CrossMark

The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016

Emanuele Lapponi[1] · Martin G. Søyland[2] ·
Erik Velldal[1] · Stephan Oepen[1]

Abstract In this work we present the Talk of Norway (ToN) data set, a collection of Norwegian Parliament speeches from 1998 to 2016. Every speech is richly annotated with metadata harvested from different sources, and augmented with language type, sentence, token, lemma, part-of-speech, and morphological feature annotations. We also present a pilot study on party classification in the Norwegian Parliament, carried out in the context of a cross-faculty collaboration involving researchers from both Political Science and Computer Science. Our initial experiments demonstrate how the linguistic and institutional annotations in ToN can be used to gather insights on how different aspects of the political process affect classification.

Keywords Computational political sciences · Computational social science · Language technology · Natural language processing · Parliamentary proceedings

✉ Emanuele Lapponi
emanuel@ifi.uio.no

Martin G. Søyland
m.g.soyland@stv.uio.no

Erik Velldal
erikve@ifi.uio.no

Stephan Oepen
oe@ifi.uio.no

1 Language Technology Group, Department of Informatics, University of Oslo, Oslo, Norway

2 Department of Political Sciences, University of Oslo, Oslo, Norway

🖄 Springer

1 Introduction

A large part of political science studies relies on text as the main source of data. Be it policy evaluation, the performance of civil service, explaining elite behavior, or analyzing international negotiations—researchers rely on vast amounts of written or spoken forms of political text and methods for systematizing and summarizing these. Quantitative methods (content analysis) have, in particular, become more and more popular in recent years due to an ever increasing availability of large amounts of data, computational power for handling them, and the methods for properly studying them.

In this work, we provide a freely available and carefully curated data set of Norwegian parliamentary proceedings, to lower the technological barrier to entry to text-based political science research and to aid replicability of results. The data set results from an interdisciplinary (informatics–political sciences) collaboration and provides a unique combination of rich, non-linguistic metadata and ready-to-use morpho-syntactic analysis of its textual content. This analysis was performed through the Language Analysis Portal (LAP; Lapponi et al. 2014), and the data set is maintained and distributed under the auspices of the Norwegian CLARIN branch.

The remainder of the paper is structured as follows. Section 2 presents related work, focusing on recent and sustained efforts and data sets. Section 3 describes the ToN data, how it was produced and how to obtain it. Section 4 provides some preliminary results on party classification using ToN, seeking to showcase the kind of insights that can be gathered when investigating results over meaningful subsets of the data. Finally, Sect. 6 presents our concluding remarks and plans for future work.

2 Related work

Text-based quantitative analysis of parliamentary proceedings is an active area of research in Political Science. Both supervised and unsupervised text classification techniques are used in tandem with non-textual data sources (e.g. roll call votes and survey results) to gather novel insights and drive the field forward. Clustering and other *unsupervised* modeling techniques have become a staple of this kind of research. Notable examples in recent years include Eggers and Spirling (2014), who show that the level of conflict in the electoral districts of a given member of parliament (MP) is important for her participation in both voting and speech-making; Bäck and Debus (2016), who use the Wordscores technique (Laver et al. 2003) to explore what causes MPs to participate more or less actively in parliament and why they sometimes deviate from the party line; Lauderdale and Herzog (2016), who demonstrate that a hierarchical approach to the Wordfish algorithm (Slapin and Proksch 2008) greatly improves its quality when applied to parliamentary speeches; and Proksch and Slapin (2015), who study parliamentary speeches from the UK, Germany, and New Zealand, showing that backbencher MPs deviate more from their party line in majoritarian than proportional representation electoral systems.

While not as ubiquitous, *supervised* classification techniques have also been adopted as a means to investigate research questions related to ideology in parliaments. Yu et al. (2008) find that training an ideology classifier is possible and fairly generalizable based on their classification results on congressional speeches in the US. Høyland et al. (2014), by using a similar approach, classify party affiliation in the European Parliament. While the results are generally less accurate, mostly because of the multi-party setting (in contrast to the two-party system of the US, where a majority baseline would yield results that are comparable to the best reported EU classifier configuration), they also demonstrate that some parties are harder to classify than others. For example, the Liberal (ELDR) Party is argued to be a hard case because it shifted coalition allegiance between parties in the period under investigation, and consisted of an ideologically heterogeneous party group based on the MEP's country of origin. In their experiments on the Canadian Parliament debates, Hirst et al. (2010) find that the driving features in party classification are those describing roles of opposition and government, suggesting that classification performance is mostly driven by the language of attack and defense, rather than a party's ideological and political profile.

The data that enables researchers to conduct studies like the ones mentioned above is typically available through public institutions. However, considerable efforts have been made in order to transform the 'raw' data into more easily digestible formats, often augmenting it with additional information. The Canadian Hansard Dataset (Beelen et al. 2017), studied by Hirst et al. (2010), is a collection of debates from the Canadian House of Commons. The data set is searchable via a web interface,[1] and available for download in a variety of formats, including a series of daily UTF-8 comma-separated value (CSV) files. Notably, while digitization of the speeches started in 1994, the data made available by this effort dates back to 1901. Pre-1994 data had to be scanned and processed. The congressional speech data[2] (Thomas et al. 2006), studied in Yu et al. (2008), collects all publicly available pages of the 2005 U.S. House record. The speeches are serialized in individual files, with underscore-separated annotations in the filenames. These include speaker party and whether or not the speaker voted in favor of the bill discussed in the session.

European politics are also covered by a number of parliamentary debates collections. Talk of Europe (ToE; van Aggelen et al. 2017) collects debates from the European Parliament. This initiative builds on the data studied in Høyland et al. (2014), and makes it available in the form of an RDF graph that connects it with additional metadata on the speakers and other facets of European politics. In the Scandiavian context, the plenary sessions of the Finnish,[3] Danish,[4] and Swedish[5] parliaments are also available to researchers. Finally, the Norwegian parliamentary

---

[1] http://www.lipad.ca/.

[2] http://www.cs.cornell.edu/home/llee/data/convote.html.

[3] http://urn.fi/urn:nbn:fi:lb-2017020201.

[4] https://clarin.dk/clarindk/item.jsp?id=dkclarin:986010

[5] https://spraakbanken.gu.se/eng/resources, under *Riksdag*.

Table 1 Basic corpus statistics for the ToN data, broken down across political party labels, also showing the corresponding abbreviated name for each party

Party/source	Abbr.	#Speeches	#Tokens	%NNO
President		72,646	2,525,733	0.70%
Labor Party	Ap	43,483	16,008,420	0.90%
Conservatives	H	32,945	11,481,762	0.20%
Progress Party	FrP	30,217	9,729,435	0.50%
Socialist Left Party	SV	19,941	7,218,136	18.00%
Christian Democrats	KrF	19,720	6,653,088	19.00%
Center Party	Sp	18,255	5,874,381	33.00%
Liberal Party	V	11,579	3,830,095	0.80%
Green Party	MDG	508	153,834	0.01%
Coastal Party	Kp	492	128,709	0.06%
Non-Partisan Deputies	TF	409	97,001	0.00%
Independent		131	38,284	0.00%
Other		47	64,715	19.00%
Total		250,373	63,803,593	19.00%

Rows are sorted by the number of speeches, but we also show the number of tokens for each class and the percentage of speeches held in the Nynorsk variant of Norwegian (NNO)

debates from 2008 to 2015 are available through Corpuscle[6] (Meurer [2012]). While offering the same core data as the ToN corpus presented in the current article, this latter effort differs from ours in several aspects: (a) it covers only part of the digitally available proceedings, while ToN speeches go back to 1998; (b) it makes available a very small subset of the available metadata information on the speeches (5 metadata variables including language identification, compared to ToN's 83); (c) it does not provide linguistic annotations.

3 The Talk of Norway data set

The Talk of Norway (ToN) data set is a collection of the digitized records from the Norwegian *Storting* (Parliament) (1998–2016), centered around the transcribed speeches of the members of parliament (MPs). It provides researchers investigating questions akin to those described in Sect. [2] with a rich set of readily available data variables, providing detailed meta-information not only on the speeches, but also on the MPs and their parties, as well as contextual information on the cabinet and the ongoing debate at the time the speech is held.

In the period covered by the data, the parliament has consisted of seven main parties, that have held seats in all of the parliamentary sessions[7] from 1998 to 2016. It also contains speeches from three smaller parties (Green Party, Coastal Party, and Non-Partisan Deputies) that occupy only a small share of seats in specific

---

[6] http://clarino.uib.no/korpuskel/corpus-list, under *Storting debates*.

[7] One parliamentary session, roughly, starts in October one year and lasts until September the next year.

parliamentary periods. The same period has also seen three prime minsters (Kjell Magne Bondevik, Jens Stoltenberg, and Erna Solberg), that have lead six distinct cabinets.[8]

At the level of the speaker, ToN provides records of the county the MP was elected from, gender, party affiliation, committee membership, and more. At the level of the party, there are variables denoting how many seats the party has at any given time frame, and whether the cabinet is part of the government at the time of the speech. At the cabinet level, ToN provides the start and end date of the cabinet and its composition. The available variables also include a variety of data on the ongoing debate at the time the speech is held, such as the responsible committee, the MP asking a question during question hour, keywords denoting the topic of the speech, and so on. The result is a data frame with 250,373 speeches over 83 variables.

The foundation for the data, the speeches, was structured and provided by *Holder de ord*[9]—an independent organization that makes available digital tools for political analysis in Norway—whereas most of the metadata on the representatives, bills, propositions, and questions was obtained through the Parliament's own API[10] and merged with the speech data. The API, however, does not make available various important sources of information, which we were able to obtain by scraping the Parliament website directly. These include attributes such as the debate subject, the questions asked during question hours and interpellations (where ministers give in-depth answers to questions on large policy areas), and the name of committees (e.g. *Transport and Communication Committee*). The scraping itself was done via exact match of the speeches in the ToN data to the raw HTML of the Parliament's website,[11] and the relevant information was retrieved by parsing the HTML markup.

We found essential metadata on several cabinet-related attributes to be missing in both the *Storting* API and website, including information on the role of parties in a given period (e.g. opposition, cabinet, and support parties), and cabinet composition (e.g. single-party or coalition). We make these variables available by merging the hand-coded data from Søyland (2017) with the ToN speeches.

Linguistic Annotations This first version of the ToN corpus also seeks to facilitate access to linguistic annotations for the speeches themselves. As so-called text-as-data approaches become increasingly prominent in political science, the field is also gradually becoming aware of the effects that pre-processing decisions have on models built on natural language data. Matthew J. Denny and Arthur Spirling (2017) show that pre-processing decisions (ranging from word tokenizer and stemmer choices to dimensionality reduction approaches) can lead to radically different analyses of the same text. They call for a choice of pre-processing steps that is informed by the nature of the problem at hand, noting that many experiments

---

[8] In line with the literature on cabinet formation, cabinets are counted by change of prime minister, party constellation, and elections. Thus, we have two Bondevik cabinets, three Stoltenberg cabinets, and one Solberg cabinet.

[9] https://www.holderdeord.no.

[10] https://data.stortinget.no.

[11] http://www.stortinget.no.

in the field simply replicate the steps taken by a handful of seminal papers. Our position is that using state-of-the-art, language-specific linguistic pre-processing is a sensible starting point for any research project in this field. However, we find that at least one prominent multi-lingual study (Bäck and Debus 2016) does not use a Norwegian-aware tokenizer for the Norwegian data. We speculate that this kind of choice is rooted in (a) the authors not being aware of available NLP tools or (b) technical challenges in installing, running, and decoding the output of less known tools.

In order to facilitate access to state-of-the-art linguistic annotations for Norwegian, ToN speeches are distributed with basic pre-processing, as detailed below. They are first run through a language identifier (Lui and Baldwin 2012), which assesses whether a speech is given in *Bokmål* or *Nynorsk*, the two official standards of written Norwegian language[12] (the percentage of speeches classified as Nynorsk is shown in Table 1, along with statistics on the number of speeches and tokens for each party). This annotation serves two purposes. One is to provide the information to potential users: because parliamentary debates are written in two languages, automatic analysis results can potentially be driven by the language rather than the actual content of the speech. This has, largely, been ignored in political science studies on Norwegian records. The other purpose of this annotation is to inform the other tool used to analyze the speeches so that it can be configured correctly. This tool, the Oslo–Bergen Tagger (OBT), annotates text with sentence and token boundaries, lemmas, parts of speech (PoS), and morphological features (Johannessen et al. 2012).

These morpho-syntactic annotations were obtained from the Language Analysis Portal[13] (Lapponi et al. 2014), an initiative that aims at providing researchers outside of NLP with easy access to state-of-the-art tools. Part of the mission of the annotation and experimental efforts in this ongoing cross-faculty collaboration is informing the system architects to allow the replication of end-to-end experimentation directly in the portal. In hope to foster more experimentation with the Norwegian parliamentary debates, we make the full ToN data set publicly available.[14]

Data Format and Utilities For ease of access across a broad range of user groups and tools, the core component of the ToN data set is a CSV file, where each line contains comma-separated values for the metadata variables, including the raw unprocessed speeches. Linguistic annotations reside in auxiliary, tab-separated value (TSV) files, one per speech. These are linked to their respective row in the main CSV by way of the file name, which is a unique id variable. In the tradition of shared tasks at the Conferences on Computational Natural Language Learning (CoNLL), tokens are separated by a single newline, while sentence boundaries are encoded as double newlines. Figure 1 displays the first two sentences of the (chronologically) first speech in the ToN data set: Tabulator characters separate

---

[12] As a safeguard, classifications outside of the *Bokmål/Nynorsk* range were corrected to *Bokmål*.

[13] https://lap.clarino.uio.no.

[14] Please see http://www.mn.uio.no/ifi/english/research/projects/ton/ for access information and updates around this project.

1	Ærede	ære	adj	fl|<perf-part>|tr1
2	medrepresentanter	medrepresentant	subst	appell|mask|ub|fl
3	!	$!	clb	<<<|<utrop>|<<<

1	Tidligere	tidlig	adj	komp
2	stortingsrepresentant	stortingsrepresentant	subst	appell|mask|ub|ent
3	Sjur	Sjur	subst	prop|mask|<*>
4	Lindebrække	Lindebrække	subst	prop|<*>
5	er	være	verb	pres|a5|pr1|pr2|<aux1/perf_part>
6	død	død	adj	ub|m/f|ent|pos
7	–	–	symb	_
8	89	89	det	fl|kvant
9	år	år	subst	appell|nøyt|ub|fl
10	gammel	gammel	adj	ub|m/f|ent|pos
11	.	$.	clb	<<<|<punkt>|<<<

Fellow honorable representatives! Former member of parliament Sjur Lindebrække is dead – 89 years old.

Fig. 1 The first two sentences of the first speech in the ToN data set, "tale000000.tsv". These five columns, from left to right, contain the following values: CoNLL-style token indices, which reset to 1 for each sentence, followed by surface forms, lemmas, parts-of-speech, and pipe character-separated morphological features

annotation fields for each token, viz. the surface form, lemma, part-of-speech, and morphological features; given the variable cardinality of the latter, each set of features is split by pipe characters (|), and occupies a single field.

The choice of file formats is motivated by common tools and workflows adopted by quantitative-oriented social scientist. We speculate that serializations such as the elaborate RDF triples from ToE or the CG3 XML format of OBT are not immediately usable for the main consumers of the data, typically relying on statistical software such as SPSS, Stata, or R. To further lower the entry barrier to text-as-data experimentation with ToN, we bundle the data with libraries to easily read and manipulate metadata and linguistic information jointly in both R and Python. The in development R-package *tonR* includes functions for reading the annotated CoNLL-like files, constructing corpora from a set of speeches, calculating F_1 scores from classification experiments, and more. The ToN Python library *ton.py* allows users to stream speeches with both metadata and linguistic annotations into Python dictionaries, making it easy to integrate ToN into existing Python workflows. Additionally, it can be used to re-serialize the data into JSON, using the JSONlines file format.[15] Both libraries are available throught the projects github pages.[16]

Finally, ToN is accessible in the Corpuscle corpus management application, where the data set can be queried with an array of language analysis tools,[17] and the csv file with the variable as well as the tsv files with the annotations can also be obtained through a CLARINO repository (Lapponi and Søyland 2016).[18]

---

[15] http://jsonlines.org/.

[16] https://github.com/ltgoslo/talk-of-norway.

[17] http://clarino.uib.no/korpuskel.

[18] https://repo.clarino.uib.no/xmlui/handle/11509/123.

Table 2 Party-wise results for the best performing classifier configuration (with the best score for each metric in bold), also showing macro-averaged F_1 for all parties and overall accuracy, to be compared to the majority class baseline

| Party	Abbr.	P	R	F_1	Accuracy
Socialist Left Party	SV	0.578	0.490	0.531	–
Labor Party	Ap	0.471	0.624	0.537	–
Center Party	Sp	0.618	0.527	0.569	–
Christian Democrats	KrF	0.578	0.433	0.495	–
Liberal Party	V	**0.637**	0.351	0.452	–
Conservatives	H	0.503	0.485	0.494	–
Progress Party	FrP	0.603	**0.665**	**0.632**	–
Baseline		0.035	0.142	0.056	0.248
Macro		0.570	0.511	0.538	0.539

4 Preliminary experiments

We here report on a first suite of preliminary experiments on the the Talk of Norway corpus, training a maximum-margin classifier to assign party labels to individual speeches. Our aim here is to provide an example of how the linguistic and institutional data in the corpus can be taken advantage of in Political Science research; the reported results themselves and their (preliminary) discussion is the initial output of ongoing quantitative research on Norwegian party politics. The experiments are performed on a subset of the ToN data where we exclude all speeches lacking a party identifier (for instance, everything uttered by the president). We also remove all speeches from parties that do not appear across all sessions, such as the Green Party (MDG), and speeches comprised of less than 100 tokens. We then divide the resulting data set into six folds—each comprised of speeches held under a given cabinet—and perform six-fold cross validation experiments. Recall that the ToN data set encompasses the last six Norwegian governments.

Speeches are represented as TF-IDF weighted vectors, filtering out common Bokmål and Nynorsk stop words as well as the 100 tokens with the highest IDF values.[19] We use the Linear SVM implementation available through Scikit Learn (Pedregosa et al. 2011), a widely adopted Python package for machine learning. We performed empirical tuning of various feature configurations and hyperparameters, including the SVM regularization parameter (C) governing the trade-off of training error and margin-size.[20]

---

[19] The full set of ToN tokens and associated IDF values is available in the ToN github repository.

[20] This was done using Scikit Learn's grid search functionality on each training fold individually, testing with exponentially larger values of C (ranging from 0.1 to 1000). For all folds, the returned best value was consistently 1.
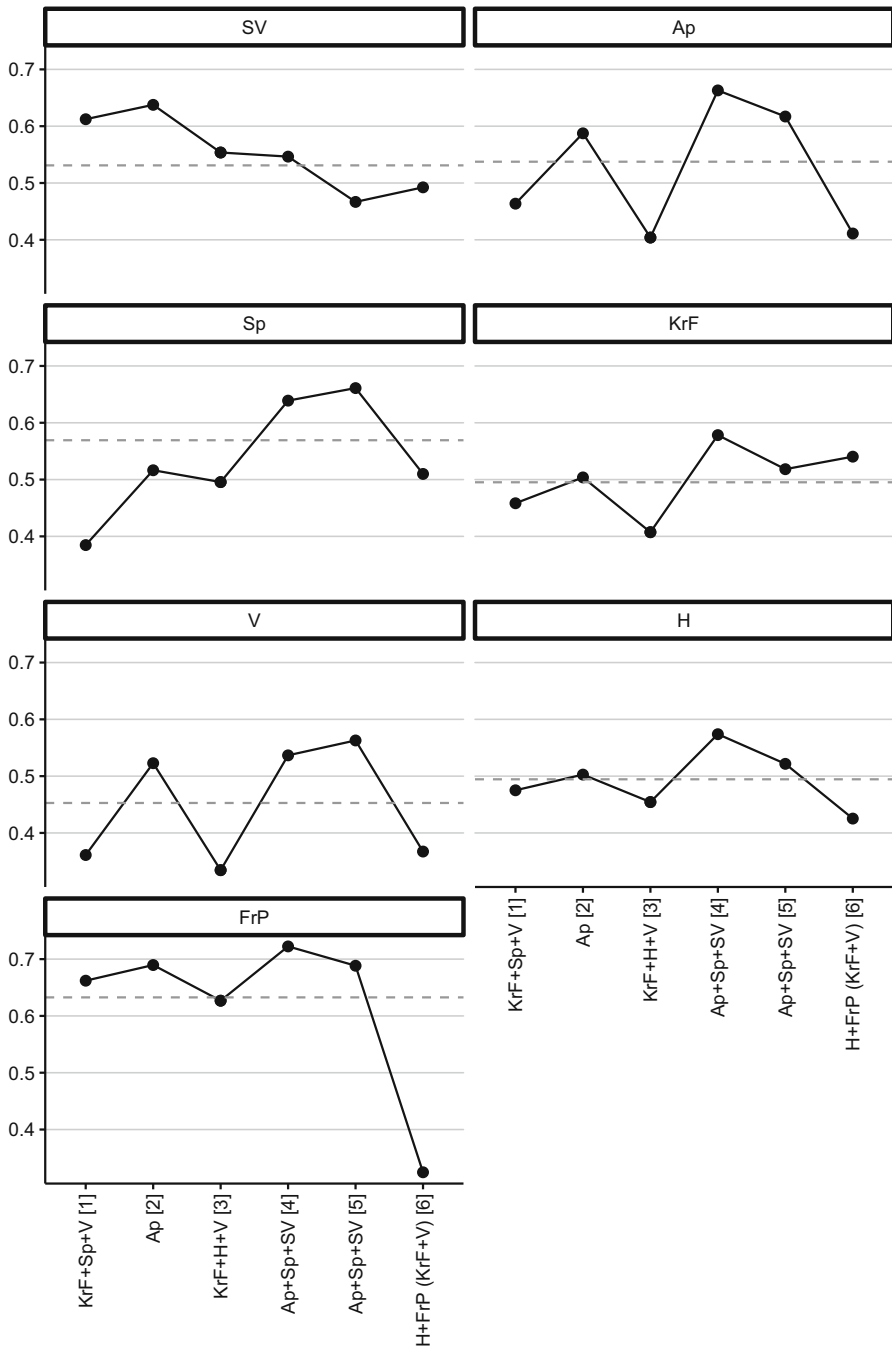
Fig. 2 Party-wise F_1 scores for different cabinet periods, where the points show F_1 for the party under each of the six cabinets and the dashed line shows F_1 for the party over all speeches in the full sample. The x-axis is ordered by cabinet sessions, and the party of the Prime Minister is the first on each tick label

Actual party	Cabinet							Opposition						
SV	44.3	26.8	3.6	5.0	2.0	13.4	4.8	53.9	18.7	2.8	4.2	1.7	8.2	10.5
Ap	4.6	62.6	4.2	5.0	2.2	15.9	5.5	5.9	62.1	4.4	5.2	1.2	9.8	11.4
Sp	3.3	23.9	54.6	4.5	1.6	7.7	4.4	4.5	21.4	49.5	6.5	1.3	7.6	9.2
KrF	4.4	37.7	9.6	32.8	0.9	9.2	5.5	5.3	15.9	4.1	50.6	1.1	11.4	11.5
V	4.4	48.0	6.4	2.6	23.2	11.0	4.4	7.7	13.0	4.8	4.7	47.7	11.4	10.7
H	5.6	44.5	3.0	3.4	0.9	35.2	7.4	4.1	13.4	2.4	2.6	1.1	61.1	15.3
FrP	5.8	40.1	7.9	3.9	1.7	12.7	27.8	2.6	8.6	1.3	2.3	0.6	10.0	74.5
	SV	Ap	Sp	KrF	V	H	FrP	SV	Ap	Sp	KrF	V	H	FrP
								Predicted party						

Fig. 3 Confusion matrices for two subsets of the data, one comprised of speeches uttered by MEPs in cabinet and one by MEPs in opposition. Rows sum to 100%, so that the cells contain percentage of speeches classified in each predicted class relative to the true class

Closely mirroring the set-up of Høyland et al. (2014), we report results for the best performing configuration[21] in Table 2, using a heterogeneous set of both (a) basic linguistic and (b) non-linguistic features: Set (a) comprises token and lemma n-grams (ranging from unigrams to trigrams) and parts of speech, while set (b) encodes metadata variables such as speaker gender and county of provenance, the type of debate (minutes, question hour, interpellations, and so on), its keyword (for instance, "taxes", "research", "immigration", and so on), the name of the committee leading the debate, and finally the type of case (general issue, budget, law). In addition to party-wise F_1 scores, we report macro-averaged F_1 and accuracy for all parties. As a point of reference we also include results for a majority class baseline, corresponding to simply assigning the Labor Party (Ap) as the class label for all speeches.

These results compare favorably to previously published results for multi-party systems (Høyland et al. 2014). We are not aware of any inter-annotator agreement studies for party classification, making it hard to compare classification scores to human performance. We speculate that this is a relatively hard task even for humans, since there is significant ideological overlap between different clusters of parties on many topics.

Looking more closely at the classifier performance in Table 2 and comparing it with the corpus statistics in Table 1 we see that class size disparities do not seem to have much direct effect on classifier performance, which is not proportional to the amount of speeches available for each party. The Liberal Party (V) is an exception to this trend, being both the one where the classifier delivers the poorest performance and the party with the least amount of speeches in the subset of the data used in our experiments.

---

[21] While we have experimented with different feature sets, our primary goal here is to establish a plausible 'strong baseline'; broader feature ablation experiments are beyond the scope of this project note, and are currently being studied in the context of the aforementioned ongoing research on Norwegian party politics.

It does, however, see the highest numbers in term of precision; while this is certainly in part true because the classifier is very conservative with assigning V labels, it also means that these are the ones where the performance of the classifier is most reliable. Furthermore, the classifier is far better at classifying the vocal Progress Party (FrP) than e.g. the more moderate Labor Party (Ap), which indicates that parties with a clear (and polarizing) political profile are easier to classify.

While looking at overall classifier performance can be informative in itself, more insights can be gained by comparing performance for various subsets of the data. In the next section we break down the classifier predictions—correct and incorrect— along various dimensions of the ToN metadata.

5 Discussion

Figure 2 plots the F_1 scores for each party under each cabinet period, with the party average across periods shown by the dotted line. The x-axis labels the periods by the parties comprising the cabinet, first listing the party that holds the prime minister (with supporting parties in parenthesis).

For most parties the trend appears to be that party affiliation is more reliably predicted when not in cabinet. In Fig. 2, this pattern is perhaps most distinctly manifested for the Liberal Party (V), though we can also see the same trend for the Conservatives (H) and Christian Democrats (KrF). Also evidence of the same trend, the single most abrupt shift is observed for the far-right Progress Party (FrP): Of all parties, the classifier obtains the highest average F_1 for FrP while in opposition, but it plummets to the lowest observed F_1 score (0.325) when in cabinet.
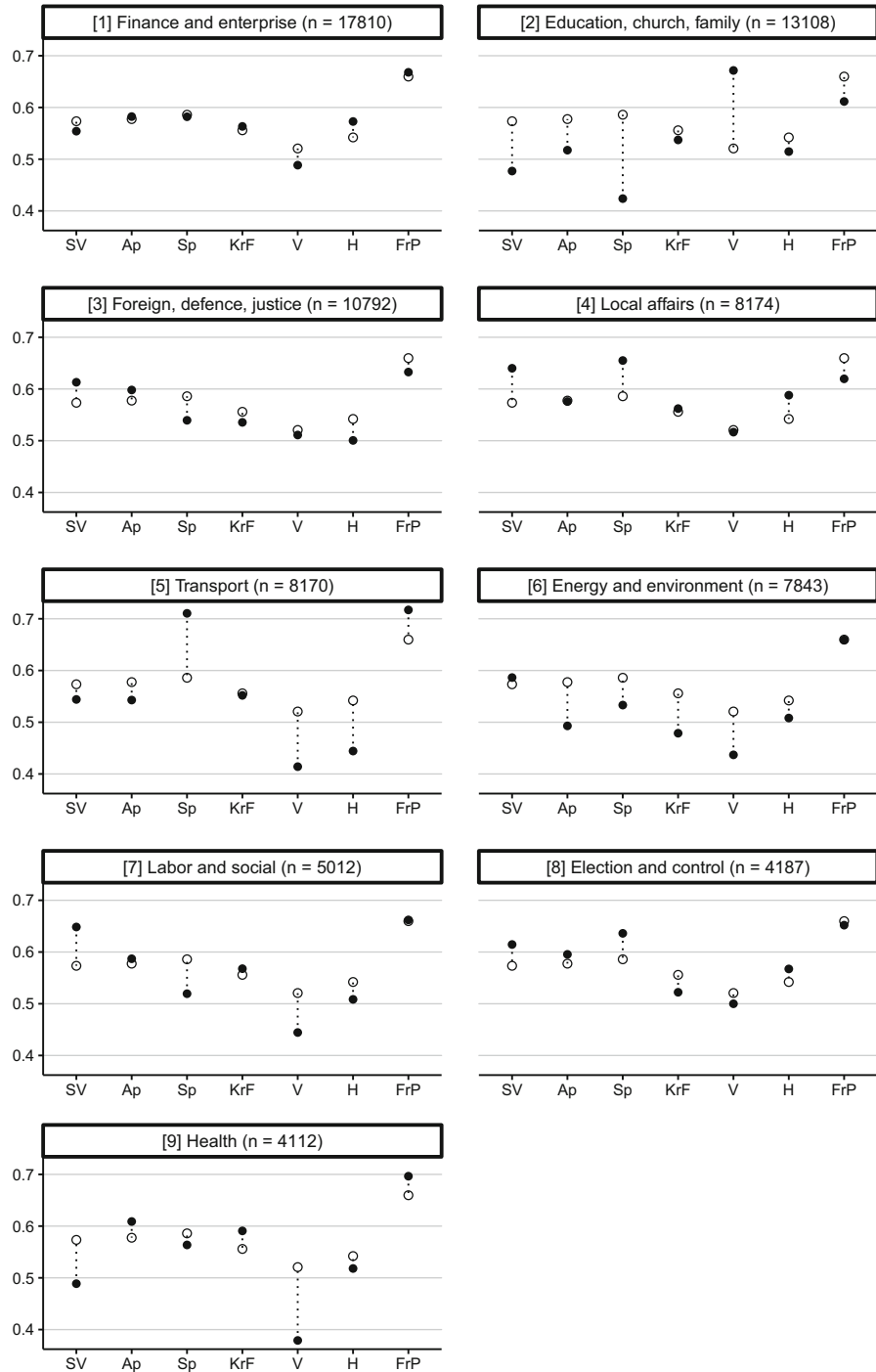
The trend that party prediction is easier when in opposition than in cabinet is less clear for the agrarian Center Party (Sp) and the Socialist Left Party (SV). For the latter the trend breaks for the last two time points, making it seem like the prediction just gets harder over time. For the Labor Party (Ap), finally, the trend is entirely reversed: The F_1 score is above its average in all three Stoltenberg (Ap) cabinets, and under the average in the Solberg (H) and two Bondevik (KrF) cabinets.

In sum, we can say that the performance of our party classifier is to a large degree driven by the role of the party under a given cabinet. This result also harmonize well with the party classification results for the Canadian House of Commons mentioned in Sect. 2.

The confusion matrices in Fig. 3 shed more light on the trends seen in Fig. 2. The horizontal rows show the predicted label distribution for speeches collected for each party while in government (left) and in opposition (right).[22] An effect that immediately stands out is that labels for all parties tend to move towards the center[23] of the political spectrum when in government: Comparing the second columns across the matrices, we see that the misclassifications towards the moderate Ap party makes a large jump when parties move from opposition into position.

---

[22] In the Solberg cabinet (the last cabinet in the sample), KrF and V are formally labeled as *support parties*, but are in this particular instance re-coded as opposition.

[23] Here, *center* does not refer to the the traditional economic left–right policy dimension, but rather to how conventional parties are. The two largest parties, H and Ap, are more conventional than for example SV, KrF and FrP.

[1] Finance and enterprise (n = 17810)

[2] Education, church, family (n = 13108)

[3] Foreign, defence, justice (n = 10792)

[4] Local affairs (n = 8174)

[5] Transport (n = 8170)

[6] Energy and environment (n = 7843)

[7] Labor and social (n = 5012)

[8] Election and control (n = 4187)

[9] Health (n = 4112)

Although this effect can be observed for all parties (including Ap itself), it gets gradually more pronounced as one moves towards the right, culminating with FrP where the misclassification rate towards Ap jumps from 8.6% when in opposition to 40.1% when in cabinet. Moreover, we see that while FrP has the lowest misclassification rate by far of all parties when in opposition, it is one of the parties with the highest error rate when in government. Overall, these trends seem to align well with the intuition that it is easier to maintain a sharp ideological profile when in opposition, and that there is a pull towards the center when in position. At the same time, for all parties we observe that when they are in opposition, their misclassification rate towards the far-right FrP party roughly doubles. The fact that this effect occurs across the full left–right spectrum would seem to indicate that the classifier to some degree also picks up on the same attack-and-defense dynamics reported for the experiments of Hirst et al. (2010).

Finally, we set out to discover how certain topics affect classifier performance across parties, by calculating F_1 scores for speeches uttered under debates led by different committees. To maximize the amount of speeches for each set, we use the ToN metadata to join together committees dealing with related topics (for instance, the "transport" committee and the "transport and communication" committee). Figure 4 shows party-wise scores for the resulting 9 subsets (solid dots) to be compared to the corresponding scores on the full data (hollow dots, same as reported in Table 2). The intuition here is that the former should be higher where the party has more distinct policies, which should translate into speeches that are easier to classify.

We find this intuition is met for several topics, perhaps most clearly in 5 (Transport). Here we see a large spike in classification accuracy for Sp and FrP, for whom this issue is notoriously crucial. Sp, a party whose voter population is in large part from rural Norway, will often call for measures that improve existing infrastructure to benefit rural and peripheral communities, rather than central areas. FrP on the other hand is a zealous advocate of the construction of highways connecting large cities (often battling the environmental concerns raised by other parties), and regards highway tolls as a central topic in their anti-tax policies. Analogously, plot 7 (Labor and Social) sees a better performance for SV, Ap and KrF, three parties who are traditionally associated with labor and social issues; the same is true for Ap, KrF and FrP in plot 9 (health). Further, for the subsets in plots 1 (Finance and Enterprise), 4 (Local Affairs), and 8 (Election and Control) classification appears to be easier in general (save the minor score drop for FrP in 4). We expect topic 1 (the largest topical subset in the data) to be a salient issue for all parties, which we see reflected in classifier performance. The same is also true for 4, traditionally a pivotal issue in Norwegian politics.

We find plots 2 (Education, Church and Family), 3 (Foreign Affairs) and 6 (Energy and Environment) to yield the most surprising results. In 2, while we do see an improvement for KrF (who is expected to hold a distinctive position on church and family issues) and V (who has a strong profile in education), we would have expected to see a

similar trend for SV. The latter sees its largest margin of improvement in 3, which can be attributed to its distinctive positions in international politics (SV is the only elected Norwegian party who is anti-Nato), while Sp's scores drop (expectedly, given their 'local' profile); Ap's improvement is somewhat surprising here, as the two largest parties (Ap and H) hold very similar positions on this topic. Finally, for 6, we see inverse trends for the two parties with the most distinct pro-environmental profile: SV goes up, while V goes down. FrP's score sees an improvement as well, which does not come as surprise, given their clear and distinctive position on areas such as oil drilling and global warming.

6 Future work

There are several avenues for future work that we would like to pursue. In terms of the ToN data set itself, we plan on enriching the available linguistic information with syntactic annotations. This can facilitate tracing relations between words in the text of the speeches, for instance helping to disambiguate the meaning of keywords when they occur as subject or object of a given verb. With this information available, we want to further develop the linguistic feature engineering for our classifier to continue improving its performance, as using syntactic information has already been proven beneficial in other text classification tasks (Johansson and Moschitti 2013). On the level of the speeches, we plan on adding automatically derived sentiment polarity scores, based on the emerging resources for Norwegian sentiment analysis currently being developed by the SANT project,[24] such as the Norwegian Review Corpus (NoReC; Velldal et al. 2017). We would also like to annotate the text with named entities, which would enable a host of new analyses of the speeches. For instance, identifying targets in fine-grained sentiment analysis, or analyzing whether MPs use speech for communicating concerns about the constituency they were elected from; this has often been analyzed through voting behavior in parliaments, but less explored with speech data in electoral systems, such as the Norwegian, where voting unity is high due to strong political parties. Unfortunately, at the time of writing, no off-the-shelf tools for Norwegian named entity recognition exist.

On the experimental side of the project, ongoing work is focused on evaluating the effects of different text representation techniques and experimental setups on Political Science research. By evaluating party classification results across increasingly more linguistically informed models, and testing with different cross-validation splits of the data, we seek to investigate how different classification workflows affect the conclusions drawn by political scientists in the kind of experiments presented above. We also plan on comparing our current setup to one that uses a distributional semantics approach to represent the speeches based on word- and document embeddings (i.e. low-dimensional dense vectors). This kind of technique has seen a surge of popularity in recent years and would allow us to model the meaning of the words in the speeches using unsupervised methods applied to external and unlabeled data, such as the vast amounts of text found in the Norwegian Newspaper Corpus Andersen (2012) and the Norwegian Web Corpus Guevara (2010). This kind of approach has been shown to improve on the state of the art of text classification tasks (Le and Mikolov 2014).

---

[24] For more information on SANT, see http://www.mn.uio.no/ifi/english/research/projects/sant/.

7 Conclusion

This paper presented the Talk of Norway data set, a collection of Norwegian parliamentary debates from 1998 to 2016. The speeches are for the first time made available to the research community together with a large array of unified metadata variables collected from a number of sources. These include detailed information on speakers, parties, cabinets, and the speeches themselves. Moreover, the actual content of the speeches is enriched with automatically obtained linguistic annotations, including language labels and sentence, token, lemma, part of speech, and morphological feature annotations. The public availability of the data set aims to enhancing comparability and replicability of research based on Norwegian parliamentary proceedings, and to encourage broader use of 'basic' morpho-syntactic analysis (as included in the ToN annotations) in support of text-based research in the computational social sciences.

Based on this data, we presented a pilot study on political party classification in the Norwegian Parliament using supervised machine learning methods. Using a combination of linguistic and non-linguistic features, our initial results are well above a majority class baseline and compare favorably to party classification results in the European Parliament, a multi-party system akin to the Norwegian one. Finally, we showcased the use of additional ToN metadata, to investigate classification results further by looking at performance and error across different cabinet periods, party roles, and topics. We find that the performance of our party classifier is to a large degree driven by institutional roles: Most parties are easier to classify when they are in opposition, while the converse is true for other parties. We inspect this effect further by looking at classification errors when parties are in position and opposition, and observe that (a) most of the misclassifications for government parties fall to the largest party in Norwegian politics (Ap), and (b) parties are in general easier to classify when they are in opposition. Looking at F_1 scores across debates led by different committees, we observe that classification performance oscillates for parties depending on the topic of the discussion. In general, scores tend to be higher when parties regard a policy area as salient, which indicates that the position–opposition dynamic is not the only driving force behind classification.

We distribute the ToN data publicly and will prepare new, extended versions regularly (see Sect. 3 for access information). We do so in the hope of enabling quantitative Political Science research on Norwegian parliamentary records and in particular seek to make possible the use of state-of-the-art basic morpho-syntactic analysis (by non-experts in NLP) in such studies, as well as to further replicability and reproducibility of results.

Appendix

The table below lists the all names of ToN metadata variables, accompanied by short explanations of the associated values, as well as short examples of their content (or simply *text* where the value is running text).

Index	Variable description	Example value	Source
1	Unique id for a speech	tale000000	♡
2	Unique id for representatives in online urls	SONS	▶
3	Unique id for representatives from original API	SONS	◆
4	First name	Sonja Irene	◆
5	Last name	Sjøli	◆
6	Full name	Sonja Irene Sjøli	◆
7	Start of speaker mandate	1997-10-01	▶
8	End of speaker mandate	2001-09-30	▶
9	Type of representative	Representant	▶
10	County of provenance	Akershus	◆
11	Position on the party list at the last election	5	▶
12	Unique id for a party	H	◆
13	Full name of the party	Høyre	◆
14	Parliament role of the party	Opposition	◇
15	Number of speaker party's seats in parliament	23	◇
16	Pet name for the current cabinet	Bondevik I	◇
17	Start date of the cabinet	1997-10-17	◇
18	Cabinet composition	Coalition	◇
19	End date of the cabinet	2000-03-16	◇
21	Date of birth	06.06.1949	◆
22	Date of death	05.12.2005	◆
20	Gender	mann	◆
23	Election cycle	1997-2001	◆
24	Total amount of seats in parliament	165	▶
25	Party seats in the upper chamber (where applicable)	6	▶

Index	Variable description	Example value	Source
26	Party seats in the lower chamber (where applicable)	17	▶
27	Speaker committees (mutliple committes separated by ";")	Sosialkomiteen ; Valgkomiteen	▶
28	Date of membership in committees (mutliple separated by ";")	21.10.1997-30.09.2001 ; 21.10.1997-30.09.2001	▶
29	Speaker role in committees (multiple roles separated by ";")	Varamedlem ; Nestleder	▶
30	Debate reference	*url*	◆
31	Unique id of the case	15631	◆
32	Title of the debate	Formalia	◆
33	Subject of the debate	Kredittmeldinga 1998	◆
34	Type of debate	voteringer	◆
35	Unique id of proposition	5060	◆
36	Proposition full text	*text*	◆
37	Underlying debate document group	innstillingssaker	◆
38	Short subject description of document	Nødhjelp til Kosovo	◆
39	Short description of decision made on the case under debate	*text*	◆
40	Any notes attached to the underlying document of debate	Forslag fra (H)	◆
41	Source representative ids (rep_id) for the case	AG ; IVK ; ON, HTN ; JIA	◆
42	Representative id (rep_id) for the chair of the case	AG ; IVK ; ON, HTN ; JIA	◆
43	Type of case	VB	◆
44	Description of decision on the case	Vedtak i samsvar med innstillingen	◆
45	For questions, the question number	1	◆
46	For questions, who asked the question (rep_id)	SONS	◆
47	For questions, who the question was asked to (rep_id)	DH	◆
48	For questions, who answered the question (rep_id)	DH	◆
49	For questions, id of ministry the answering minister is under	SHH	◆
50	For questions, title of ministry the answering minister is under	helseministeren	◆

Index	Variable description	Example value	Source
51	Id of the subjects under debate (multiple separated by ";")	185 ; 203	◆
52	Names of the subjects under debate (multiple separated by ";")	STATSBUDSJETTET ; STORTINGET	◆
53	Is this the main subject? (multiple separated by ";")	false ; false	◆
54	Id of the main subject	187	◆
55	Committee for this subject (id)	JUSTIS	◆
56	Committee for this subject (name)	Justiskomiteen	◆
57	Debate's number on this day's agenda	1	◆
58	Reference to the case (from agenda data)	(Innst. S. nr. 1 (1998-99))	◆
59	Text for the case (from agenda data)	*text*	◆
60	Type of case (from agenda data)	FORO	◆
61	Agenda number of this meeting	4	◆
62	Meeting id (to match meeting data from API)	349	◆
63	All procedures the debate has gone through (id)	KOMBEH ; DEBATT	◆
64	All procedures the debate has gone through (name)	Komitebehandling ; Debatt og vedtak	◆
65	All procedures the debate has gone through (step number)	1 ; 2, 1 ; 2 ; 3	◆
66	Publication export id for underlying case	NA ; s160204	◆
67	Publication text for underlying case	Innst. S. nr. 1 (1998-99) ; Stortingsreferat 20.10.1998	◆
68	Publication url for underlying case	*url*	◆
69	Publication type for underlying case	referat ; innstilling ; referat	◆
70	Publication subtype for underlying case	proposisjon ; storting	◆
71	Id of related cases	61779	◆
72	Type of related cases	fellesinnst ; sammekilde	◆
73	Short titles of related cases	Dokument nr. 8:92 (1997-98)	◆
74	One word keyword for debate	Statsbudsjettet	◆
75	All keywords for debate	EØS-avtalen ; Flyulykker	◆

Index	Variable description	Example value	Source
76	Language identifier	nob	♠
77	A date variable that distiguish day and night meetings	s981020a	♣
78	Speech order in a given day	1	♣
79	Timestamp of the speech	1998-10-19 22:00:00 UTC	♣
80	Parliament session	1998-1999	♣
81	Date of the speech	1998-10-20	♣
82	Role of the speaker	President	♣
83	Raw speech speech	*text*	♣

♢ Generated internally

♥ Scraped and merged from the Norwegian Parliament homepage at www.stortinget.no

◇ Handcoded by Søyland (2017)

◆ Merged from the Norwegian Parliament API at https://data.stortinget.no

♣ Merged from the https://www.holderdeord.no database, via the Norwegian Parliament API

♠ Automatically annotated with langid. py

References

Andersen, G. (Ed.). (2012). *Exploring newspaper language. Using the web to create and investigate a large corpus of modern norwegian, studies in corpus linguistics* (Vol. 49). Amsterdam: John Benjamins Publishing.

Bäck, H., & Debus, M. (2016). *Political parties, parliaments and legislative speechmaking*. Basingstoke: Palgrave Macmillian.

Beelen, K., Thijm, T. A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., et al. (2017). Digitization of the canadian parliamentary debates. *Canadian Journal of Political Science/Revue canadienne de science politique*, *50*, 1–16.

Denny, M. J., & Spirling, A. (2017). *Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it*. Unpublished manuscript.

Eggers, A. C., & Spirling, A. (2014). Electoral security as a determinant of legislator activity, 1832–1918: New data and methods for analyzing british political development. *Legislative Studies Quarterly*, *39*(4), 593–620.

Guevara, E. (2010). NoWaC: A large web-based corpus for Norwegian. In: *Proceedings of the NAACL HLT 2010 sixth web as corpus workshop*, Los Angeles, CA, USA (pp. 1–7).

Hirst, G., Riabinin, Y., & Graham, J. (2010). Party Status as a confound in the automatic classification of political speech by ideology. In: *Proceedings of the 10th international conference on statistical analysis of textual data* (pp. 731–742).

Høyland, B., Godbout, J. F., Lapponi, E., & Velldal, E. (2014). Predicting party affiliations from European Parliament debates. In: *Proceedings of the 52nd meeting of the association for computational linguistics: Workshop on language technologies and computational social science*, Baltimore, MD, USA (pp. 56–60).

Johannessen, J. B., Hagen, K., Lynum, A., & Nøklestad, A. (2012). *OBT+Stat: a combined rule-based and statistical tagger* (pp. 51–65). Amsterdam: John Benjamins.

Johansson, R., & Moschitti, A. (2013). Relational features in fine-grained opinion analysis. *Computational Linguistics*, *39*(3), 473–509.

Lapponi, E., & Søyland, M. G. (2016). *Talk of Norway*. http://hdl.handle.net/11509/123, *common language resources and technology infrastructure norway (CLARINO) Bergen Repository*.

Lapponi, E., Velldal, E., Oepen, S., & Knudsen, R. L. (2014). Off-road LAF: Encoding and processing annotations in NLP workflows. In: *Proceedings of the 9th international conference on language resources and evaluation*, Reykjavik, Iceland (pp. 4578–4583).

Lauderdale, B. E., & Herzog, A. (2016). Measuring political positions from legislative speech. *Political Analysis*, *24*(3), 374–394.

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

Le, Q.V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In: *Proceedings of the 31st international conference on machine learning, Beijing, China, proceedings of machine learning research* (Vol. 32, pp. 1188–1196).

Lui, M., & Baldwin, T. (2012). *langid.py: An off-the-shelf language identification tool*. Jeju, Republic of Korea (pp. 25–30).

Meurer, P. (2012). Corpuscle–A new corpus management platform for annotated corpora. In G. Andersen (ed) (pp. 31–50).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Proksch, S. O., & Slapin, J. B. (2015). *The politics of parliamentary debate*. Cambridge: Cambridge University Press.

Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705–722.

Søyland, M. G. (2017). Survival of the ministers: On ministerial durability in postwar Norway. *Scandinavian Political Studies*, *40*(2), 182–206.

Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote. Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 327–335).

van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2017). The debates of the European Parliament as linked open data. *Semantic Web*, *8*(2), 271–281.

Velldal, E., Øvrelid, L., Bergem, E. A., Stadsnes, C., Touileb, S., & Jørgensen, F. (2017). NoReC: The Norwegian review corpus. arXiv preprint arxiv: 1710.05370

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, *5*(1), 33–48.