The MoPE Corpus – *M*entions *o*f the *P*eople and the *E*lite Datasheet

I. Motivation For Dataset Creation

A. Why was the dataset created?

The main goal of the corpus is to serve as training data for a classifier that can detect references to *The People* and *The Elite* in political text, as a measure of *thin populism* (see Jagers and Walgrave, 2007) [**?**]. Previous datasets for measuring populism have either approximated the construct by weakly supervised labels based on party affiliation, or have been focussing on stance and emotions towards a small subset of groups, or have been restricted in size and interpretability, if available at all. Our dataset tries to address those limitations by encoding the building blocks of populism, i.e., references to *The People* and *The Elite*, thus yielding interpretable results that are also more fine-grained than the original conceptualisation of *thin populism*, which allows users to study populist rhetoric in different contextual settings.

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

The raw data included in MoPE is freely availabe and has already been used in different projects and publications, mostly by political scientists.

C. What (other) tasks could the dataset be used for?

Beyond populism detection, we expect that MoPE will also be interesting for investigations of anti-elitism in parliamentary debates. Furthermore, we expect that being able to detect mentions of different social groups in political text will also be useful for many other research questions in the political and social sciences.

In addition, we assume that MoPE might also be interesting for corpus-based investigations of political communication.

D. Who funded the creation dataset?

The research has been supported by the Ministry of Science, Research and the Arts Baden Württemberg, Germany.

II. Dataset Composition

A. What are the instances?

MoPE is a text corpus that includes political speeches by members of the German parliament. We provide the data in a tabular format, similar to the well-known CoNLL format. Each instance is a text sequence (paragraph) with annotations on the token level. Each token can have one or more annotations. MoPE includes nested annotations.

B. How many instances are there in total?

The dataset includes text from 267 speeches held in the German Bundestag by 196 different speakers (213,617 tokens). The time frame covers the 19th legislative term (2017–2021). The data has been split into train, development and test data and has been converted into a "flat" version where nested annotations have been ignored (e.g., for nested coordination "[[children $_{pAge}$] and [adolescents $_{pAge}$] $_{pAge}$]", we only consider the largest span "[children and adolescents $_{pAge}$]"). This version of MoPE includes 7,422 annotated mentions (22,479 annotated tokens).

We will also release a version of MoPE that includes all nested annotations.

C. What data does each instance consist of?

Each instance consists of the text of one paragraph in a tabular format, with one token per line (similar to the CoNLL format for Named Entity Recognition (NER) data. Each line includes a paragraph and token id, the word form, the annotations for level 1-3 (our annotation scheme includes hierarchical annotations on three levels) and meta-information (file name, speaker name, party affiliation for the speaker, date of the debate and speech id).

D. Is there a label or target associated with each instance? If so, please provide a description.

For more information on the annotation scheme, please refer to the annotation guidelines in the supplementary materials and our paper submission (links to those documents will be added to the final version of the datasheet upon publication).

E. Is any information missing from individual instances?

The dataset was created from the transcripts of the parliamentary debates and should thus be considered as a

normalised version of the original speech data. We do not include the audio files in the corpus (those are, however, available at https://www.bundestag.de/dokumente/textarchiv/. We also removed all comments from the speeches so that we could be sure that all speech events in a specific speech has been produced by the speaker.

F. Are relationships between individual instances made explicit?

The relation between individual speakers can be inferred through the meta-information provided in the data (e.g., party affiliation of the different speakers). The information on date and agenda item also allows to reconstruct which speeches have been given on the same day and topic (however, the topic itself is only specified on an abstract leve, e.g., "agenda item 1").

G. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset is a sample of the speeches from the 19th legislative term (2017–2021) of the German Bundestag. The distribution of topics in MOPE is not representative of the larger data but has been sampled to cover a more diverse range of topics, with contributions from all parties distributed over the whole legislative term. Below, we describe the sampling procedure in more detail.

a) Sampling procedure: We extracted a dataset of parliamentary debates from the German Bundestag, covering a time period from the 19th legislative term (2017 to 2021).[1] The corpus includes speeches by 807 different speakers, with over 900,000 sentences and over 16 mio tokens. From this corpus, we selected individual speeches for annotation as follows. Our goal was to create a gold standard, controlled for topic and including speeches for each of the political parties. In addition, we wanted the texts to be evenly distributed over the time span of the legislative term (2017–2021). To achieve this goal, we selected specific agenda items that covered a range of topics, and then sampled all speeches that belong to this specific agenda item, to increase the comparability of the contributions made by speakers from different parties.

b) CAP topics: We based our topic selection on the coding scheme developed in the Comparative Agendas Project (CAP) [**?**]. The coding scheme includes 21 major topics (see Table I) and more than 200 fine-grained subtopics. The topics we selected have been annotated as major CAP topics, which allowed us to use the annotated CAP data to train a topic classifier.

c) Training a CAP topic classifier: For training data, we used the Parliamentary Question Database[2], a data set with more than 10,000 major and minor interpellations posed by parliamentarians to the government. The data set ranges over the 8th to the 15th legislative periods (1976–2005). Each interpellation has been assigned to a major and a minor topic, according to the CAP coding scheme.

1	Cultural Policy Issues
2	Defense
3	Domestic Macroeconomic Issues
4	Education
5	Environment
6	Health
7	Immigration and Refugee Issues
8	Law, Crime, and Family Issues

TABLE I
MAJOR TOPICS FROM THE COMPARATIVE AGENDAS PROJECT THAT WE SAMPLED TO BE INCLUDED IN OUR DATA SET.

Before training, we did some standard preprocessing and clean-up of the data where we lower-cased the text and used a number of regular expressions to remove non-ascii characters, listings of politicians' names, header and footer information and so on. We also removed stopwords and punctuation and extracted a tokenised and lemmatised version of the speeches.[3] This resulted in a training set with 10,033 interpellations, with an average length of 388 tokens per interpellation. We then trained a feature-based classifier, based on tf-idf weighted bag-of-words (BOW) features. We experimented with different classifiers provided by the scikit-learn library[4] and found that the linear SVM gave us best results for predicting topics on the interpellations. For the 21 major topics, our classifier achieves a micro F1 of 72.9% on the indomain interpellation data.

d) Sampling based on predicted CAP topics: We then used the classifier to predict topics for each speech in the parliamentary debates, after applying the same preprocessing steps to the data. This gives us topic predictions for each individual speech. To guide our sampling process, we aggregated the predictions for all speeches belonging to the same agenda item. We call the topic based on a "majority vote" for each agenda item the *major topic* of the agenda. Our assumption is that all speeches given on the same agenda item should belong to the same major topic. As a result, we obtained a distribution of topics over all speeches for each respective agenda item. We sorted the predictions and *manually selected and validated* agenda items for each of the CAP topics in Table I, where the majority of the speeches for this agenda item have been predicted as belonging to this topic.

We only selected agenda items where each of the political parties participated in the debate, and also aimed at selecting items that are roughly evenly distributed over the time period of the legislative term, to *ensure that our data set is as representative as possible, covering a range of different topics, distributed over the whole legislative term and including speeches from all different parties on the same set of topics*.

[1]The data is freely available from https://www.bundestag.de/services/opendata, the Open Data service of the German Bundestag.

[2]https://www.comparativeagendas.net/datasets_codebooks

[3]For lemmatisation, we used the spaCy library: https://spacy.io with the de_core_news_sm model.

[4]https://scikit-learn.org

H. Are there recommended data splits (e.g., training, development/validation, testing)?

We provide the train/dev/test splits used in our experiments (ref-to-paper-submission). Table II shows the distribution of labels in the different data splits (train/development/test) for each level in our hierarchical annotation schema. Please note that we assured that none of the agenda items in the test set are included in the training set. This results in a more realistic setting as compared to distributing speeches from the same agenda item into training and test set.

I. Are there any errors, sources of noise, or redundancies in the dataset?

While we removed comments from the speeches to avoid including speech events that have been produced by persons other than the speaker, the speeches might include some interposed questions or closing remarks not properly marked in the XML version of the data.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained and does not rely on other external resources. But note that the audio and video data for the speeches can also be accessed at `https://www.bundestag.de/`.

The raw data is in the public domain. The annotated version of the data will be made available under the Creative Commons BY-SA 4.0 license (`https://creativecommons.org/licenses/by-sa/4.0/`).

III. Collection Process

A. How was the data collected?

The data has been downloaded from the open data service of the German Bundestag who provide the transcripts of all recent debates in XML format: `https://www.bundestag.de/services/opendata`.

B. If the dataset is a sample from a larger set, what was the sampling strategy?

See Section II, G.

C. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Co-authors of the paper and two student assistants with background in political/social science were involved in the data creation process.

D. Over what timeframe was the data collected?

The data was collected in January 2021 and annotated from January to March 2021.

IV. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done?

We removed comments and tokenized the data, to convert it into a one-token-per-line format.

B. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The raw data is available from `https://www.bundestag.de/services/opendata` in XML format.

C. Is the software used to preprocess/clean/label the instances available?

We used spaCy for tokenization `https://spacy.io/`, to convert the data into a one-token-per-line format.

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?

Predicting references to *The People*, based on a classifier that has been trained on the dataset, is successful in identifying Jagers and Walgrave [**?**]'s concept of *thin populism* in large amounts of text and agrees well with expert ratings for people-centrism from the Populism and Political Parties Expert Survey (POPPA).[5] We observed a positive correlation (r=.94, p=.005) between the POPPA expert ratings for people-centrism and our predicted counts for mentions of *The People* (Level 1 in our hierarchical annotation scheme).

While we could show that our results correlate with expert ratings from survey tools for German, the number of instances for the infrequent classes is not sufficient to achieve a high accuracy for those labels. In addition, the robustness of our models on data from different domains and text types still needs to be validated.

V. Dataset Distribution

A. How will the dataset be distributed?

We will make the data available via our university's GitHub account:
`https://github.com/umanlp/mope.git`.

[5] `http://poppa-data.eu/`

		Dataset distribution							
(lr)3-10		train		dev		test		total	
	Label	**#ment.**	**#token**	**#ment.**	**#token**	**#ment.**	**#token**	**#ment.**	**#token**
Level 1									
Elite	ELI	2603	8028	438	1342	1049	3302	4090	12672
People	PEO	1510	5093	134	501	656	2503	2300	8097
Level 2									
People	PPEO	1510	5093	134	501	650	1894	2300	8097
Organisation	EORG	1571	4421	267	769	656	2503	2488	7084
Person	EPER	1033	3607	172	573	402	1408	1607	5588
Level 3 *Elite-Person*									
Domain:									
politics	EPPOL	969	3293	157	493	370	1316	1496	5102
science	EPSCI	31	150	3	9	32	146	46	204
culture	EPKULT	8	50	2	3	8	17	15	77
military	EPMIL	4	44	6	37	67	149	5	46
finance	EPFINANZ	2	5	None	None	1	8	7	41
economy	EPWIRT	4	14	9	35	12	31	13	37
movement	EPMOV	5	19	None	None	None	None	13	36
NGOs	EPNGO	4	19	3	11	9	24	5	24
media	EPMEDIA	5	11	5	36	6	53	6	19
religion	EPREL	1	2	None	None	None	None	1	2
Level 3 *Elite-Organisation*									
Domain:									
politics	EOPOL	1318	3612	121	183	125	368	2031	5524
finance	EOFINANZ	76	279	1	3	1	2	117	441
military	EOMIL	70	192	6	30	21	156	148	414
economy	EOWIRT	50	148	11	48	68	319	90	346
NGOs	EONGO	25	82	4	13	74	209	40	124
media	EOMEDIA	15	37	40	160	1	2	33	97
science	EOSCI	9	36	1	5	3	4	17	93
movement	EOMOV	7	33	None	None	None	None	11	40
religion	EOREL	1	2	None	None	None	None	3	5
Level 3 *People*									
Domain:									
function	PFUNK	736	2771	202	491	4	18	1125	4354
age	PAGE	252	720	16	43	9	23	388	1136
social	PSOZ	201	652	7	32	164	231	228	845
ethnicity	PETH	72	266	2	4	11	28	149	620
national	PNAT	113	348	77	292	511	1421	194	611
generic	PGEN	138	336	8	52	65	220	221	531
geo-pol.ent.	GPE	725	1296	16	46	312	1291	1010	1710

TABLE II

LABEL DISTRIBUTION (PER ANNOTATED TOKEN AND PER MENTION) FOR THE TRAIN/DEV/TEST SPLITS FOR DIFFERENT LEVELS OF ANNOTATION.

B. When will the dataset be released/first distributed?

MOPE will be released upon publication of our research paper "Our kind of people? A new dataset for detecting populist references in political debates" that introduces the dataset. The data will be published under the Creative Commons BY-SA 4.0 license (`https://creativecommons.org/licenses/by-sa/4.0/`).

C. Are there any copyrights on the data?

No.

D. Are there any fees or access/export restrictions?

No.

VI. DATASET MAINTENANCE

A. Who is supporting/hosting/maintaining the dataset?

The dataset will be distributed via the GitHub account of [anonymised] university.

B. Will the dataset be updated?

No.

C. If the dataset becomes obsolete how will this be communicated?

We do not foresee a scenario where the dataset will become obsolete.

D. Is there a repository to link to any/all papers/systems that use this dataset?

No.

E. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?

The data is available via the Creative Commons BY-SA 4.0 license, so others may extend/augment/build on this dataset, given that they also make the new resource available under the same license.

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

B. Does the dataset contain data that might be considered confidential?

No.

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

The data might include racist and discriminating statements by specific politicians that might be considered as offensive to the user.

D. Does the dataset relate to people?

Yes.

E. Does the dataset identify any subpopulations (e.g., by age, gender)?

The dataset includes speeches by members of the German parliament, held in the Bundestag. The data collection was conducted by the Bundestag itself and all speakers were aware of the data collection and consented to it.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

Yes, all speakers are known.

G. Were the individuals in question notified about the data collection?

The data collection was conducted by the German Bundestag and all speakers were aware of the data collection and consented to it. In addition, the recordings of all debates are freely available on the Bundestag website.

H. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

No.