller. 2020. Signaling expertise through the media? Measuring the appearance of corporations in political news through a complexity lens. *Journal of European Public Policy*. https://doi.org/10.1080/13501763.2020.1797144.

Berkhout, J., J. Beyers, C. Braun, M. Hanegraaff, and D. Lowery. 2018. Making inference across mobilisation and influence research: Comparing top-down and bottom-up mapping of interest systems. *Political Studies* 66: 43–62.

Binderkrantz, A.S. 2012. Interest groups in the media. bias and diversity over time. *European Journal of Political Research* 51: 117–139.

Binderkrantz, A.S., L. Chaqués-Bonafont, and D. Halpin. 2017. Diversity in the news? A study of interest groups in the media in the UK, Spain and Denmark. *British Journal of Political Science* 47: 313–328.

Binderkrantz, A.S., P.M. Christiansen, and H.H. Pedersen. 2015. Interest group access to the administration, parliament and media. *Governance: An International Journal of Policy Administration, and Institutions* 28: 95–112.

Binderkrantz, A.S., P.M. Christiansen, and H.H. Pedersen. 2020. Mapping interest group access to politics: A presentation of the INTERARENA research project. *Interest Groups & Advocacy* 9: 290–301.

Braun, C. 2012. The captive or the broker? Explaining public agency-interest group interactions. *Governance: An International Journal of Policy Administration, and Institutions* 25: 291–314.

Bunea, A., R. Ibenskas, and A.S. Binderkrantz. 2017. Estimating interest groups' policy positions through content analysis: a discussion of automated and human-coding text analysis techniques applied to studies of EU lobbying. *European Political Science* 16: 337–353.

Dahl, R.A. 1998. *On democracy*. New Haven: Yale University Press.

Danielian, L.H., and B.I. Page. 1994. The heavenly chorus: Interest group voices on TV news. *American Journal of Political Science* 38: 1056–1078.

De Bruycker, I., and J. Beyers. 2015. Balanced or biased? Interest groups and legislative lobbying in the European news media. *Political Communication* 32: 453–474.

Dimitrova, D.V., and J. Strömbäck. 2009. Look who's talking. Use of sources in newspaper coverage in Sweden and the United States. *Journalism Practice* 3: 75–91.

Fraussen, B., T. Graham, and D.R. Halpin. 2018. Assessing the prominence of interest groups in parliament: A supervised machine learning approach. *The Journal of Legislative Studies* 24: 450–474.

Garlick, A., and J. Cluverius. 2020. Automated estimates of state interest group lobbying populations. *Interest Groups & Advocacy* 9: 396–409.

Grant, J., F.R. Baumgartner, J.D. McCarthy, S. Bevan, and J. Greenan. 2012. Tracking interest group populations in the US and the UK. In *The scale of interest organization in democratic politics. Data and research methods*, ed. D. Halpin and G. Jordan, 141–160. Chippenham and Eastbourne: Palgrave Macmillan.

Grimmer, J., and B. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21: 267–297.

Klüver, H. 2015. The promises of quantitative text analysis in interest group research: A reply to Bunea and Ibenskas. *European Union Politics* 16: 456–466.

Lowery, D., F.R. Baumgartner, J. Berkhout, J.M. Berry, D. Halpin, M. Hojnacki, H. Klüver, B. Kohler-Koch, J. Richardson, and K.L. Schlozman. 2015. Images of an unbiased interest system. *Journal of European Public Policy* 22: 1212–1231.

Nadeau, D., and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30: 3–26.

Schattschneider, E.E. 1975. *The semisovereign people*. Thomson Learning, US: A Realist's View of Democracy in America.

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.

Thrall, T.A. 2006. The myth of the outside strategy: Mass media news coverage of interest groups. *Political Communication* 23: 407–420.

Tiffen, R., P.K. Jones, D. Rowe, T. Aalberg, S. Coen, J. Curran, et al. 2013. Sources in the news. *Journalism Studies* 15: 374–391.

Van Atteveldt, W.H. 2008. *Semantic network analysis: Techniques for extracting, representing, and querying media content*. Charleston, SC: BookSurge Publishers.