

National Centre of Competence in Research (NCCR)  
*Challenges to Democracy in the 21<sup>st</sup> Century*

Working Paper No. 95

# **SIFT – A Language Technology Toolkit to Assess the Print Media Coverage of New Forms of Governance**

Bruno Wüest (Department of Political Science, University of Zurich)\*

Michael Amsler (Institute of Computational Linguistics, University of Zurich)

Gerold Schneider (Institute of Computational Linguistics, University of Zurich)

November 2017

# SIFT – A Language Technology Toolkit to Assess the Print Media Coverage of New Forms of Governance

Bruno Wüest<sup>1</sup>, Michael Amsler<sup>2</sup> & Gerold Schneider<sup>2</sup>

<sup>1</sup> Department of Political Science, University of Zurich

<sup>2</sup> Institute of Computational Linguistics, University of Zurich

## Abstract

For social scientists, it is increasingly important to explore large text collections without time-consuming human intervention. We are presenting a language technology tool kit that allows researchers of the NCCR Democracy Module 1 to extract information on various forms of governance from a comprehensive multilingual corpus. The tool kit called SIFT<sup>1</sup> allows searching for governance entities and measuring their salience, tonality, issue context and media frames. In substantial terms, our tool pipeline enables scholars of governance to extend their research focus to the previously neglected area of the public communication of democratic legitimacy and accountability of various forms of governance.

## 1 Introduction

Automated approaches to analyze unstructured text data have made tremendous progress in computational linguistics in the last decades (Jurafsky and Martin, 2009). They allow us to conduct research at a larger scale, with advanced statistical models and giving thorough quantitative support for intuition-based hypotheses. At the same time, social scientists are increasingly in need of such approaches, since the number of large, digitally available text collections is constantly growing. The research agenda in Module 1 of the NCCR Democracy (*NCCRDm1*) is no exception to this. Centered on researching the democratic legitimacy of new forms of governance, the *NCCRDm1* focuses on a multitude of governance actors, policy fields and geographic areas. The amount of available and potentially relevant digital texts for this research agenda is tremendous,

<sup>1</sup>SIFT is an acronym for Salience, Issues, Frames and Tonality.

which diminishes the feasibility of manual content analysis procedures for large scale studies. The obvious task in this case is to transfer and adapt the comprehensive computational linguistic tool set to social science (Wueest et al., 2011). Accordingly, we present in the following a pipeline of language technologies that allows the analysis of big text data for research on the various forms of democratic governance.

The denationalization and privatization of democratic governance poses formidable challenges to the traditional, territorially grounded forms of democratic authorities (Zürn, 1998). At the European and international level, new modes of governance such as supra-national and inter-governmental bodies as well as transgovernmental networks have come to supplement classic intergovernmental governance (Abbott and Snidal, 2008). At the sub-national level, regulatory agencies and public-private partnerships increasingly spread across metropolitan regions by transforming traditional regional and local state institutions (Kelleher and Lowery, 2009).

These various new forms of governance have in common that they organize political authority along functional rather than territorial lines, which also implies that they are comparatively decoupled from representative democratic control. This is why some observers often declare see challenges for the accountability of single institutions or even more general for the democratic legitimacy for the political system as a whole (Follesdal and Hix, 2006; Keohane et al., 2009). Other scholars point to formal accountability mechanisms such as governmental and parliamentary oversight as well as judicial review and highlight that they can at least partly compensate a deficit in democratic legitimacy (Lodge, 2002).

It is without controversy that the media are the most important source of information for most people in established democracies around

the world (Bennett and Entman, 2001; Strömbäck, 2008; Walgrave et al., 2008; Müller, 2014; Arnold, 2004). But, politics has not only become predominantly mediated, the search for news value by the media also increasingly dominate political processes. This mediatization means that media attention increasingly influences the timing and substance of policy making processes.

Given this importance of the media, it is surprising that extant research in political science only rarely focuses on the role of the media for two core concepts of today's established democracies: democratic legitimacy of the political system as a whole and the public accountability of single actors and institutions (but see Christmann et al., 2015; Maggetti, 2012 and Hurrelmann et al., 2009, for exceptions). The few studies that exist limit their focus to a few media outlets – most frequently, one or two outlets from the quality press are analysed – or entities – the existing studies are mainly case studies of a few actors or institutions. However, we are now in a situation where the advent of automated content methods enable the systematic analysis of large-scale text collections (Grimmer and Stewart, 2013).

The media can broadly hold two distinct functions for both democratic legitimacy and public accountability. On the one hand, the media can foster the quality of the different forums that debate and decide on legitimacy and accountability (Christmann et al., 2015; Jacobs and Schillemans, 2016). Hence, by conveying the necessary information, media provide the prime connecting mechanism between the different entities monitoring government and governance such as parliaments, courts and, of course, the citizenry and the actors and institutions that are to be held accountable. Empirical studies and theoretical accounts alike acknowledge the critical role of of freely operating media in maintaining a public sphere in which the performance of a democratic system and its actors is debated and negotiated (Habermas 2008, Hurrelmann).

On the other hand, the media has traditionally been conceived as an accountability forum in its own right, too (Graber, 2003). The media is more than a 'watchdog' in this perspective, since there are direct consequences if media uncover a violation of norms or standards (Christmann et al., 2015). Reputation is a indispensable resource for political actors in established democracies (Car-

penter, 2010), so the limelight in the media can sometimes be enough to provoke corrective measures such as resignations of the responsible policy makers or reversion of policies.

There are, however, several limitations to both the supporting and the self-supporting function of the media for democratic legitimacy and public accountability. Most importantly, the media can never completely substitute formal processes of democratic control (Bovens, 2007b). Obviously, journalists and publishers are not elected. And they also do not possess any direct means to sanction or remove the rascals among the political elite. As welcome it is that media do regularly pay critical attention to governance processes, they can only encourage the formation of an informed public opinion (O'Donnell, 1998). It is then only the threat of electoral punishment or popular protest that mounts the ultimate pressure on governance actors to explain, justify and – if necessary – correct their conduct.

Further, the strive of the media to hold the powerful accountable may increase the risk of dysfunctional outcomes. With growing media attention, politicians are, for instance, inclined to care more about blame-avoidance strategies and a positive self-presentation than about genuine accountability (Papadopoulos, 2010).

In the following, we present *SWIFT*, a comprehensive corpus and a largely automated language technology pipeline, which enable political scientists to assess these questions. The paper begins by presenting our operationalization of indicators that allow the reliable measurement of governance accountability in a large-scale text analysis. Subsequently, we will describe the software pipeline and language technologies necessary to implement the operationalization, before we present a case study highlighting the feasibility of our approach.

## **2 Measuring media coverage on public accountability and democratic legitimacy**

So far, mediatized accountability mechanisms have only been dealt with in conceptual elaborations or comparative case studies that entailed manual content analyses (Maggetti, 2012; Coglianese and Howard, 1998; Gerhards and Roose, 2007). Although these contributions are theoretically insightful and empirically rich, their focus on a narrow set of actors, geographical units

or media sources always faces the necessity to justify why their cases provide more than just idiosyncratic evidence. An automated large-scale analysis, by contrast, helps to achieve a more broad analytical support on the question whether and how media scrutinize on the accountability of governance processes.

## 2.1 Sample

The anchor of the analysis is a large gazetteer of pre-defined entities related to governance (see Figure 1). These entities refer to actors (collective actors and individuals), policy fields and regulation such as treaties or directives. At the moment, a comprehensive gazetteer of entities for 3257 queries is integrated in the document retrieval. The entities cover a large variety of forms of governance: transgovernmental networks, independent as well as private regulatory authorities, metropolitan bodies, supranational parliaments and international environmental governance outcomes. A gazetteer of entities needs to be hand-crafted for every new form of governance that is to be studied<sup>2</sup>.

Table 1: Entities of interest by research project

Project	Description of entities	Queries
IP1	Metropolitan bodies on public transport and (local) economic promotion	26
IP2	International organizations, supranational parliaments and general international regulatory activity (directives, regulations etc.)	140
IP3	Regulatory activity on international environmental governance (treaties, conventions, protocols etc.)	216
IP4	Transnational networks (committees, institutions, associations, persons etc.)	238
IP5	Independent and private regulatory authorities (organisations, authorities, firms etc.)	57

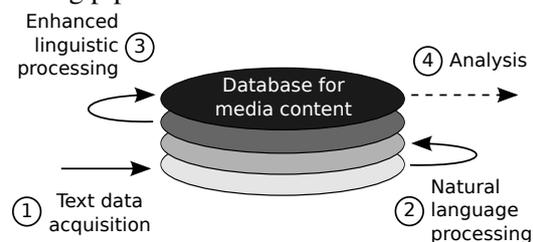
Most of the queries are additionally translated into the three languages (English, French and German), for which the analysis is set up. The *European Economic Area*, which is a relevant entity for the project IP2, for example, is searched as *Espace économique européen* in French sources and *Europäischer Wirtschaftsraum* in German sources. Also, where applicable, the queries have been extended with more general keywords such as abbreviations to a wide query. In English newspaper articles, the *European Economic Area* often is

<sup>2</sup>We are especially thankful to Michael Buess, Reto Wüest, Jofre Rocabert, Christian Ewert, Michael Strebel and Steffen Mohreberg who established, refined and delivered these gazetteers.

only mentioned with its abbreviation, *EEA*, which is why this is included into the search query as well.

The gazetteer of queries that map to governance entities is the only manual input of *SIFT* that is indispensable. After the governance entities of interests have been operationalized into search queries, the basic data generation therefore runs fully automated. Saliency, Tonality and Issues can be generated automatically as well, whereas researches can optionally provide manual input for the topic models that calculate the issues. Only for the frame analysis, an active learning framework is implemented, which means that manual input in the form of labeled documents is required.

Figure 1: Stylized workflow in the language processing pipeline



In a first step, a comprehensive corpus of the following newspapers, newswires and online sources is established by retrieving all articles for the keyword gazetteer via API accesses to media content databases such as Lexis Nexis (① in Figure 1).

- *Quality*: Frankfurter Allgemeine, Süddeutsche Zeitung, Welt, Tageszeitung (Germany); Figaro, Le Monde (France); Neue Zürcher Zeitung, Le Temps (Switzerland); The Guardian, London Times, Independent (UK)
- *Tabloid/Freesheets*: Bild (Germany); Aujourd'hui en France, 20 minutes (France); Blick, Le Matin, 20 Minuten (Switzerland); Daily Mail, Daily Mirror, Metro (UK)
- *Magazines*: Spiegel, Stern, Zeit (Germany); Nouvel Observateur, L'Express (France); Weltwoche, Wochenzeitung, L'Hedbo (Switzerland); New Statesman, Spectator, Economist (UK)
- *Regional*: Berliner Zeitung, Stuttgarter Zeitung, Stuttgarter Nachrichten (Germany); Le Parisien, Le Progrès (France); Tagesanzeiger, Berner Zeitung (Switzerland); London Evening Standard, City A.M., Birmingham Mail, Birmingham Post (UK)
- *Online sources*: Spiegel Online (Germany), Figaro Online, Le Monde Online (France); 20 Minuten Online (Switzerland); BBC News Online (UK)

- *Newswires*: Associated Press, Agence France Presse, Deutsche Presse Agentur, BBC Monitoring, Europolitics, ENP Newswire, AWP

Since different types of media systems (Hallin and Mancini, 2004), as well as different types of media (Strömbäck and Kaid, 2008) possibly cover governance in different ways, the media sources are sampled so that there is a balanced set of outlets in our four countries (Switzerland, Germany, France, and United Kingdom). From each type of outlets, the outlet with the highest circulation (or website visits in the case of the online sources) was chosen. As far as possible, we also cover other potential variations such as different ideological leanings. In addition to these country-specific media samples, we also include a range of internationally operating newswires, which provide us with information on the general reporting on governance in disregard of specific journalistic cultures in single media outlets.

Subsequently, an additional layer of data consisting of the compressed documents along with initial meta-data (source, date-of-publication etc.) is added to the database (② in Figure 1). At a third stage, we employ a full natural language processing chain, which includes morphological analysis, tagging, lemmatizing, and dependency parsing (③). Finally, a fourth layer of enhanced linguistic analysis – named entity recognition, coreference resolution, sentiment detection, opinion mining and topic modeling – is implemented to calculate the indicators of interest we will discuss in the following (④).

## 2.2 Saliency

The attention media pay to specific forms of governance is the obvious starting point of the data analysis. No media attention is the worst case in terms of questions regarding the public accountability and legitimization of governance, since ‘quiet politics’ (Culpepper, 2010) implies absent interest by the public and, correspondingly, high leverage for particular interests and dishonest conduct in governance processes. The first necessary measure in the *SWIFT* analysis therefore is saliency, defined as the visibility of specific governance entities in the media.

## 2.3 Tonality

A second crucial information on governance entities is the media’s evaluation of these governance entities in terms of tonality. The tone of

media reports on governance entities yields useful results if changes in tonality signify reactions to events on the governance processes under concern (Maggetti, 2012). For example, if a corruption scandal shakes a governance actor, we expect media reports to shift to a negative tone. This also implies that tonality has to be measured at the level of the specific entity and not at the level of text documents as a whole.

We expect that the saliency, tonality and issues in media reports on governance entities already reveal crucial evidence on whether and how media coverage entails mechanisms of accountability. More precisely, if media adjust their attention according to events related to specific governance entities, if media react to failure with a negative tone – and to success with a positive tone –, and if the media really cover the issues related to the area of responsibility of these governance entities, media coverage actually constitute an accountability forum’ for this governance entity (Bovens, 2007a).

## 2.4 Issues

Governance entities may draw media attention for different reasons, but not all are relevant for the research objective. If a sports magazine reported on the passion of the head of the Swiss Financial Markets Supervisory Agency (Finma) for windsurfing (which arguably is true), hardly any political analyst would deem this information relevant to understand financial market regulation in Switzerland. More generally, evidence on the thematic context in which governance entities are mentioned is key to assess whether media reports on specific entities are actually covering the governance processes of interest. In addition, an issue analysis caters to the growing interest in the study of the relationship between media and politics from an agenda-setting perspective (Vliegenthart et al., 2016).

## 2.5 Frames

What is still missing is information on the reasons why the media report on governance entities, i.e. which interpretations and problem definitions journalists convey to the reader. To this aim, we additionally conduct an analysis of the frames in the media (Entman et al., 2009). To start from the classical definitions, frames are *schemata of interpretation* (Goffman 1974, 21), or, very similarly, *interpretive packages* (Gamson

and Modigliani 1989, 2) that are constructed in order to influence how people think about ‘reality’. In the context of mega, we further specify frames as generic schemata of interpretation that refer to the main source of democratic legitimacy of governance entities as it is reported in the text documents. We are aware that many statements may not refer to any source of democratic legitimacy at all, however, for the study of media accountability of governance, evidence on the democratic legitimacy of governance is of the utmost significance.

To be precise, we separate input-oriented legitimacy frames from output-oriented ones. The concepts of output- and input-oriented legitimacy have their origin in Abraham Lincoln’s famous dictum about democracy requiring government *by the people, of the people and for the people* (Gopnik, 2007). Hence, for democratic governance to work properly, the participation of the potentially affected citizens, the representation of the citizen’s interests, and the authorities’ effectiveness is required. The labels for these concepts, in contrast, have diffused from systems theories, particularly from the work of Easton (1965), into the broader literature. Accordingly, the citizens’ support, be it directly through elections or indirectly through shared identities or diffuse system legitimization, constitute the input into the political system (see Schmidt 2013). Input legitimacy is thus present if media refer to participatory aspects, popular support and democratic accountability in general, or public interest representation with regards to governance processes. Output, on the other hand, is defined as the governments’ decisions and actions (Easton 1965). Output legitimacy, accordingly, refers to the efficiency, effectiveness and the due process of governance.

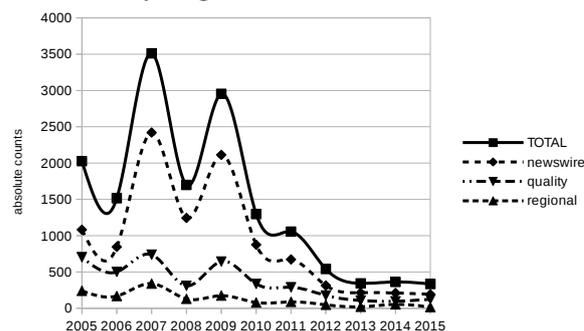
### 3 Empirical implementation and case study on the media coverage of the Kyoto protocol

#### 3.1 Saliency

For this case study, we measure saliency as the occurrence of articles in the media coverage across the timeline. Although it is a simple measurement, saliency reveals on the one hand important insights about the presence of the respective entity and, on the other hand, offers the opportunity to closely scrutinize the content near the peaks.

As can be seen in Figure 2, the visibility manifests itself with two clear peaks in 2007 and

Figure 2: Saliency of articles referring to Kyoto Protocol (only English articles; n=15,849)



2009. A closer investigation of the respective coverage points towards the importance of the Fourth Assessment Report of the United Nations Intergovernmental Panel on Climate Change (IPCC) in 2007 and the 2009 United Nations Climate Change Conference in Copenhagen which triggered each an increased attendance to the subject. We also see that most coverage stems from newswires, while print media are more selective in their reporting. Also, people who only have access to regional newspapers, we only observe a modest increase for the reporting of the main events,

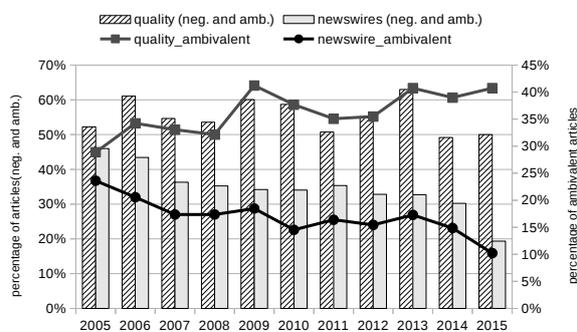
In addition to this straightforward operationalization of saliency, the number of articles found can additionally be weighted by document length. As an alternative to the number of documents, saliency can also be expressed with the number of keyword hits, although we only recommend this for clearly delimitable entities.

#### 3.2 Tonality

To measure tonality in the media coverage, we apply a linguistically informed sentiment analysis system, similar to (Taboada et al., 2011). More precisely, we use domain-adapted and extended lexical semantic resources in combination with a dependency parse that infers the syntactic structure of the sentence. The latter allows for composing tonality from singular words into chunks of the sentence (e.g., noun phrases) which in turn are then taken into account to derive the tonality analysis for the core of the sentences, i.e., the main verbs. A more detailed description can be found in (Klenner et al., 2014). The system used for this task was evaluated in another case study for the tracking of coverage tonality which yielded good results (see Wueest et al. (2014)). Although the tonality can be derived for singular entities in the

given texts, we aggregate in this case study on the document level since the thematical focus is narrowed by the data acquisition process (i.e. the query to the media databases).

Figure 3: Comparison of negative and ambivalent tonality between media types *quality* and *newswire*



In Figure 3 we focus on the difference of tonality regarding the level of critique considering different media types: the bars show the percentage of articles of negative and ambivalent tonality (ordinate on the left-hand side). It is obvious that the coverage in quality papers is much more critical than in the newswire articles. The lines show the percentage of only the ambivalent articles (ordinate on the right-hand side) which reveals that the difference between the two media types mainly stems from the much higher percentage of ambivalent articles, that is, articles which discuss the topic under different perspectives, considering chances and risks as well as progress and failure in the implementation process.

### 3.3 Issues

#### 3.3.1 Topic Models

We apply structural topic models (STM) (Roberts et al., forthcoming) to explore the thematic context in which the media writes about governance. STM is a data-driven technique, which allows us to estimate document probabilities for latent variables, called topics. STM builds on the Latent Dirichlet Allocation, a hierarchical mixed-membership model in which the document-topic and word-topic probabilities have a common prior drawn from a Dirichlet distribution (Blei et al., 2003). One of the major innovations of STM is that the prior distribution of topics (i.e. topic prevalence) can be influenced by covariates. In the following analysis, we use the newspaper names and a b-spline with 10 degrees of freedom on a monthly

trend variable to control for unwanted linguistic differences across news outlets and over time. In addition, we apply a parametric evaluation of the most probable topic-word vectors in order to find the optimal number of topics. To this purpose, we use word2vec (Mikolov and Dean, 2013), which learns and aggregates term similarities through a shallow neural network process. These term similarities can then be used to compare topic coherence and exclusiveness across different topic models. For the Kyoto protocol corpus, word2vec suggests a granularity of 19 for a candidate range of 3 to 20 topics.

Figure 4: Dynamics of selected topics

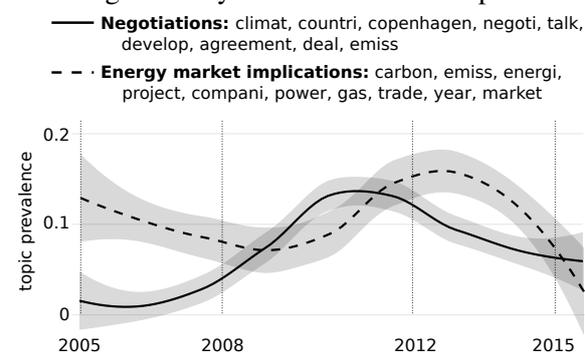


Figure 4 shows the trends in the prevalence of two especially meaningful topics over time. In addition, the list of the 10 most probable word stems for each topic is listed.

The first topic summarizes the different negotiation rounds on the Kyoto protocol, most notably the first commitment period from 2008 until 2012 with the Copenhagen summit in 2009 as key event. Reports on the different negotiations accordingly peak in this period. The second topic, in contrast, highlights the consequences of the Kyoto protocol on the energy markets and emission trading. Quite intuitively, this topic becomes most prevalent in the aftermath the big policy decisions from 2011 on.

#### 3.3.2 Collocations

Issues can also be detected by various keyword algorithms. While they are less suitable to detect broad topics than topic modeling approaches, they reliably report keywords and key phrases. They typically perform better than topic modeling in situations where the amount of data is relatively small.

As a keyword detection method which does not need reference corpora, collocations measures can

Table 2: Top 50 collocations from the Kyoto protocol articles with  $O/E$

<i>Collocation</i>	<i>Rank</i>	<i>O/E</i>	<i>Collocation</i>	<i>Rank</i>	<i>O/E</i>
jacques chirac	1	24'335.1	de boer	26	5'386.8
christiana figueres	2	16'616.7	ban ki	27	5'081.6
connie hedegaard	3	15'802.7	stephen harper	28	4'883.7
nfiona harveyguardian	4	15'555.1	nobel peace	29	4'832.4
arnold schwarzenegger	5	14'979.9	intergovernmental panel	30	4'815.7
saudi arabia	6	13'053.9	solar panels	31	4'549.3
angela merkel	7	12'940.7	todd stern	32	4'359.9
ki moon	8	12'108.1	nfull textthe	33	4'341.0
nfiona harvey	9	9'849.3	njohn vidal	34	4'210.8
inconvenient truth	10	9'286.6	debt relief	35	4'080.0
prince albert	11	9'248.3	peace prize	36	4'074.4
polar bears	12	8'498.5	west midlands	37	3'943.5
downing street	13	7'893.5	liberal democrat	38	3'903.8
nmichael mccarthy	14	7'842.0	light bulbs	39	3'899.2
njohn vidalguardian	15	7'777.6	ed miliband	40	3'824.5
margaret beckett	16	7'282.2	sir nicholas	41	3'781.7
vladimir putin	17	7'219.0	al gore	42	3'569.0
tar sands	18	6'635.2	road map	43	3'461.2
ndavid adam	19	6'545.0	hadley centre	44	3'434.1
chris huhne	20	6'366.0	ice sheets	45	3'315.3
yvo de	21	6'252.6	pm bst	46	3'086.1
mrs beckett	22	6'131.9	kevin rudd	47	3'023.1
rural affairs	23	6'116.5	pm gmt	48	2'941.2
environmentally friendly	24	5'876.5	german chancellor	49	2'804.9
east anglia	25	5'387.8	pm aest	50	2'585.6

be used to detect multi-word terms. Technical terms, also in less technical genres such as newspapers, are often multi-word concepts. According to Master (2003, 2), the multi-word units consisting of nouns are particularly important, as they are directly linked to neologisms and new concepts: “A noun compound is a grammatical structure in which nouns are linked together to indicate a new concept”. In information retrieval, multi-word expressions often offer a better unit of analysis than single words (Schwartz and Ungar, 2015, 84)

There exist a variety of collocation measures, see Evert (2009) and Pecina (2009) for a comprehensive overview. We use information-theoretic collocation measures based on observed frequency (O) of word pairs and expected frequency (E) of word pairs if the words in a text were randomly shuffled, i.e. the drawing of each of the two words in a collocation pair were an independent event. The corresponding measures, for example O divided by E (known as  $O/E$  or O over E) are easy to interpret.

Applied to the articles on the Kyoto protocol,  $O/E$  delivers the ranking given in Table 2. We only list the top 50 entries here, but the first several hundred entries contain many true positives, which can be easily filtered by manually sifting the lists.  $O/E$  has a well-known tendency to over-represent rare events (see e.g. Smith (2002)),

which can be counteracted by using measures such as  $O^2/E$  (which scored best in Bartsch and Evert (2014)),  $O^3/E$  or measures based on significance testing, such as  $T$ -score. The top collocations for  $O^3/E$  are given as an example in Table 3. These lists can also be automatically filtered for named entities (such as place names or persons) or other key concepts related to the entity under concern.

### 3.4 Frames

While we have focused on purely empirical data for the other indicators, we will report on the evaluation for the methodological approach used for the framing prediction.

#### 3.4.1 Annotation

In contrast to the measurements for the other indicators which are derived generically, we rely on annotated data for the framing. More precisely, we annotate the frames using the brat annotation tool (Stenetorp et al., 2012). For the whole corpus three annotators build valuable training data of 18'548 frames<sup>3</sup>. After an intensive training phase, inter-annotator agreement is constantly high (micro-averaged F1-scores for fine-grained frame categories that range between 0.66

<sup>3</sup>We thank Michelle Amman, Anna Sigrist and Anna-Lina Müller for their excellent work for the manual annotation data.

Table 3: Top 50 collocations from the Kyoto protocol articles with  $O^3/E$

Collocation	Rank	$O^3/E$	Collocation	Rank	$O^3/E$
climate change	1	$8.66e + 10$	united states	26	$1.38e + 09$
kyoto protocol	2	$2.29e + 10$	de boer	27	$1.34e + 09$
per cent	3	$2.05e + 10$	long term	28	$1.25e + 09$
global warming	4	$1.33e + 10$	tony blair	29	$1.17e + 09$
greenhouse gas	5	$1.04e + 10$	the us	30	$1.13e + 09$
prime minister	6	$9.08e + 09$	to be	31	$1.09e + 09$
of the	7	$7.16e + 09$	it is	32	$1.04e + 09$
legally binding	8	$5.51e + 09$	barack obama	33	$1.02e + 09$
don t	9	$4.29e + 09$	george bush	34	$9.79e + 08$
greenhouse gases	10	$4.19e + 09$	saudi arabia	35	$9.73e + 08$
developing countries	11	$3.36e + 09$	rather than	36	$9.67e + 08$
carbon dioxide	12	$3.00e + 09$	power stations	37	$9.58e + 08$
fossil fuels	13	$2.96e + 09$	nmichael mccarthy	38	$9.50e + 08$
in the	14	$2.86e + 09$	mr blair	39	$8.82e + 08$
such as	15	$2.41e + 09$	have been	40	$8.66e + 08$
intergovernmental panel	16	$2.25e + 09$	world s	41	$7.83e + 08$
will be	17	$2.24e + 09$	framework convention	42	$7.75e + 08$
has been	18	$1.97e + 09$	european union	43	$7.55e + 08$
the world	19	$1.90e + 09$	downing street	44	$7.44e + 08$
ki moon	20	$1.89e + 09$	ndavid adam	45	$7.26e + 08$
tar sands	21	$1.84e + 09$	trading scheme	46	$7.08e + 08$
gordon brown	22	$1.77e + 09$	the kyoto	47	$7.03e + 08$
white house	23	$1.75e + 09$	al gore	48	$6.91e + 08$
more than	24	$1.68e + 09$	fossil fuel	49	$6.89e + 08$
gas emissions	25	$1.50e + 09$	angela merkel	50	$6.67e + 08$

for 23 documents during and 0.71 for 5 documents at the start of the annotation).

### 3.4.2 Detection

First attempts have revealed that the recognition of frames is a challenging task, especially since we encounter a skewed distribution in the data (i.e. paragraphs containing frames vs. paragraphs without frames). Additionally, the distribution between the different types of frames is skewed as well (i.e. some frames occur much more than others), which then again complicates the task for a supervised learning approach.

Hence we design the automated recognition of frames as follows: in the first stage we apply a model that tries to detect paragraphs with mentions of democratic legitimacy (as a generic category). Second, we differentiate then between input frames output frames and throughput frames and apply in parallel the fine-grained frame classification.

Our baseline consists of a paragraph-based bag-of-words (BoW) model. For this baseline, we report an accuracy of 0.85 and a F1-score of 0.58 (macro-averaged) for the binary classification of the first stage (whether there is a frame in a paragraph or not). While the overall accuracy is encouraging, especially the recall for frames is low due to the data imbalance. To counteract this un-

desired outcome, we apply different techniques: firstly, we introduce on the one hand concept features to the model. For these features, a small set of seed words ( $\sim 15$  words) for every frame category was extended using word embeddings (using the word2vec algorithm (Mikolov and Dean, 2013) and the gensim framework (Řehůřek and Sojka, 2010)) to  $\sim 250$ -400 concept feature words, which were then added for the classification task. Additionally, we downsample the paragraphs not containing frames. Lastly, we push the classifier in the ensemble scenario towards recall at cost of precision (system “rb”). In this way, we increase the recall of the frames up to 0.76.

Table 4 shows precision, recall, and F1 scores for the individual categories.

Table 4: Evaluation for the detection of frames in paragraphs on unseen testdata

	Frame			No Frame			ALL
	Prec.	Rec.	F1	Prec.	Rec.	F1	Acc.
EN: BoW	0.32	0.22	0.26	0.90	0.93	0.91	0.85
EN: rb	0.19	0.76	0.30	0.94	0.52	0.67	0.55
FR: BoW	0.33	0.32	0.33	0.85	0.85	0.85	0.75
FR: rb	0.31	0.64	0.42	0.89	0.67	0.77	0.67
DE: BoW	0.30	0.27	0.29	0.89	0.90	0.89	0.81
DE: rb	0.23	0.70	0.35	0.93	0.62	0.74	0.63

For the frame detection in paragraphs in the first stage, we apply the following preprocessing steps. First, we embed the sentences of the paragraphs using word embeddings derived from the full text

corpus of the project for all three languages. Note that we filter here for content words only as there is evidence that this enhances the process of sentence content representation (cf. (Lau and Baldwin, 2016)). Second, we compare the embedded sentence to quantized representations (i.e., clustered to 10 centroids per frame category) of the aforementioned concept features as well as we calculate raw counts and ratios for those features. Based on this comparisons, we produce scorings for each frame category on the fine-grained level as well as on the coarse-grained level. After the preprocessing step, we train an soft voting ensemble classifier, consisting of a random forest classifier, a logistic regression, and a naive bayes estimator. To furthermore counteract the data imbalance leading to a low recall for the frames, we down-sample the paragraphs not containing a frame according to their distribution in the feature space (see “rb” version in Table 4)

It must also be noted that the baseline approach which is bag-of-words based is far less transferable to other projects since it makes heavy use of domain adaptation via the used vocabulary. However, since the requirements of the involved projects are different, we allow for shifting the decision boundary of the classifier ex-post to reach a higher overall accuracy while maintaining an acceptable recall for the frames. Table 5 shows an example for such a threshold based scenario where we reach an overall accuracy above 72% while keeping the recall for frames above 0.5.

Table 5: Evaluation for the detection of frames using a threshold

	Frame			No Frame			ALL	
	Prec.	Rec.	F1	Prec.	Rec.	F1	Acc.	TH
EN: rbT	0.26	0.61	0.36	0.93	0.75	0.83	0.73	0.7
FR: rbT	0.34	0.53	0.42	0.88	0.76	0.82	0.72	0.8
DE: rbT	0.26	0.54	0.35	0.91	0.75	0.82	0.72	0.8

### 3.4.3 Classifying Frames

While the system predicts frames on the unit of a paragraph, we identified during the annotation the slice of the text which was the core of the frame, i.e., the text passage that triggered the annotation of the presence of the specific legitimacy frame subclass. Since this piece of information contains the most secure information of how such a frame is textually manifest, we rely on this spans to create the seeds for the aforementioned lexicon generation process. To corroborate the robustness of the system to differentiate on the coarse-

grained and fine-grained level we evaluate on this textual passages only. Table 6 shows that the system performs above the accuracy of 0.8 on coarse-grained level and between 0.58 and 0.62 on the fine-grained level. Note that we report weighted average values.

Table 6: Evaluation for coarse-grained and fine-grained frame classification on frame passages

	Coarse-grained				Fine-grained			
	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.
EN	0.81	0.80	0.81	0.80	0.59	0.58	0.58	0.58
FR	0.83	0.83	0.83	0.83	0.61	0.62	0.61	0.62
DE	0.83	0.83	0.83	0.83	0.60	0.61	0.60	0.61

In contrast, the general prediction task is carried out on the unit of the paragraph, not the frame text passage itself. Nevertheless, the identification of the latter is crucial to derive the resources on which the system relies.

### 3.4.4 Generalization and Transferability

To test the transferability of the framing detection and classification component, we evaluated the performance separately on another annotated dataset which was sampled from entities of governance which were not included in the dataset we developed and trained on. This dataset contains another 1’516 paragraphs which were annotated on the fine-grained level<sup>4</sup>. We report here an overall accuracy for the frame detection of 0.72. If we predict directly the coarse-grained level for paragraphs, we score at 0.63. For the fine-grained level, the accuracy is 0.54.

## 4 Conclusion

The corpus and language technology pipeline *SIFT* was built around the assumption that the salience, tonality and issues in media reports on governance entities reveal crucial evidence on whether and how media support or diminish democratic legitimacy and public accountability of governance. In this paper, we discussed the methodological decisions and the technical implementation necessary to acquire and analyse about 600 governance entities in roughly 4.5 mio. news articles. The case study on the Kyoto protocol has shown, that we can draw meaningful conclusions from the results provided by *SIFT*. At least under certain conditions, media adjust their attention according to events related to specific governance

<sup>4</sup>We thank Rolf Badat for his great work for the manual annotation as well the evaluation of the analysis.

entities; react to failure with a negative tone – and to success with a positive tone –; and convey the issues and frames related to the area of responsibility of the respective governance entity.

We have shown how news documents can be mined to uncover the media’s role for public accountability and democratic legitimacy. More substantially oriented research is necessary to make full use of *SIFT* in order to further disentangle this role. Possible research questions, to name only a few, are whether and to what effect media coverage varies across different governance entities; whether increased monitoring by the media has the intended, positive effect on legitimacy and accountability; or whether media are merely an aggregator and amplifier for other legitimacy and accountability forums or constitute an own, independent forum.

## Acknowledgments

Research for this paper was funded by the NCCR Democracy at the University of Zurich.

## References

- K. W. Abbott and D. Snidal. 2008. The governance triangle: regulatory standards institutions and the shadow of the state. In Walter Mattli and Ngaire Woods, editors, *The Politics of Global Regulation*. Princeton University Press, Princeton, NJ.
- R. D. Arnold. 2004. *Congress, the press, and political accountability*. Princeton University Press, Princeton, NJ.
- S. Bartsch and S. Evert. 2014. Towards a firthian notion of collocation. In A. Abel and L. Lemnitzer, editors, *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, number 2/2014 in OPAL – Online publizierte Arbeiten zur Linguistik, pages 48–61. Institut für Deutsche Sprache, Mannheim.
- W. L. Bennett and R. M. Entman. 2001. *Mediated Politics: Communication in the Future of Democracy*. Cambridge University Press, Cambridge, MA.
- D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning*, 3:993–1022.
- M. Bovens. 2007a. Analysing and assessing accountability: a conceptual framework. *European Law Journal*, 13(4):447–468.
- M. Bovens. 2007b. New forms of accountability and eu-governance. *Comparative European Politics*, 5:104–120.
- D. Carpenter. 2010. *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton University Press, Princeton, NJ.
- A. Christmann, D. Kübler, K. Hasler, F. Marcinkowski, and C. Schemer. 2015. New regionalism – not too complex for the media watchdog: Media reporting and citizens’ perception of democratic legitimacy in four european metropolitan areas. *Urban Affairs Review*, 51:676–707.
- C. Coglianese and M. Howard. 1998. Getting the message out: Regulatory policy and the press. *Politics*, 3:39–55.
- P. D. Culpepper. 2010. *Quiet Politics and Business Power. Corporate Control in Europe and Japan*. Cambridge University Press, Cambridge, MA.
- D. Easton. 1965. *A Systems Analysis of Political Life*. Wiley, New York, NY.
- R. M. Entman, J. Matthes, and L. Pellicano. 2009. Framing politics in the news: Nature, sources and effects. In K. Wahl-Jorgensen and T. Hanitzsch, editors, *Handbook of Journalism Studies*. Routledge, London.
- S. Evert. 2009. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook, article 58*. Mouton de Gruyter, Berlin.
- A. Follesdal and S. Hix. 2006. Why there is a democratic deficit in the EU: A response to Majone and Moravcsik. *JCMS: Journal of Common Market Studies*, 44:533–562.
- W. A. Gamson and A. Modigliani. 1989. Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, 95(1):1–37.
- A. Offerhaus Gerhards, J. and J. Roose. 2007. Die öffentliche zuschreibung von verantwortung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 59:105–124.
- E. Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harper & Row, New York, NY.
- A. Gopnik, 2007. *Angels and Ages. Lincoln’s language and its legacy*. The New Yorker, May 28, 2007.
- D. Graber. 2003. The media and democracy: Beyond myths and stereotypes. *Annual Review of Political Science*, 6:139–60.
- J. Grimmer and B. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.
- D. C. Hallin and P. Mancini. 2004. *Comparing Media Systems: three Models of Media and Politics*. Cambridge University Press, Cambridge, MA.

- A. Hurrelmann, Z. Krell-Laluhová, F. Nullmeier, S. Schneider, and A. Wiesner. 2009. Why the democratic nation-state is still legitimate: A study of media discourses. *European Journal of Political Research*, 48:483–515.
- S. Jacobs and T. Schillemans. 2016. Media and public accountability: typology and exploration. *Policy & Politics*, 44:23–40.
- D. Jurafsky and J.H. Martin. 2009. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, Upper Saddle River, NJ.
- C. A. Kelleher and D. Lowery. 2009. Central city size, metropolitan institutions and political participation. *British Journal of Political Science*, 39(1):59–92.
- R. O. Keohane, S. Macedo, and A. Moravcsik. 2009. Democracy-enhancing multilateralism. *International Organization*, 63(1):1–31.
- M. Klenner, M. Amsler, and N. Hollenstein. 2014. Verb polarity frames: a new resource and its application in target-specific polarity classification. In *Proceedings of KONVENS 2014*, pages 106–115.
- J.H. Lau and T. Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany, August. Association for Computational Linguistics.
- M. Lodge. 2002. The wrong type of regulation? regulatory failure and the railways in Britain and Germany. *Journal of Public Policy*, 22:271–297.
- M. Maggetti. 2012. The media accountability of independent regulatory agencies. *European Political Science Review*, 4(3):385–408.
- P. Master. 2003. Noun compounds and compressed definitions. *English Teaching Forum*, 41(3), July.
- K. Chen G. Corrado Mikolov, T. and J. Dean, 2013. *Efficient Estimation of Word Representations in Vector Space*. CoRR, abs/1301.3781.
- L. Müller. 2014. *Patterns of Media Performance: Comparing the Contribution of Mass Media to Democratic Quality Worldwide*. Palgrave Macmillan, Houndmills, UK.
- G. A. O’Donnell. 1998. Horizontal accountability in new democracies. *Journal of Democracy*, 9:112–126.
- Y. Papadopoulos. 2010. Accountability and multi-level governance: More accountability, less democracy? *West European Politics*, 33:1030–1049.
- P. Pecina. 2009. *Lexical Association Measures: Collocation Extraction*, volume 4 of *Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Charles University in Prague, Praha, Czech Republic.
- Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. <http://is.muni.cz/publication/884893/en>.
- M. E. Roberts, B. M. Stewart, and M. E. Airoidi. forthcoming. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*.
- V. A. Schmidt. 2013. Democracy and legitimacy in the european union revisited: Input, output and ‘throughput’. *Political Studies*, 61:1467–9248.
- H. A. Schwartz and L. H. Ungar. 2015. Data-driven content analysis of social media: A systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, 659(1):78–94, apr.
- D. A. Smith. 2002. Detecting events with date and place information in unstructured text. In *Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries*, pages 191–196, Portland, OR.
- P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April. Association for Computational Linguistics.
- J. Strömbäck and L. L. Kaid. 2008. *The Handbook of Election News Coverage Around the World*. Routledge, New York, NY.
- J. Strömbäck. 2008. Four phases of mediatization: An analysis of the mediatization of politics. *The International Journal of Press/Politics*, 13(3):228–246.
- M. Taboada, J. Brooke M., Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.
- R. Vliegthart, S. Walgrave, F. R. Baumgartner, S. Bevan, C. Breunig, S. Brouard, L. C. Bonafont, E. Grossman, W. Jennings, P. B. Mortensen, A. M. Palau, P. Sciarini, and A. Tresch. 2016. Do the media set the parliamentary agenda? a comparative study in seven countries. *European Journal of Political Research*, 55(2):283–301.
- S. Walgrave, S. Soroka, and M. Nuytemans. 2008. The mass media’s political agenda-setting power: A longitudinal analysis of media, parliament, and government in Belgium (1993 to 2000). *Comparative Political Studies*, 41:814–836.

- B. Wueest, S. Clematide, A. Bünzli, D. Laupper, and T. Frey. 2011. Electoral campaigns and relation mining: Extracting semantic network data from swiss newspaper articles. *Journal of Information Technology and Politics*, 8(4):444–463.
- B. Wueest, G. Schneider, and M. Amsler. 2014. Measuring the public accountability of new modes of governance. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 38–43, Baltimore, MD, USA, June. Association for Computational Linguistics.
- M. Zürn. 1998. *Regieren Jenseits des Nationalstaats. Globalisierung und Denationalisierung als Chance*. Suhrkamp, Frankfurt a. M., DE.