Semantic Network Analysis of Contested Political Concepts

Paul Nulty
Centre for Research in Arts, Social Science and Humanities,
University of Cambridge
`pgn26@cam.ac.uk`

Abstract

This work presents methods for exploring the lexical environment of political concepts using interactive network visualisations of corpus-derived grammatical relations and word associations. The conceptual relations consist of part-of-speech tagged words connected by typed, weighted, edges indicating the strength of relations between words, as measured by pointwise mutual information of different types of co-occurrences. An interactive animated interface allows users to adjust the node degree directly, or to specify edge-weight thresholds, and observe the resulting effect on the network. The system can be searched by neighbourhood sub-graphs ('ego graphs') of particular sets of query terms. The force-directed layout of the network highlights conceptual structure, as terms connected by many relations are drawn together, and the user can select which subsets of relations and sub-corpora to display. Community detection (cliques) and centrality measures provide additional comparative measures of the use of contested concepts in diverse political communities. As an example of such a system for exploring the structure of political concepts, an implementation on comments from libertarian and socialist partisan online communities is presented. The work is motivated by the extensive theoretical treatment of political conceptual morphology but limited computational implementations extant in the literature.

1 Introduction

Distributional semantic methods that use aggregate syntactic and textual word co-occurrence behaviour have been applied successfully to many natural language processing tasks. These methods are often deployed as part of a pipeline to solve an engineering problem, and evaluated by classification or correlation performance against human judgements on sub-tasks. This paper focuses on a descriptive or exploratory application of the data generated from distributional semantic analysis, allowing researchers to examine terms of interest in aggregate contexts across different kinds of co-occurrence relations and association measures.

In the social sciences, concordance views, topic models, and dictionary analysis provide a simple digest of the use of particular words in digital text collections. Although not often discussed as a specific method, in practice keyword search and snippet-views of large digital book collections are a widely used in historical and theoretical political research. The system described in this paper shows that interactive tables and network diagrams may be used to present summaries of linguistic features and word associations that allow for a descriptive interpretation of how concepts are deployed in ordinary political discourse. This can serve as a level of analysis between a close reading of the whole text collection and a fully automated bag-of-words based classification, dictionary analysis, or concordance.

This paper emphasises political concept analysis as a particularly useful application of this kind of method, due to the extensive theory of political conceptual morphology but limited computational implementations extant in previous work.

1.1 Essentially contested politcal concepts

Political discourse provides an especially suitable domain for exploratory analysis of distributional semantic data. It has long been recognized that many political concepts are 'essentially contested concepts':

> When we examine the different uses of these terms and the characteristic arguments in which they figure we soon see that there is no one use of any of them which can be set up as its generally accepted and therefore correct or standard use. (Gallie, 1955)

Researchers in political theory and intellectual history have emphasised the importance of understanding political concepts in their linguistic context(Skinner, 1969). Such studies primarily consist in close reading of primary texts produced by academics, intellectuals, and political actors, in addition to consideration of the social and cultural contexts of their time and place.

Following earlier work on the subject (Berlin, 1958; Skinner, 2012) a series of political theory articles discuss the distinction between negative and positive conceptions of liberty, in part by considering the syntactic valences and semantic affordances of the words themselves and their morphological variants (Gray, 1980; MacCallum, 1967; Oppenheim, 1961)

Koselleck (1989) discusses the historical origins of debates over voting rights by pointing to the semantic distinctions drawn by the terms *bürger* and *citoyen* in the eighteenth century. De Bolla (2013) traces the history of the concept of human rights through word use in eighteenth century corpora, and proposes a method and typology for conceptual analysis that captures variation in levels of abstraction and the rhetorical or ideational functions of concepts.

Freeden (1994) refines Gallie's notion of essential contestability by emphasising that particular instantiations of concepts ('conceptions') expressed in political discourse consist of 'empirically ascertainable and describable' as well as normative parts. This chimes with the descriptive goal and empirical methods commonly now employed by lexicographers to capture the meaning of words in general. Freeden and others (e.g. Finlayson (2007); Oppenheim (1983)) urge political theorists to investigate the structure of political concepts through their actual usage in text, with reference to structures such as a substantive core and optional peripheral components, or the roles filled by concepts as they are expressed in sentences. This approach has much in common with how word meaning is modelled by computational semantics (Jackendoff, 2010; Pustejovsky et al., 1993), though the two literatures are not connected. This does not assume that analytic treatment of linguistic context can resolve any 'true' or 'correct' meaning of a contested concept, but simply that usage reflects meaning as held by the community or ideology that produces the text.

Despite the history of interest in linguistic analysis in political theory and the history of intellectual thought, quantitative political science researchers generally treat text analysis as a means to an end, whereby the distribution of words into documents allows for measurement of attention to issues, or estimation of ideological positions. Bag-of-words techniques often suffice for this task, as it is generally observed that political actors tend to reveal ideology more often through relative issue emphasis than by expressing contrasting beliefs, desires, or intentions about the same issues (Budge, 2001).

Political document scaling methods solve a practical estimation problem — how similar is each document to the others in the corpus along particular ideological dimensions? (Slapin and Proksch, 2008; Laver et al., 2003). These inferred positions can then be used in statistical models of the political institutions or processes that produced the documents. With some exceptions (Monroe et al., 2008; Sagi et al., 2013), the intention is usually not to interpret the weights or parameter estimates of the model's linguistic features in order to investigate the relationship between the language used and the implied political position. Where topic models are applied to measure issue emphasis or attention, the goal is to incorporate these measurements into a wider model of political attention or an analysis of the effects of speeches focusing on particular topics. (Grimmer and Stewart, 2013; Grimmer, 2010; Quinn et al., 2010).

Computer scientists have applied more complex methods in order to recover ideological position from text: Iyyer et al. (2014) use recursive neural networks to detect political ideology compositionally, using crowd-sourced annotations of congressional debates. However, the focus is again to expedite the data-annotation process with machine learning, rather than to present and interpret the linguistic structures that give rise to ideological differences.

Given the ongoing significance of contested concepts in public debate (for an example treatment of the term *neoliberalism*, see (Ferguson, 2010)), a computationally-assisted comparative analysis of

descriptive word meaning in various political communities can give insight into the nature of ideological disagreement and miscommunication. This work demonstrates that interactive visualisations and novel application of standard network analysis techniques such as community detection and centrality measures discover relevant ideological distinctions.

1.2 Previous work

Mixed-method approaches to text analysis in political science have made use of commercial software for manual coding or visualisation of dictionary or factor-analysis methods (Reinert, 1993) with word clouds or tables. In corpus linguistics and lexicography, perhaps the most widely used tool is the Sketch Engine (Kilgarriff et al., 2014), which presents tables summarising the selectional preferences of terms of interest gathered from dependency parsed corpora. Sketch Engine has been widely deployed in lexicography and the study of language learning, but less often for broader questions in social science (Blinder and Allen, 2015). In computational semantics, the Wordnet, Concept Net and generative lexicon projects specify the representation of concepts, but lack implementations of exploratory tools beyond presenting tables of the resulting structures, and derive their associations from 'universal' human curated databases, rather than automatically from ideologically partisan corpora.

Network representations of concepts are widely studied in cognitive science (Steyvers and Tenenbaum, 2005; Gruenenfelder et al., 2015). Lopes et al. (2010) describes a web-based interactive tool for exploring networks derived from bioinformatic data. Van Atteveldt (2008) describes an RDF-based system for representing and querying semantic network data from Dutch newspaper articles.

(Shneiderman and Aris, 2006) presents software for dynamically presenting network visualisations to allow the user control the density and visibility of nodes based on degree and node metadata, although the software is not yet publicly available [1].

2 Method

Conceptual structures are implied in text through both syntactic and non-syntactic relations. The type, attributes, and functional roles of a concept may be indicated through paradigmatic and syntagmatic grammatical relations, and also by general thematic textual co-occurrence counts as leveraged by topic models and document classification systems. The system described here implements network visualisations derived from both syntagmatic grammatical relations and textual co-occurrences. For both types of co-occurrence, the association between words is calculated using adjusted pointwise mutual information, with the context distribution smoothing method of Levy et al. (2015) used to reduce the impact of very infrequent co-occurrences.

The corpus used to construct the networks for figures in this paper and the linked R Shiny web applications is a collection of all comments from the *libertarian* and *socialism* communities on the website `reddit.com`. This data shows the contrasting way in which political concepts are deployed in natural discourse within self-selecting ideologically partisan communities. All comments from 2014-2015 with a positive rating of three or higher and containing at least two sentences are included. Reddit comments are public data available through an API provided by the website[2]

2.1 Word co-occurrence extraction

Grammatical relations are extracted with a syntactic dependency parser implemented in the SpaCy python package for natural language processing, accessed through the R spacyr package.[3] This parser has been shown to achieve state-of-the-art accuracy on part-of-speech tagging and dependency parsing evaluation datasets (Honnibal et al., 2015). To simplify the visualisation and try to focus on the

---

[1] http://www.cs.umd.edu/hcil/nvss/#software

[2] https://www.reddit.com/dev/api/

[3] https://github.com/explosion/spaCy, spacyr package: https://github.com/kbenoit/spacyr

most informative relations, the detailed dependency tagset is reduced to four general relation types, and functional categories (determiners and auxiliaries) and prepositional relations are excluded. The resulting relation types are conjunction, modification, verb-subject and verb-object. The type of semantic association is indicated by the colour of the edge, and the strength of association by the edge width. Non-syntactic associations are measured by counting co-occurrences within the same comment, but excluding co-occurrences within the same sentence. The intention is to properly separate the data gathered from sentential and non-sentential (document) relations.

2.2 Interactive network visualisation

The visualisation is implemented as an R Shiny web application. The figures in this section show screenshots from the network representatations displayed by the web app. These are 'neighbourhood' or 'ego' graphs of order two, that is, they show nodes within at most two edges from the focal node — an edge exists between two nodes if their PMI association is a above a user-specified threshold. The network is drawn using the R *visNetwork* package, using a force-directed algorithm (Fruchterman and Reingold, 1991), which models the network mechanically as repelling particles connected by springs. The result is that in a graph of suitable density and degree, nodes are spaced apart enough to be distinguished, but the edges pull together nodes into clusters that share many relations.

The figures in this section are illustrative only, intended to show the interface design and the general properties of the network. The application should be evaluated using the online interactive prototypes for the syntactic [4] and textual co-occurrence [5] data. The visNetwork package implements a drag, pan, and zoom enabled central widget, and this is combined with input boxes for search terms and sliders for setting the order of the neighbourhood graph, the maximum node degree, and the association score threshold.

Static images of these networks are of limited use when large enough to show structures larger than a few individual nodes — attempting to label all of the nodes makes them unreadable. If the number of nodes is reduced in order to make the labels legible, then the resulting network is too small to show interesting structure at a large or medium scale. Interpretation or exploration of these semantic networks is therefore best approached through an interactive interface which allows for adjustment in the scale and highlighting of particular neighbourhoods.

Figure 1 shows syntactic association graphs for the term *power* tagged as a noun for the socialism (left) and libertarian (right) communities. In the socialism community, the immediate neighbours of power are *priviledge, wealth, influence* (conjunction relations), *rule* (verb subject) and *labour* (modifier). For the libertarian community, the immediate neighbours are *corrupt, enforce, abuse, sell* (verb subject); *limit* (verb object), *wealth* and *influence* (conjunction).

---

[4]Syntactic co-occurrences: `http://52.207.96.220:3838/iwcs_app_syntax/`
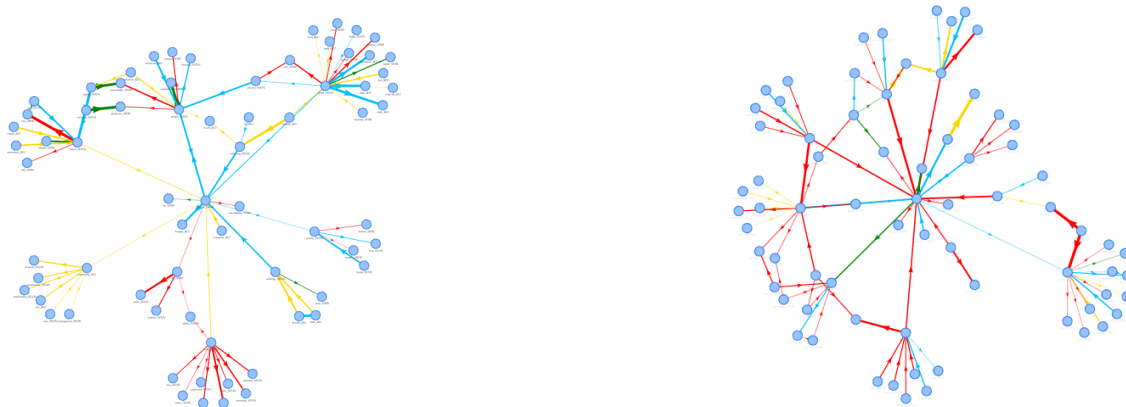[5]Textual co-occurrences `http://52.207.96.220:3838/iwcs_app_cooc/`

Figure 1: *power* (noun) in the socialism (left) and libertarian (right) subreddit

Figure 2 shows syntactic association graphs for the term *freedom* tagged as a noun for the socialism (left) and libertarian (right) communities. In the socialism community, the immediate neighbours of freedom are *democracy, wealth, equality* (conjunction relations), *attain* (verb subject) and *true, personal* (modifier). For the libertarian community, the immediate neighbours are *choose, associate, abridge, value, restrict* (verb object); *religious, personal* (modification), and *liberty* (conjunction).
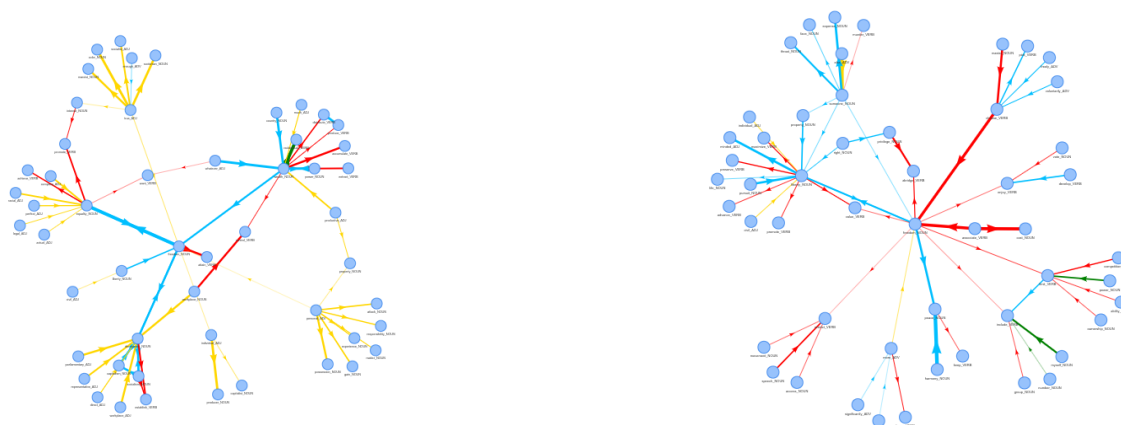
Figure 2: *freedom* (noun) in the socialism (left) and libertarian (right) subreddit

Figure 3 shows the full interface for the textual co-occurrence relationships. The sliders on the left allow control of node degree, association score threshold, and the selection of variations of the PMI association measure. The concept shown is the verb 'plan', for the socialism subreddit, where the immediate neighbours are *enterprise*, *central*, *economy*, *innovation*, and *diet*.

Figure 3: Interface for Shiny application for graph visualisation. Sliders control node degree, association score threshold, and word association measure calculation. The central node of the ego network is the verb *plan*.
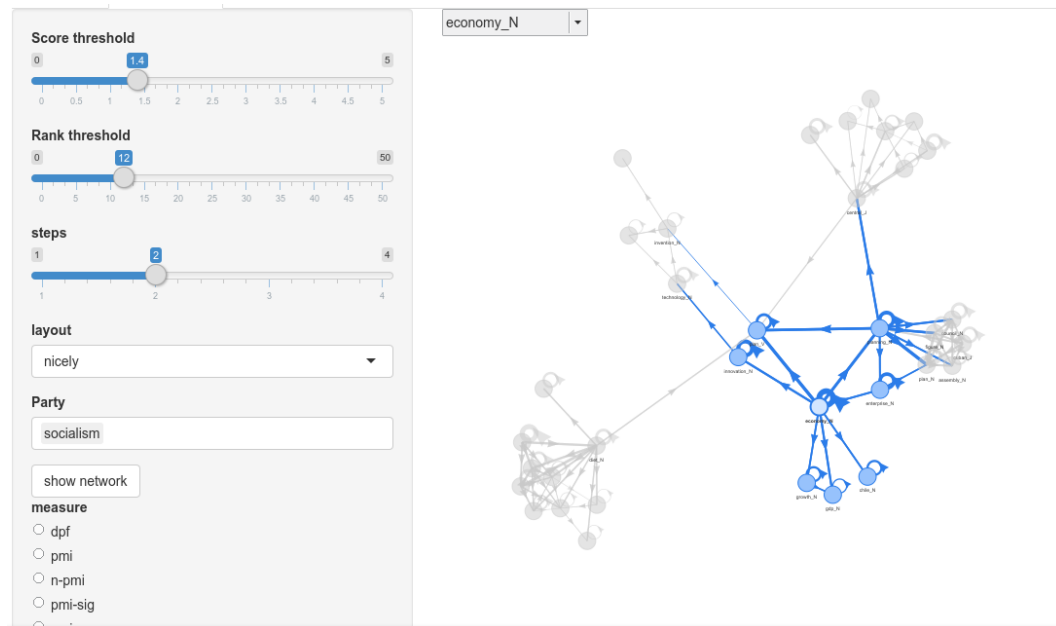
Syntactic selectional preferences require an automated parser trained on comparable text, and therefore it is not appropriate to apply this analysis to historical text, especially with additional noise resulting from optical character recognition. However, general context-based co-occurrences may be extracted from historical texts. Figure 4 shows a joint neighbourhood network of the terms *freedom* and *liberty* from the 1795 – 1800 portion of the Eighteenth Century Collections Online corpus. Edges are associations based on sentence-level co-occurrences. The user can choose from a number of lexical association measures; in this case t-score associations are shown. Lexical association measures typically trade-off the degree to which generally frequent terms are scored highly due to the significant weight of evidence resulting from high co-occurrence counts, or penalized due to the high independent probability of co-occurrence of generally frequent terms. The t-score tends to prioritise generally frequent words, while pointwise mutual information often over-estimates the significance of co-occurrence of words with low overall frequencies (Evert, 2005).

Figure 4: Joint neighbourhood network of the terms *freedom* and *liberty* from the 1795 – 1800 portion of the Eighteenth Century Collections Online corpus. Color indicates dominant part-of-speech; edge width indicates log t-score association. An online demo is available here: `http://52.207.96.220:3838/ecco-network-1/`. This demo also provides interactive interfaces to igraph's clique-finding and shortest-path functions applied to the semantic network derived from ECCO.

.

2.3 Centrality measures

Beyond network visualisation, many standard network analysis methods may be applied to semantic networks to discover the most important nodes and clusters in the graph. The table below shows the highest scoring words by two commonly-used network centrality measures — eigenvector centrality and betweenness centrality — for the libertarian and socialism subreddits. While further work is needed to go beyond face-validity, the terms seem to reflect common concerns in the ideological communities from which the semantic network is constructed.

Libertarian		Socialism	
eigen	betweeness	eigen	betweeness
detention_N	law_N	class_N	class_N
detainee_N	government_N	production_N	capitalism_N
facility_N	tax_N	revolution_N	party_N
administration_N	libertarian_N	struggle_N	socialism_N
detain_V	pay_V	relation_N	revolution_N
congress_N	year_N	movement_N	movement_N
law_N	state_N	commodity_N	society_N
amendment_N	libertarianism_N	labor_N	production_N
prosecute_V	business_N	socialism_N	war_N
authorize_V	get_V	capitalism_N	union_N
charter_N	rate_N	value_N	state_N
judge_N	crime_N	mode_N	think_V
section_N	amendment_N	development_N	power_N
senate_N	marriage_N	party_N	struggle_N
suspect_N	insurance_N	participation_N	time_N
combatant_N	property_N	produce_V	work_V
trial_N	war_N	contradiction_N	system_N
conviction_N	case_N	society_N	say_V
restore_V	regulation_N	communism_N	value_N

3   Conclusion

This paper presents an interactive web visualisation system for exploring conceptual associations derived from syntactic and textual lexical co-occurrence data. The method is motivated by the need for descriptive and exploratory methods for investigating the linguistic context of essentially contested contests in political discourse. A demonstration is presented showing the usage of political concepts in ideologically partisan comments in an online community, and an interface to neighborhood network plots and clique-detection for a portion of the Eighteenth Century Collections Online corpus. Future extension of this work will diachronically and synchronically compare modern usage of particular contested concepts in online social communities with that of political actors from records such as Hansard, and extend the analysis beyond visualisation to make further use of network analysis techniques such as community detection and centrality measures.

References

Berlin, I. (1958). Two concepts of liberty.

Blinder, S. and W. L. Allen (2015). Constructing immigrants: Portrayals of migrant groups in british national newspapers, 2010–2012. *International Migration Review*.

Budge, I. (2001). Validating party policy placements. *British Journal of Political Science 31*(01), 179–223.

De Bolla, P. (2013). *The architecture of concepts: The historical formation of human rights*. Fordham Press.

Evert, S. (2005). The statistics of word cooccurrences: word pairs and collocations.

Ferguson, J. (2010). The uses of neoliberalism. *Antipode 41*(s1), 166–184.

Finlayson, A. (2007). From beliefs to arguments: Interpretive methodology and rhetorical political analysis. *The British Journal of Politics and International Relations 9*(4), 545–563.

Freeden, M. (1994). Political concepts and ideological morphology. *Journal of Political Philosophy 2*(2), 140–164.

Fruchterman, T. M. and E. M. Reingold (1991). Graph drawing by force-directed placement. *Software: Practice and experience 21*(11), 1129–1164.

Gallie, W. B. (1955). Essentially contested concepts. In *Proceedings of the Aristotelian society*, Volume 56, pp. 167–198. JSTOR.

Gray, J. N. (1980). On negative and positive liberty. *Political Studies 28*(4), 507–526.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 1–35.

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

Gruenenfelder, T. M., G. Recchia, T. Rubin, and M. N. Jones (2015). Graph-theoretic properties of networks based on word association norms: implications for models of lexical semantic memory. *Cognitive science*.

Honnibal, M., M. Johnson, et al. (2015). An improved non-monotonic transition system for dependency parsing. In *EMNLP*, pp. 1373–1378.

Iyyer, M., P. Enns, J. Boyd-Graber, and P. Resnik (2014). Political ideology detection using recursive neural networks. In *Proceedings of the Association for Computational Linguistics*, pp. 1113–1122.

Jackendoff, R. (2010). *Meaning and the lexicon: the parallel architecture, 1975-2010*. Oxford University Press.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlỳ, and V. Suchomel (2014). The sketch engine: ten years on. *Lexicography 1*(1), 7–36.

Koselleck, R. (1989). Linguistic change and the history of events. *The journal of Modern history 61*(4), 650–666.

Laver, M., K. Benoit, and J. Garry (2003). Extracting policy positions from political texts using words as data. *American Political Science Review 97*(02), 311–331.

Levy, O., Y. Goldberg, and I. Dagan (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics 3*, 211–225.

Lopes, C. T., M. Franz, F. Kazi, S. L. Donaldson, Q. Morris, and G. D. Bader (2010). Cytoscape web: an interactive web-based network browser. *Bioinformatics 26*(18), 2347–2348.

MacCallum, G. C. (1967). Negative and positive freedom. *The Philosophical Review 76*(3), 312–334.

Monroe, B. L., M. P. Colaresi, and K. M. Quinn (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis 16*(4), 372–403.

Oppenheim, F. E. (1961). Dimensions of freedom: An analysis.

Oppenheim, F. E. (1983). Political concepts: A reconstruction.

Pustejovsky, J., P. Anick, and S. Bergler (1993, June). Lexical semantic techniques for corpus analysis. *Comput. Linguist. 19*, 331–358.

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.

Reinert, M. (1993). Les mondes lexicaux et leur logiqueà travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société 66*, 5–39.

Sagi, E., D. Diermeier, and S. Kaufmann (2013). Identifying issue frames in text. *PLoS one 8*(7), e69185.

Shneiderman, B. and A. Aris (2006). Network visualization by semantic substrates. *IEEE Transactions on Visualization and Computer Graphics 12*(5), 733–740.

Skinner, Q. (1969). Meaning and understanding in the history of ideas. *History and theory 8*(1), 3–53.

Skinner, Q. (2012). *Liberty before liberalism.* Cambridge University Press.

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.

Steyvers, M. and J. B. Tenenbaum (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science 29*(1), 41–78.

Van Atteveldt, W. (2008). Semantic network analysis: Techniques for extracting, representing, and querying media content. *PhD Thesis*.