How to do politics with words: Investigating speech acts in parliamentary debates

Ines Reinig, Ines Rehbein, Simone Paolo Ponzetto
Data and Web Science Group
University of Mannheim, Germany
{ines.reinig, rehbein, ponzetto}@uni-mannheim.de

Abstract

This paper presents a new perspective on framing through the lens of speech acts and investigates how politicians make use of different pragmatic speech act functions in political debates. To that end, we create a new resource of German parliamentary debates, annotated with fine-grained speech act types. Our hierarchical annotation scheme distinguishes between cooperation and conflict communication, further structured into six subtypes, such as informative, declarative or argumentative-critical speech acts, with 14 fine-grained classes at the lowest level. We present classification baselines on our new data and show that the fine-grained classes in our schema can be predicted with an avg. F1 of around 82.0%. We then use our classifier to explore the use of speech acts in a large corpus of parliamentary debates over a time span from 2003–2023.

Keywords: Speech acts, Language in institutions, Political Text Analysis

1. Introduction

Being able to determine the rhetorical function of speech acts in political communication has great potential for studies in the area of political text analysis, as it will allow us to investigate how different political players interact with each other in order to achieve their goals or legitimise their actions.

Many works in the political sciences have studied this question, based on qualitative content analysis (Reyes, 2011; Abrahamyan, 2020; Kashiha, 2022). However, the major disadvantage of qualitative frameworks such as Critical Discourse Analysis (Fairclough, 1995) and other types of content analysis is that they do not scale well to large data and thus do not allow us to study communicative behaviour over longer periods of time and in different contextual settings.

To address these limitations, there has been an increasing interest in semi-automatic and fully automated methods for the study of research questions from the political and social sciences, using text as data. We follow this line of research, with the goal of investigating rhetorical strategies in parliamentary discourse through the lens of speech acts.

Our contributions are the following:

- We present a new annotation scheme for the analysis of speech acts in political text.
- We create a new dataset of German parliamentary debates annotated for speech acts, with more than 12,900 annotated instances.
- We provide strong baselines for the following two sub-tasks on our data: (i) utterance segmentation and (ii) speech act classification.
- We showcase how our new schema and resources can be used for political text analysis.

We make the code, models and annotation guidelines available in the follow GitHub repository: `https://github.com/umanlp/speechact`.

2. Related Work

Discourse units and their pragmatic functions have been the topic of many research projects interested in discourse analysis. Prominent examples are Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) and its application (Carlson et al., 2001), or the somewhat shallower view on discourse relations taken in the Penn Discourse Treebank (Miltsakaki et al., 2004; Prasad et al., 2008). While those look at the relation *between* different discourse units, other work has focussed on individual utterances and their pragmatic function in the discourse.

Many studies have looked at speech acts in academic texts, most noteworthy the work on Argumentative Zoning (Teufel et al., 1999), or models of interaction in academic discourse (Hyland, 2005; Hyland and Tse, 2004). Other work has been focussing on spontaneous speech and dialogue structure (see, e.g., the VerbMobil, DAMSL, and SWBD-DAMSL projects) (Alexandersson et al., 1995; Allen and Core, 1997; Jurafsky et al., 1997; Stolcke et al., 2000). Bunt et al. (2019) aimed at developing a unified and generic schema for dialogue act tagging while others have targeted task-oriented dialogue in applications (Stolcke et al., 2000; Mesgar et al., 2023; Moghe et al., 2023; Zhang et al., 2020, 2021).

In our work, we look at speech acts in argumentative text, specifically, in political communication.

LREC-COLING 2024, pages 8287–8300

While there is some overlap between the "classical" speech acts of Austin (1962) and Searle (1969) and our work (see §3), our annotation scheme offers a taxonomy for classifying institutional communication into more fine-grained classes that describe the rhetorical functions of speech acts in parliamentary debates.

2.1. Speech acts in political text

Many studies have used speech act theory for political text analysis, often based on a qualitative analysis of a small number of manually coded texts (Akinkurolere, 2020; S. Abiola, 2021; Gani et al., 2020; Ramanathan et al., 2020). Artés (2013) uses speech acts to investigate if and when Spanish politicians keep their promises. Schueler and Marx (2023) present a dataset of Dutch COVID-19 press conference releases, annotated for speech acts. They use the annotations to investigate changes in the distribution of speech acts over time and relate them to real-world events during the COVID-19 crisis. Zhang et al. (2017) analyse question-answer sessions of the UK parliament, using an unsupervised approach, and automatically identify different rhetorical functions of questions in their data. They observe a correlation between question type and the career status of the speaker.[1] Their approach is tailored toward the genre of interactive question-answer sessions while we look at longer political speeches held in parliament.

Subramanian et al. (2019) automatically classify speech acts in media releases and speeches from two Australian political parties during the 2016 Australian federal election campaign. Their schema is similar to the traditional speech acts of Searle (1969) but includes some refinements inspired by the political science literature. In their definition of the task, the authors jointly predict the speech act category and the target party (i.e., the political party that is the target of the utterance; either LABOUR, LIBERAL, or NONE) for each utterance.

Most relevant for us is the work of Kondratenko et al. (2020) who present a qualitative analysis of parliamentary discourse from the Ukrainian parliament (2004–2019), focussing on the communication strategies and tactics employed by different politicians (see §3). We apply and extend their schema, resulting in a hierarchical taxonomy that is more fine-grained than the ones described above. Our schema is tailored towards the communicative functions used in parliamentary debates. Similar to Subramanian et al. (2019), we present experiments for speech act segmentation and classification.

[1]"Younger opposition members tend to contribute more condemnatory questions compared to older members [...], who disproportionately favor concede/accept questions". (Zhang et al., 2017, p.1566)

3. Speech act schema

In the paper, we use the term *speech act* to refer to individual discourse units that serve one or more specific, pragmatic functions in communication. Psycholinguistic evidence suggests that information encoded in matrix clauses is perceived as more salient by the reader compared to information included in sentential complements (Ruppenhofer and Schneevogt, 2014). We, therefore, define speech acts on the sentence level as matrix clauses, including sentential complements; coordinated sentences are split into multiple speech-act units.[2]

Our work is based on the linguo-pragmatic schema developed by Kondratenko et al. (2020) to study interactions in institutional communication and, in particular, parliamentary debates. Their hierarchical schema classifies interactions into two classes, i.e., *cooperative* speech-acts and those used in *conflict* (Table 1). On the next level, the cooperative interactions are further split into *regulatory, informational* and *consolidation* communication strategies while conflict communication is subdivided into *declarative, confrontational* and *argumentative-critical* interactions. Below, we provide a short description of each category.[3]

3.1. Cooperative interactions

Cooperative–regulatory The regulatory category includes all speech acts used for moderation of the debate, managing interposed questions, or announcing the beginning and end of a debate.

Cooperative–informational This type of interaction informs the parliamentarians about new and operational information, e.g., about the results of a working group or committee. While the main objective is to inform members of the parliament, this interaction can also be used to convince parliamentarians of the speaker's point of view.

Cooperative–consolidating Consolidation interactions can be further divided into (i) *the pursuit of the common unity of people and politicians* and into (ii) *expressing support to specific participants in parliamentary discourse and relevant bills* (Kondratenko et al., 2020, p.23). The second subclass includes interactions where the speakers express their support for a specific political position.

[2]Detailed annotation instructions for segmentation are included in the annotation guidelines, pp. 7-10.

[3]The annotation guidelines are available in the GitHub repository: https://github.com/umanlp/speechact/blob/main/Guidelines_speechacts.en.pdf.

Cooperation		
regulatory	*informative*	*consolidating*
Macro	Report	Request
Expressive	Question	Support
Subjective-statement		

Conflict		
declarative	*confrontational*	*argumentative-critical*
Self-representation	Rhetorical-question	Demand
		Rejection
Promise	Accusation	
		Bad-outcome
Subjective-statement		

Table 1: Our linguo-pragmatic taxonomy for speech acts in parliamentary discourse, based on and extending Kondratenko et al. (2020).

3.2. Conflict communication

Conflict communication is characterised by “a lack of willingness for a common result” and verbal aggression (Kondratenko et al., 2020, p.24).

Conflict–declarative The first subclass, *declarative* interactions, can be described as the positive self-presentation of a political actor or party or the negative representation of others, where “politicians comment on their own position and demonstrate image characteristics” (ibid.). Declarative speech acts are typically not directed towards other members of parliament but to a larger audience that follows the debate in the media. Therefore, declarative speech acts can be seen as performances directed towards potential voters and often use emphatic expressions and emotional speech.

Conflict–confrontational speech acts are characterised by an unconstructive communicative behaviour. This involves a frequent use of accusations, blaming and attacks of the political opponent and the use of offensive language.

Conflict–argumentative-critical This class can be described as “rational speech behaviour, directed not on the agreement with the interlocutor but on criticising them or denying their words” (Kondratenko et al., 2020, p.26). In contrast to the confrontational interaction type, the speaker follows the rules of argumentation, basing her arguments on facts, reasoning and inference.

3.3. Our fine-grained speech acts

We expand the taxonomy developed by (Kondratenko et al., 2020) by adding 14 fine-grained subclasses for the different communication types (see Table 1).[4] Below, we highlight subtle differences between some of the subclasses. Table 10 in the Appendix provides additional examples for each of the 14 classes. More detailed information can be found in the annotation guidelines.

Promise vs. Self-representation Our annotation scheme distinguishes between PROMISE and SELF-REPRESENTATION, with the first class comprising cases where the speaker commits to working towards the achievement of a certain goal or the performance of an action. Examples are “We stand by small farms, and we want to preserve their structures.” (see Figure 1). SELF-REPRESENTATION, on the other hand, does not include any concrete proposals or actions but focusses on rather abstract values and ideals (“We are on the side of freedom and justice.”).

Self-representation vs. Support While SELF-REPRESENTATION focusses on the speaker’s *own* achievements or values, SUPPORT emphasizes support for the goals or proposals put forward by *another* person or party (“We support the Greens’ motion.”) and is thus considered as Cooperation–consolidating.

Bad-outcome vs Accusation We use the class BAD-OUTCOME for the negative consequences of a political goal, action, or event. Note that this does not necessarily imply an accusation (“Emissions are a health hazard.”). The primary function of ACCUSATION, on the other hand, is not to argue but to insult and/or apportion blame (“You have betrayed the interests of the people.”). For instances that include both an accusation and a description of negative consequences, we assign both labels (“The Greens’ policy is jeopardizing the security of supply.”).

Subjective-statement This class includes evaluation and subjective statements, which amount to a large proportion of sentences in our data. These instances can fulfill different functions in the discourse and can thus not be unambiguously classi-

[4] We also experimented with a label for irony, sarcasm and humour but, as we only encountered few instances in our data, excuded this class from our experiments.

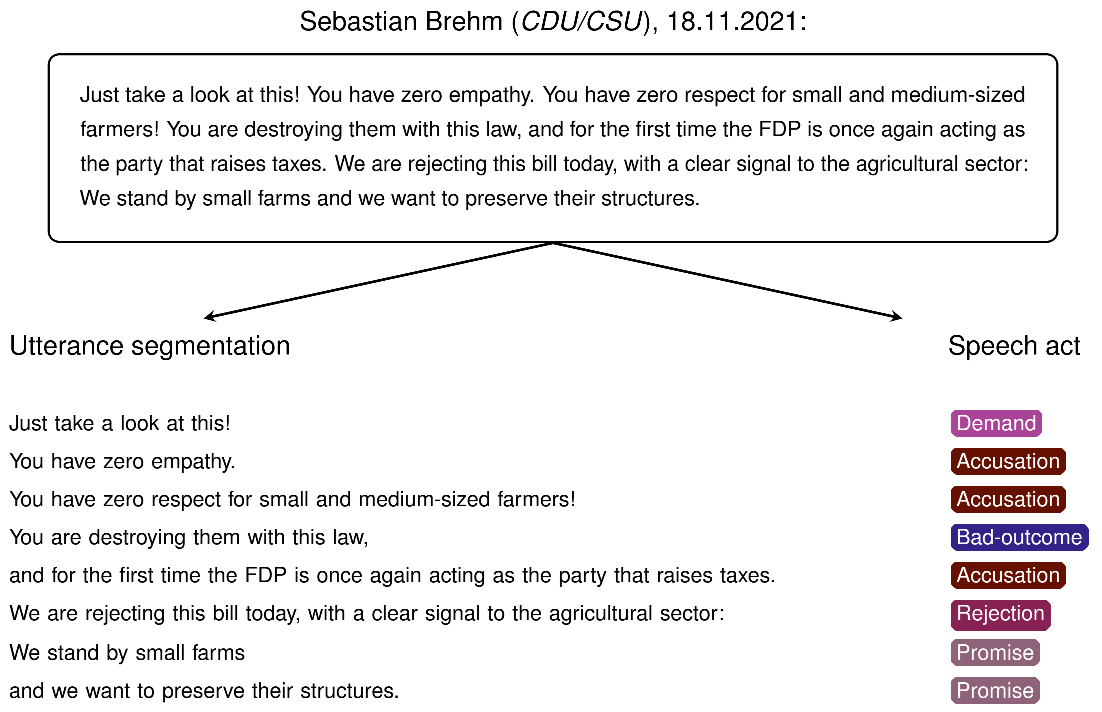


Figure 1: Translated exempt from a parliamentary debate and its corresponding speech act annotation. Speech acts are color coded following Table 1. The original German text is provided in Figure 5.

fied as either cooperation or conflict (see Table 1 and the more detailed examples in the guidelines).

4. Speech act annotation

Data In our work, we use a large corpus of parliamentary debates from the German Bundestag, spanning a time period from Jan 2003 to Sep 2023. The first part of the data (until 2018) comes from the Parlspeech corpus[5] (Rauh and Schwalbach, 2020), the last four years have been downloaded from the open data service of the German Bundestag (BT).[6] For all data, we add metadata about the speakers, their party affiliation and the date of the speech.

Annotation The data sampled for annotation includes 250 speeches from the 19th legislative term (2017-2021). The annotation has been conducted by four coders, two of them students from communication science/linguistics and two experts in linguistic annotation, both with a degree in computational linguistics. We started with the hierarchical schema of Kondratenko et al. (2020) described in §3 and defined fine-grained subclasses for each of the six pragmatic communication types (Table 1). We then used a data-driven approach to expand our label set, so that all utterances could be classified. We had regular meetings where we discussed and refined our schema and guidelines. For annotation, we used the browser-based INCEpTION platform (Klie et al., 2018). The annotation is a multi-class, multi-label task and includes two steps:

1. segmenting utterances into speech acts
2. assigning one or more speech act label(s).

Each document has been annotated independently by two coders and disagreements between the two coders have then been resolved by a third coder (one of the expert annotators). After the annotation phase has been completed, extensive consistency checks have been carried out where we searched for inconsistent annotations and corrected them.[7]

Inter-annotator agreement (IAA) As our data includes multiclass, multilabel annotations, we report Krippendorff's alpha with MASI distance (Passonneau, 2006). Table 2 shows α scores for different samples in our data. During the annotation of the first samples we were still improving our schema and adapting the guidelines, which is reflected in the lower α scores. IAA increased substantially during the annotation process. One of the student annotators, A4, only took part in the annotation of the first samples and then left the project. When

[5] From https://doi.org/10.7910/DVN/L4OAKN.
[6] https://www.bundestag.de/services/opendata.
[7] More details on the consistency checks are provided in Section A.2 in the Appendix.

Sample	Krippendorff's α
Sample 1	0.559
Sample 2	0.583
Sample 3	0.602
Sample 4	0.642
Sample 5	0.669

Table 2: IAA (Krippendorff's α with MASI distance) for all 4 coders for different samples in our data.

excluding A4's annotations, IAA for *all samples* increases from 0.612 to 0.643 α.[8] We also observed some variation for individual coder pairs, with an α in the range of 0.62 (A1-A2) to over 0.70 (A2-A3).

Given the difficulty of the task and the subjectivity involved in the decisions for some of the labels, we consider this agreement as satisfactory. The major sources for disagreements between the coders were (1) different segmentation decisions made by the annotators, mostly regarding the in- or exclusion of conjunctions; (2) the distincion between rhetorical and regular questions;[9] (3) different interpretations of the utterance, as illustrated below:

(1) We will continue to support this alliance. (A1: PROMISE, A2: SUPPORT)

(2) We stand for this alliance as a guarantor of peace and freedom. (A1: SELF-REPRESENTATION, A2: SUPPORT)

(3) After the financial crisis, we rebuilt the European financial architecture. (A1: SELF-REPRESENTATION, A2: REPORT)

This ambiguity challenges the concept of "ground truth" for our data, as we would argue that none of the labels in Examples 1–3 are incorrect. Instead, they either capture different aspects of the utterance meaning or describe different interpretations of the same utterance.

5. Experiments

Figure 1 shows an exempt from our data, illustrating the two sub-tasks: automatically segmenting text into speech acts (§5.1) and determining the label(s) of a given speech act (§5.2). For both tasks, we conduct model selection on the dev set.

5.1. Speech act segmentation

For segmentation, we extract gold speech act boundaries from our annotated data and encode the beginning and end of a speech act segment with the BIO schema, using **O** to mark tokens that are not part of a speech act (mostly vocatives[10]). We then train a BERT-based token classification model to predict the speech act boundaries.[11]

Testset	Prec	Rec	F1	Support
B	0.92	0.93	0.92	2,238
I	0.98	0.99	0.99	29,421
O	0.95	0.70	0.81	2,180
Macro-avg	0.95	0.88	0.91	33,839

Table 3: Token-based results for automatic segmentation on the test set (precision, recall, Micro-F1 per class and averaged Macro-F1; all results averaged over 3 runs with different initialisations).

Table 3 shows results well over 0.90% Macro-F1 for automatic segmentation, evaluated on the token level and averaged over three independent runs with different initialisations.

Error analysis An inspection of the data showed that some of the errors are due to inconsistencies in the annotations where the coders did not strictly follow the guidelines for speech act segmentation, mostly concerning the treatment of conjunctions in sentence coordination. Another frequent error type concerns vocatives that should have been tagged as **O** but have sometimes been overlooked and included in the speech act. Here the model often predicts the correct boundaries and gets punished in the evaluation.

5.2. Speech act classification

We view speech act classification as a multi-class, multi-label text classification task where each utterance is assigned zero, one or more speech act labels, and compare different models on our data.

Majority class As a naive baseline, we define a classifier that only predicts the majority class.

TF-IDF and SVM For this baseline, we remove punctuation, apply stemming and encode the utterance using TF-IDF scores. These serve as input features to a Support Vector Machine classifier with a linear kernel. The SVM classifier is adapted to the multi-label scenario by fitting one binary classifier for each class.

[8] We carefully checked all of A4's annotations during the consistency checks to assure the quality of the data.

[9] We decided to merge the two question classes and leave the distinction between the labels for future work.

[10] We exclude vocatives from the annotation span because they are typically used as openers in parliamentary speeches and do not contain relevant pragmatic information ("Madam President, ladies and gentlemen!").

[11] For details on experimental settings, hyperparameters and model selection, see Section A.3 in the Appendix.

Setting	Utterance	Input to BERT	Token Type IDs
BERT	['Das', 'wissen', 'auch', 'Sie', '.']	['Das', 'wissen', 'auch', 'Sie', '.']	[0, 0, 0, 0, 0]
BERT$_{context}$	['Das', 'wissen', 'auch', 'Sie', '.']	['Es', 'scheitert', 'am', 'Speicher-', 'und', 'Flächenproblem', '.', **'Das', 'wissen', 'auch', 'Sie', '.'**, 'Wie', 'Frau', 'Weisgerber', 'erwähnt', 'hat', ',', 'ist', 'der', 'Anteil', 'Deutschlands', 'an', 'den', 'weltweit', 'einsparbaren', 'CO', '2', '-Emissionen', 'marginal', '.']	[0, 0, 0, 0, 0, 0, 0, **1, 1, 1, 1, 1**, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Table 4: Input format for our BERT model w/o and with context (engl. translation: “You know that, too)”.

BERT We initialise the BERT baseline with a German pretrained language model.[12] We use a maximum length of 256 subwords and pass the utterance through the BERT language model. A classification head on top of BERT then performs multi-label classification. The model's logits are passed through a sigmoid function to obtain probability-like scores for each class. We use a threshold of 0.5 to decide whether to predict a label.[13] The loss function is defined as the binary cross entropy between the gold labels and the predicted probabilities.[14]

BERT with textual context We hypothesize that the context surrounding the utterance can provide useful information to the classifier and test a modified version of BERT which additionally encodes the previous and the following sentence (see Table 4). The target utterance is indicated by setting the `token_type_ids` for the corresponding tokens to `1` and the ones for the context tokens to `0`.

Larger BERT We also evaluate a large BERT model to investigate the role of model size for speech act classification.[15]

Evaluation Table 5 shows results for the different models on the dev and test sets. Micro-F1 for the majority baseline is around 44% while the SVM outperforms this baseline by ca. 15%. The BERT-base models show a superior performance, yielding F1 scores of 77% (w/o context) and 78% (with context). The large BERT model further improves F1 by around 4 percentage points, pushing results close to 82%. Results for individual classes are in the range of 45 to 98% (Table 6).

Model	Results (%) Micro-F1	EMR
majority	46.47	43.82
TFIDF-SVM	62.97	52.80
BERT	79.99$\pm$ 0.390	76.28$\pm$ 0.313
BERT$_{context}$	80.18$\pm$ 0.468	76.33$\pm$ 0.553
BERT-large$_{context}$	84.01$\pm$ 0.730	80.70$\pm$ 0.701
majority	43.83	40.70
TFIDF-SVM	58.95	48.22
BERT	77.01$\pm$ 0.219	72.08$\pm$ 0.415
BERT$_{context}$	78.14$\pm$ 0.304	73.00$\pm$ 0.462
BERT-large$_{context}$	81.96$\pm$ 0.009	77.36$\pm$ 0.127

Table 5: Results on the BT dev (top) and test sets (bottom). EMR: Exact Match Ratio (proportion of utterances where all labels have been classified correctly. Note that this strict metric ignores partially correct predictions). BERT results are averaged over 5 runs (base models) and 3 runs (large model).

6. Analysis

We now apply our best model to predict speech acts in the entire dataset of parliamentary debates. We use the predictions to investigate two hypotheses regarding the rhetoric strategies of members of the government and parliamentary opposition in our data.

H1: We expect that members of the parliamentary opposition will use more speech acts of type *conflict* than members of the government.

H2: We expect that members of the government will produce more *consolidating* utterances than members of the opposition.

We investigate both hypotheses by comparing proportions of the entire dataset of German parliamentary debates from 2003 to 2023, as described below. Please note that our data thus represents the whole population (all existing debates given in that time period) and *not* samples taken from that population. This means that the differences in proportions that we observe in the data reflect real differences in the population, as we can exclude sampling errors, which makes a significance test unnecessary.

[12] We use the following model from the Transformers library: `https://huggingface.co/dbmdz/bert-base-german-cased`.

[13] We also experimented with a threshold optimized for each class in an n-fold cross-validation on the train data, but could not improve results over the 0.5 threshold.

[14] More precisely, we use the `BCEWithLogitsLoss` implemented in the PyTorch libary: `https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html`.

[15] `https://huggingface.co/deepset/gbert-large`.

Class	F1	Support
Subj-statement	84.50$_{\pm\,0.458}$	1,413
Report	80.65$_{\pm\,0.534}$	686
Macro	84.69$_{\pm\,1.972}$	250
Accusation	65.11$_{\pm\,0.393}$	216
Demand	80.54$_{\pm\,0.977}$	206
Request	82.35$_{\pm\,0.614}$	146
Expressive	93.58$_{\pm\,0.263}$	142
Question-All	98.21$_{\pm\,1.112}$	92
Bad-outcome	58.18$_{\pm\,2.327}$	43
Promise	67.93$_{\pm\,4.302}$	38
Rejection	67.56$_{\pm\,3.816}$	24
Self-representation	45.07$_{\pm\,3.316}$	22
Support	78.33$_{\pm\,2.887}$	9

Table 6: F1-scores per class for BERT-large$_{context}$ on the BT test set (Support: no. of instances in test).

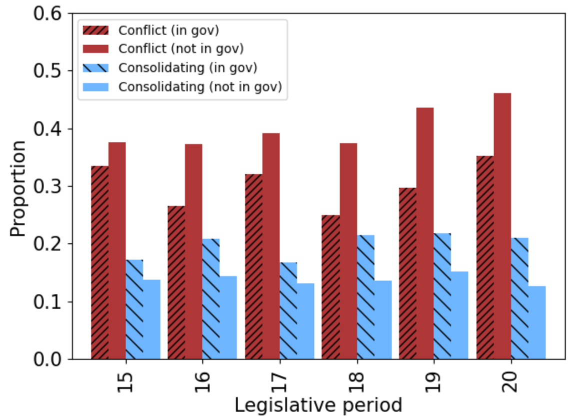


Figure 2: Proportion of conflict and consolidating speech acts by legislation term, for parties in government (hatched) and opposition parties (plain).

6.1. Data

Filtering for relevant labels For this analysis, we filter out utterances that have not been assigned a label and labels that are irrelevant for our analysis. This includes the regulatory speech acts and also SUBJECTIVE-STATEMENT which can be attributed to either cooperative or conflict communication. Finally, we filter out the merged class QUESTION-ALL which includes both informative and rhetorical questions and consider only the 9 labels that can be unambiguously interpreted as either cooperative or conflict communication. Table 7 provides an overview of the data used in our analysis, before (`All speech acts`) and after filtering (`Relevant`).

6.2. H1: Conflict communication

Setup Our fine-grained labels can be mapped to the binary classes *conflict* and *cooperation* in the taxonomy (see Table 1). For a coarse view on the discourse strategies employed by political parties, we categorize all utterances in the `Relevant` subset as belonging to *conflict*, *cooperation* or *both*. The third category *both* results from our multi-label setup. Consider the following example: an utterance is assigned the labels `['REPORT', 'ACCUSATION']` by the classifier. REPORT belongs to the category *cooperation* while ACCUSATION belongs to *conflict*. Such an utterance thus belongs to the category *both*. This category is very small however, with only 0.7% of the `Relevant` utterances. The categories *cooperation* and *conflict* include 64.8% and 34.5% of the `Relevant` utterances, respectively.[16]

[16]The high percentage of *cooperation* can be attributed to the fact that half of the `Relevant` utterances have the label REPORT.

Let $S = \{conflict, cooperation, both\}$ be our set of coarse categories and n^t_{conflict} be the number of utterances with the coarse label *conflict* for a given legislative term t. We analyze confrontational communication by means of the proportion of *conflict* utterances:

$$p^t_{\text{conflict}} = \frac{n^t_{\text{conflict}}}{\sum_{s \in S} n^t_s} \quad , 0 \leq p^t_{\text{conflict}} \leq 1 \qquad (1)$$

A proportion p^t_{conflict} close to 1.0 would indicate that nearly all utterances are of type *conflict*.

Results Figure 2 (brown bars) shows the proportion of *conflict* speech acts p^t_{conflict} for parties in government and for opposition parties for each legislative term t. The figure indicates that, as expected, *conflict* communication is more prevalent amongst members of the opposition than members of the government across legislative terms.

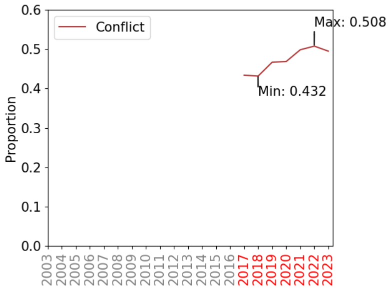
Figure 3 provides a more detailed picture of the parties' use of *conflict* speech acts over the years. In particular for the fractions of *CDU/CSU*, *Gruene* and *SPD*, we can see that the change from opposition to government corresponds to a decrease in the use of *conflict* speech acts while changing from government to opposition increases the proportion of *conflict* speech acts. This is in line with our expectations (**H1**), thus providing evidence for the validity of our manual annotations and automatic predictions.

6.3. H2: Consolidating speech acts

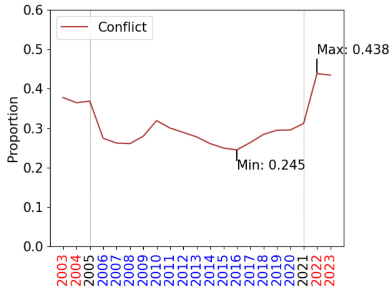
Setup In order to investigate **H2**, we focus on the *consolidating* speech act types. Here we exclude the cooperative-informative speech act class to test for robustness and to assure that our results are not solely based on the REPORT class which dominates the cooperative speech acts. We now use

	AfD	SPD	Gruene	FDP	CDU/CSU	PDS/Linke	Total
`All speech acts`	212,015	1,336,477	757,911	588,980	1,689,299	603,824	5,188,506
`Relevant`	78,415	470,987	275,631	210,211	605,059	236,343	1,876,646
`Single label`	63,417	436,292	243,471	187,923	560,676	210,449	1,702,228

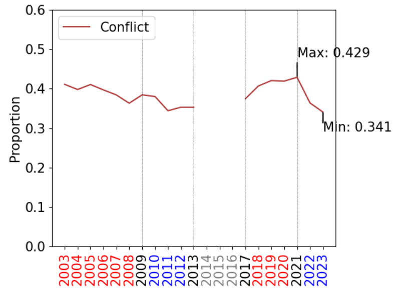
Table 7: Number of speech acts per party before and after filtering.



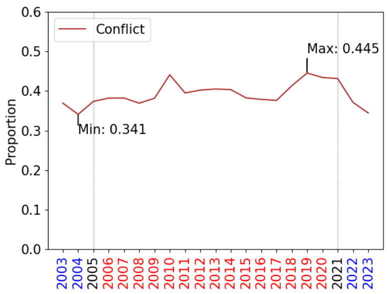
(a) AfD



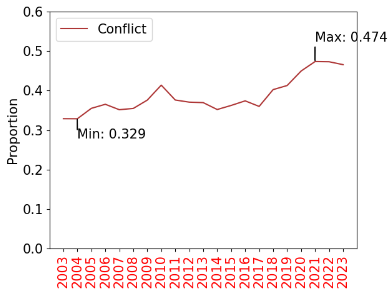
(b) CDU/CSU



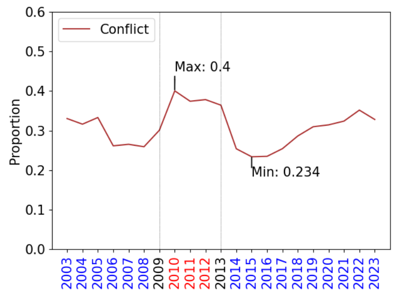
(c) FDP



(d) Gruene



(e) PDS/Linke



(f) SPD

Figure 3: Proportion of conflict speech acts over the years for each party. Labels on the x-axis are color coded: red corresponds to years in opposition, blue to years in government and grey to years not in parliament. Black colored years and black dotted vertical lines mark years in which the party went either into opposition or into government. Note that the y-scale always ranges from 0 to 0.6.

the `Single label` subset for our analysis. In analogy to p^t_{conflict}, we define p^t_{consol} as the proportion of speech acts that have a label of type *consolidating*.

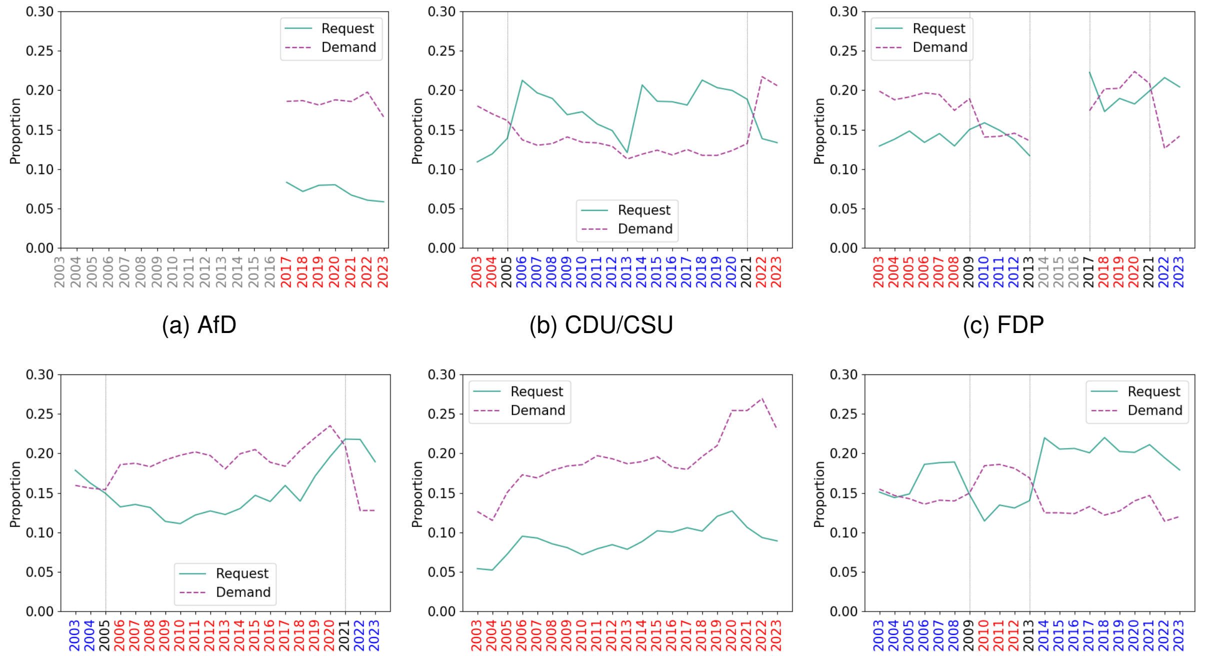
Results Figure 2 (blue bars) shows the proportion of *cooperative-consolidating* speech acts amongst members of the opposition and members of the government across legislative terms. The proportion of *consolidating* speech acts is consistently higher for members of the government than for members of the opposition, across legislative terms, thus providing evidence for **H2**.

6.4. How do parties persue their goals?

DEMAND and REQUEST constitute two substantially different ways for political actors to persue their goals. We define DEMAND as a confrontational, insistent and peremptory request, asked authoritatively as if by right. A REQUEST, on the other hand, is defined as a consolidating speech act that invites others to work together toward a common goal. We hypothesize that members of the opposition will express their goals more often using DEMAND than the cooperative REQUEST. Likewise, we expect members of the government to prefer REQUEST over DEMAND.

Setup We compare the proportions of DEMAND and REQUEST speech acts for all parties over the years, in compliance with their respective status as either member of the government or opposition. Here we consider only instances that have been assigned a single label (either DEMAND or REQUEST).

Results Figure 4 shows the proportions of the two speech acts for each party over time. Overall, we observe a more frequent use of the consolidating REQUEST when the party is in government and a higher use of DEMAND as soon as the party ends up in opposition. The graphs 4a and 4e show that for the two parties that have never been in government, *AfD* and *PDS/Linke*, the proportions



(a) AfD (b) CDU/CSU (c) FDP

(d) Gruene (e) PDS/Linke (f) SPD

Figure 4: Proportion of demand and request speech acts over the years for each party. Labels on the x-axis are color coded in the same way as in Figure 3. Note that the y-scale always ranges from 0 to 0.3.

of DEMAND are higher than REQUEST for the entire time period.

Summary Our analysis confirms our expectations regarding the different rhetorical strategies used by political actors in government and in opposition. This initial exploration suggests that our annotation schema and speech act classifier can prove useful for political text analysis.

7. Conclusions and future work

In the paper, we presented a new speech act taxonomy that captures different types of cooperation and conflict communication in political debates, with 14 fine-grained classes on the lowest level of the hierarchy. We applied our new schema to create a corpus of German parliamentary debates, with over 12,900 manually annotated speech acts. We showed that a BERT-based classifier can predict the fine-grained classes with an avg. Micro-F1 of around 82% and used our classifier to predict the pragmatic functions of utterances in German parliamentary debates over a time range of 20 years. Our initial analysis explored the use of cooperation and conflict communication on different levels in our annotation scheme, using the coarse-grained binary classes as well as the more fine-grained labels in our taxonomy. The results were consistent and robust across different operationalisations, showing that parties use different rhetorical strategies, depending on whether they are in government or part of the opposition. This provides evidence for the validity of our schema and the reliability of the predicted annotations.

In future work, we want to improve our classifier for the classes where only a small number of training instances are available. We also plan to extend our analysis by applying it to more data. We are also interested in the use of speech acts at the finer level of individual legislative terms, as we expect more confrontational behaviour toward the end of the legislative term. Another interesting line of research could investigate the interaction between speech act usage and real-world events.

8. Limitations

The most severe limitation of our work is the classifier's low accuracy for classes where only few training instances are available. We therefore strongly recommend potential users not to rely on the classifier's predictions for the rare classes (e.g., BAD-OUTCOME, SELF-REPRESENTATION, PROMISE). We would also like to stress that the classifier has not been tested across domains. We thus do not know how well it would perform when applied to political online debates that might use a more informal style, or to historical debates.

Acknowledgements

The work presented in this paper is funded by the German Research Foundation (DFG) under the UNCOVER project (PO1900/7-1 and RE3536/3-1). We would like to thank our student annotators for their dedicated work.

9. Bibliographical References

Samvel Abrahamyan. 2020. Peculiarities of Manipulative Strategies in English Political Discourse. *Armenian Folia Anglistika*, 16(1 (21)):66–81.

Susan Akinkurolere. 2020. Exploring the Significance of Context in Meaning: Speech Act Features of Performative Political Speeches of President Umaru Musa Yar' Adua. *Athens Journal of Humanities & Arts*, 7(1):63–84.

Jan Alexandersson, Elisabeth Maier, and Norbert Reithinger. 1995. A robust and efficient three-layered dialogue component for a speech-to-speech translation system. In *The 7th European Meeting of the Association for Computational Linguistics (EACL)*, pages 188–193.

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers.

Joaquín Artés. 2013. Do Spanish politicians keep their promises? *Party Politics*, 19(1):143–158.

John L. Austin. 1962. *How to do things with words*. Oxford: Clarendon Press.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang, and Kars Wijnhoven. 2019. The dialogbank: Dialogues with interoperable annotations. *Lang. Resour. Eval.*, 53(2):213–249.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Norman Fairclough. 1995. *Critical Discourse Analysis: The Critical Study of Language*. Longman.

Erizal Gani, Agustina, Yasnur Asri, and Liza Halimatul Humaira. 2020. Realization of Criticism in Political Discourse and Its Impact on Advance Threats and Polite Language:. In *Proceedings of the 3rd International Conference on Language, Literature, and Education (ICLLE 2020)*, Padang, Indonesia. Atlantis Press.

Ken Hyland. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, 7(2):173–192.

Ken Hyland and Polly Tse. 2004. Metadiscourse in Academic Writing: A Reappraisal. *Applied Linguistics*, 25(2):156–177.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation: coders manual.

Hadi Kashiha. 2022. On persuasive strategies: Metadiscourse practices in political speeches. *Discourse and Interaction*, 15(1).

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Natalia V. Kondratenko, Anastasiia A. Kiselova, and Liubov V. Zavalska. 2020. Strategies and tactics of communication in parliamentary discourse. *studies about languages*, 36:17–29.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts. *IPRA Papers in Pragmatics*, 1:1–21.

Mohsen Mesgar, Thy Thy Tran, Goran Glavaš, and Iryna Gurevych. 2023. The devil is in the details: On models and training regimes for few-shot intent classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1846–1857, Dubrovnik, Croatia. Association for Computational Linguistics.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Nikita Moghe, Evgeniia Razumovskaia, Liane Guillou, Ivan Vulić, Anna Korhonen, and Alexandra Birch. 2023. Multi3NLU++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3732–3755, Toronto, Canada. Association for Computational Linguistics.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Renugah Ramanathan, Shamala Paramasivam, and Tan Bee Hoon. 2020. Discursive Strategies and Speech Acts in Political Discourse of Najib and Modi. *Shanlax International Journal of Education*, 8(3):34–44.

Christian Rauh and Jan Schwalbach. 2020. The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. Available from Harvard Dataverse.

Antonio Reyes. 2011. Strategies of legitimization in political discourse: From words to actions. *Discourse & Society*, 22(6):781–807.

Josef Ruppenhofer and Daniela Schneevogt. 2014. Influence of information structure on the salience of opinions. In *Proceedings of the 12th Edition of the Konvens Conference Hildesheim, Germany October 8–10, 2014-Hildesheim*, pages 96–105. Universität Hildesheim.

Kalejaiye S. Abiola. 2021. A Speech Act Analysis Of Obama's Speech: Politics And Faith. *International Journal of Scientific and Research Publications (IJSRP)*, 11(8):372–377.

Daan Schueler and Maarten Marx. 2023. Speech acts in the Dutch COVID-19 press conferences. *Language Resources and Evaluation*, 57(2):869–892.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, London.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2019. Target based speech act classification in political campaign text. *NAACL HLT 2019*, page 273.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Haode Zhang, Yuwei Zhang, Li-Ming Zhan, Jiaxin Chen, Guangyuan Shi, Xiao-Ming Wu, and Albert YS Lam. 2021. Effectiveness of pre-training for few-shot intent classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1114–1120.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Discriminative Nearest Neighbor Few-Shot Intent Detection by Transferring Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082, Online. Association for Computational Linguistics.

Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. 2017. Asking too much? The rhetorical role of questions in political discourse. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1572, Copenhagen, Denmark. Association for Computational Linguistics.

Supplementary Material

A. Appendix

A.1. Dataset class distributions

Table 8 shows the class distributions in our speech acts classification dataset.

Class	Train	Dev	Test
Accusation	432	61	216
Subj-statement	4,003	576	1,413
Request	583	51	146
Promise	110	14	38
Bad-outcome	133	26	43
Report	1,528	262	686
Self-representation	132	9	22
Support	65	11	9
Demand	547	76	206
Rejection	48	7	24
Question-All	280	42	92
Expressive	355	41	142
Macro	654	83	250
Total	8,583	1,214	3,150

Table 8: Number of instances in train, development and test splits for the speech act classification task.

A.2. Consistency checks

After the annotation phase (see Section 3), extensive consistency checks have been carried out in order to find and correct inconsistent annotations. We first fine-tuned and evaluated a `BERT-base` model on the concatenation of the BT train, development and test sets. We then focussed on instances where the model predicted a label that differed from the manually assigned label, assuming that those cases might be due to inconsistencies in the annotation and thus might include annotation errors.

We found that the following classes were often confused by the model:

- ACCUSATION and SUBJECTIVE-STATEMENT
- REPORT and SUBJECTIVE-STATEMENT
- DEMAND and REQUEST

We therefore manually inspected and, if necessary, corrected each gold label for which the model predicted a label not included in the manual annotations. This semi-automated consistency check provides a fast and efficient method to detect potential outliers in the data that stem from inconsistencies and errors in the human annotation.

A.3. Training details and hyperparameters

Table 9 shows the hyperparameters we used to fine-tune models in our experiments. Hyperparameters were selected using a hyperparameter sweep, optimizing for validation loss. For the segmentation model, we use (Chan et al., 2020)'s large German BERT model, available from `https://huggingface.co/deepset/gbert-large`.

Models were fine-tuned on one RTX A6000 GPU. Training one BERT-base classification model took approximately 6 minutes.

A.4. Original German example

Figure 5 shows the original German text for Figure 1, taken from a speech by Sebastian Brehm (*CDU/CSU*), delivered on November 18, 2021 in the German Bundestag.

A.5. Additional examples for each label

Table 10 shows additional examples for each of the speech act classes in our taxonomy.

Hyperparameter	Segmentation model	Classification model
Optimizer	AdamW (Loshchilov and Hutter, 2019)	AdamW (Loshchilov and Hutter, 2019)
Training epochs	10	4
Batch size	16	16
Learning rate (LR)	$2.693154582157772 \times 10^{-5}$	$2.9206589963284678 \times 10^{-5}$
AdamW epsilon	$5.45374378277376 \times 10^{-7}$	1×10^{-8}
Weight decay	0.019840937077311938	0.01
Max. grad. clip. norm	1.0	1.0
Warmup ratio for LR	0	0.1

Table 9: Hyperparameter settings used in our speech act segmentation and classification experiments. Max. grad. clip. norm stands for maximum gradient clipping norm.

Sebastian Brehm (*CDU/CSU*), 18.11.2021:

Schauen Sie sich das doch einmal an! Sie haben null Komma null Empathie, Sie haben null Komma null Respekt vor den kleinen und mittleren Landwirten! Die machen Sie mit diesem Gesetz kaputt, und die FDP fungiert zum ersten Mal wieder als Steuererhöhungspartei. Wir lehnen diesen Gesetzentwurf heute ab, mit einem klaren Signal an die Landwirtschaft: Wir stehen zu den kleinen bäuerlichen Betrieben, und wir wollen ihre Strukturen erhalten.

Utterance segmentation	Speech act
Schauen Sie sich das doch einmal an!	Demand
Sie haben null Komma null Empathie,	Accusation
Sie haben null Komma null Respekt vor den kleinen und mittleren Landwirten!	Accusation
Die machen Sie mit diesem Gesetz kaputt,	Bad-outcome
und die FDP fungiert zum ersten Mal wieder als Steuererhöhungspartei.	Accusation
Wir lehnen diesen Gesetzentwurf heute ab, mit einem klaren Signal an die Landwirtschaft:	Rejection
Wir stehen zu den kleinen bäuerlichen Betrieben,	Promise
und wir wollen ihre Strukturen erhalten.	Promise

Figure 5: Exempt from a parliamentary debate (original German text) and its corresponding speech act annotation. The English translation is shown in Figure 1.

Class	**Example**
ACCUSATION	[. . .] und die designierte, neue Ampelkoalition ist hier bislang untätig. *[. . .] and the designated new traffic light coalition has so far failed to act.*
SUBJECTIVE-STATEMENT	Deshalb ist das Infektionsschutzgesetz genau der richtige Schritt gewesen. *That is why the Infection Protection Act was exactly the right step.*
REQUEST	Lassen Sie uns gemeinsam den Soli abschaffen. *Let us abolish the solidarity surcharge together.*
PROMISE	Und zu diesen Zielen stehen wir nach wie vor. *And we still stand by these goals.*
BAD-OUTCOME	Die sogenannte Energiewende ist eine Energiewende ins Nichts. *The so-called energy transition is an energy transition to nowhere.*
REPORT	Kinder und Jugendliche werden regelmäßig in Kitas und Schulen getestet. *Children and young people are regularly tested in daycare centers and schools.*
SELF-REPRESENTATION	Als Linke übernehmen wir Verantwortung. *As a left-wing party, we take responsibility.*
SUPPORT	Wir befürworten den vorliegenden Antrag *We are in favor of this proposal*
DEMAND	Machen Sie endlich Ihre Arbeit! *Do your job at last!*
REJECTION	Und genau deshalb lehnt die AfD den Gesetzentwurf ab. *And this is precisely why the AfD rejects the bill.*
QUESTION	Worum geht es im Einzelnen? *What is it about in detail?*
EXPRESSIVE	Vielen Dank. *Thank you very much.*
MACRO	Nächster Redner: für Bündnis 90 / Die Grünen Kai Gehring. *Next speaker: for The Greens Kai Gehring.*

Table 10: Additional examples for each of our 14 fine-grained speech act classes. Examples are taken from the BT data and translations are provided in italics.