Our kind of people? Detecting populist references in political debates

Christopher Klamm **Ines Rehbein** **Simone Paolo Ponzetto**
Data and Web Science Group
University of Mannheim, Germany
{christopher,ines,simone}@informatik.uni-mannheim.de

Abstract

This paper investigates the identification of populist rhetoric in text and presents a novel cross-lingual dataset for this task. Our work is based on the definition of populism as a "communication style of political actors that refers to the people" but also includes anti-elitism as another core feature of populism. Accordingly, we annotate references to *The People* and *The Elite* in German and English parliamentary debates with a hierarchical scheme. The paper describes our dataset and annotation procedure and reports inter-annotator agreement for this task. Next, we compare and evaluate different transformer-based model architectures on a German dataset and report results for zero-shot learning on a smaller English data. We then show that semi-supervised tri-training can improve results in the cross-lingual setting. Our dataset can be used to investigate how political actors talk about *The Elite* and *The People* and to study how populist rhetoric is used as a strategic device.

1 Introduction

The rise of populism in Europe and throughout the world has been noted not only in politics and the media but also has been the subject of many studies in political science and related areas (see, among others, Mudde (2007)). The concept of populism, however, is complex and vague and eludes a strict definition. So far, only limited agreement exists on the exact properties of the construct, despite numerous efforts to provide a clear definition.

In the literature, populism has been described as an ideology (McRae, 1969; Mudde, 2004), a rhetoric (Abts and Rummens, 2007) or style (Moffitt, 2016), as a political strategy (Weyland, 2001, 2021; Hawkins and Kaltwasser, 2017) and as a discourse (Laclau, 1977; Aslanidis, 2016), amongst others (see Aslanidis (2018) for a short overview). The *Oxford Handbook on Populism* (Rovira Kaltwasser et al., 2017) groups existing work into three dominant approaches to analyzing populism, i.e., (i) the ideational approach of Mudde (2004), (ii) the socio-cultural approach (Ostiguy, 2017), and (iii) the political-strategic approach (Hawkins and Kaltwasser, 2017), each one capturing a different view on populism.

Nevertheless, most studies agree that *anti-elitism* and *people-centrism* are amongst the core dimensions of populist rhetoric, and the two dimensions are therefore included as features in most survey tools used to measure the degree of populism of political parties and actors (Polk et al., 2017; Rooduijn et al., 2019a; Meijers and Zaslove, 2020). One major drawback of surveys, however, is that they only provide us with one score for each party or actor and can not be used to study how populist rhetoric is used as a strategic tool in different contextual settings.

As a result, more and more efforts have been made recently to measure populist and anti-elitist attitudes directly from text (Rooduijn and Pauwels, 2011; Dai, 2018; Aslanidis, 2018; Ernst et al., 2019; Hawkins et al., 2019; di Cocco and Monechi, 2021; Vaughan and Heft, 2022). This has the advantage of providing us with more fine-grained and context-dependent measures that enable us to investigate when and how anti-elitist rhetoric is used as a strategic tool in party competition (Vaughan and Heft, 2022). In addition, it has been suggested that populist rhetoric targeting political elites might function "as a form of ethnoracial dog-whistle politics" (Bonikowski and Zhang, 2023, p.2). Evidence for this claim comes from the frequent co-occurrence of right-wing populism with nativist messages, as shown in Example 1.1 below, taken from a parliamentary speech of a far-right politician in the German Bundestag.

Ex. 1.1 *Because* **the Merkel government** *has lied* **to the people** *about how long* **refugees and illegal migrants** *will actually be with us [...] (N. Kleinwächter, AfD, 15/11/2019)*

Findings of the Association for Computational Linguistics: EACL 2023, pages 1227–1243

This example illustrates the different dimensions of populist rhetoric where anti-elitism is combined with a Manichean worldview that separates society into two antagonistic camps, *the corrupt elite* and *the pure people* (Mudde, 2004). This divide into *Us-versus-Them*, also known as *Othering*, is a well-known strategy for creating in- and outgroups, used to conceptualize specific groups as outsiders and to depict them as inferior or even as dangerous. Example 1.1 uses *Othering* to transfer the message that "refugees and illegal migrants" are not part of *The People* and that an immoral political elite is acting against *The People*'s general interest ("the Merkel government has lied to the people").

While there is no shortage of studies on various aspects of populism, only a few works have tried to develop robust and reliable measures of populism that can be used for empirical research at scale to quantify the degree of populism expressed by political actors, such as politicians and parties. Being able to assess populism from a quantitative standpoint using large amounts of data, e.g., text, has the potential, in turn, to help us understand the causes and consequences of populism by allowing us to track its spatial and temporal distribution.

In the paper, we provide a methodology to detect and quantify references to *The People* and *The Elite* in large amounts of text. We present a novel dataset of German and English political debates where instances of *The People* and *The Elite* have been manually annotated and use this data to learn to predict those references in monolingual and cross-lingual settings. We then show that these predictions align with the results of expert surveys for measuring populism but, crucially, provide us with *more fine-grained and context-sensitive* information that can be used to study left- and right-wing populism in parliamentary debates at large scale. We make all data and models available at `https://github.com/umanlp/mope.git`.

2 Related Work

2.1 Defining Populism

Defining populism is an intellectual challenge *per se*. Most scholars, however, agree that populism is a multi-dimensional construct and that *anti-elitism* and *people-centrism* are two of the core characteristics of populist discourse (Mudde, 2004; Hawkins, 2009; Dai, 2018; Schulz et al., 2017). Many studies have adapted Mudde's view of populism as "a thin-centered ideology that considers society to be ultimately separated into two homogeneous and antagonistic camps, 'the pure people' versus 'the corrupt elite"' (Mudde, 2004, p. 543).

Another influential view distinguishes between *thin* and *thick* populism, where the former is considered as a "communication style of political actors that refers to the people" (Jagers and Walgrave, 2007, pp.322). *Thick* populism, on the other hand, is similar to Mudde's definition and combines people-centrist references with anti-elitism and the exclusion of certain minority groups from *The People*. Our operationalization of populist rhetoric is most similar to Jagers and Walgrave (2007)'s *thin populism*. Still, it can also be used within other conceptual frameworks that rely on people-centrism and anti-elitism as defining features of populism.

So far, a variety of approaches have been proposed for analyzing populism. Some works rely on **expert opinions and surveys** (Rooduijn et al., 2019b; Meijers and Zaslove, 2021a) to obtain theoretically grounded measurements of populism. This approach, however, only yields scores on the level of parties or organizations but defies a more fine-grained or graded analysis on the text or sub-text level (Aslanidis, 2018). **Text-based approaches**, on the other hand, have the potential to identify context-sensitive manifestations of populism and its characteristics and, in turn, profile political actors along multiple dimensions.

2.2 Measuring populism in text

Text-based methods for measuring populism can be classified into four main approaches. The first is based on **manual content analysis** where a larger text is segmented into smaller units, and trained human coders inspect each unit and search for populist cues (Jagers and Walgrave, 2007; Hawkins, 2009, *inter alia*). While this approach can obtain high content validity, it is also extremely time-consuming and, depending on the categories in the codebook, does not necessarily generalize well across different topics, geographical and cultural specificities, or time periods.

A second approach, called **holistic coding**, also involves human annotation where trained coders read the document and, based on the comparison to a small set of anchor texts, decide whether the text as a whole should be considered as populist or not (Hawkins and Castanho Silva, 2018; Hawkins et al., 2019, *inter alia*). Document-level analysis is less fine-grained, and often it is not evident why a

Level 1	*Elite* E				*People* P	
Level 2	*Person* P		*Organisation* O		–	
Level 3	Domain:	Label:	Domain:	Label:	Domain:	Label:
	Politics	EPPOL	Politics	EOPOL	Nation	PNAT
	Economy	EPECON	Economy	EOECON	Ethnicity/religion	PETH
	Finance	EPFIN	Finance	EOFIN	Profession/function	PFUN
	Media	EPMED	Media	EOMED	Age	PAGE
	Science	EPSCI	Science	EOSCI	Social variables	PSOC
	Religion	EPREL	Religion	EOREL	(gender/class/...)	
	Culture	EPCULT	Culture	EOCULT	Generic	PGEN
	Military	EPMIL	Military	EOMIL		
	NGOs	EPNGO	NGOs	EONGO		
	Movements	EPMOV	Movements	EOMOV		
Other:	references to own person EPOWN		geo-political entity GPE			

Table 1: Hierarchical annotation of references to *The People* and *The Elite*.

particular text has been coded as populist. Furthermore, assigning scores to documents offers limited interpretability for analysis.

The third approach for measuring populism applies **computer-assisted content analysis**, based on dictionaries that contain cue words related to populist rhetoric, such as *people, elite, establishment, corrupt*, etc. (e.g. Jagers and Walgrave (2007); Caiani and della Porta (2011); Vasilopoulou et al. (2014); March (2017); Pauwels (2011); Rooduijn and Pauwels (2011); Bonikowski and Gidron (2016)). While dictionary-based approaches are fast and scale easily, they are less valid and reliable than manual content analysis (Grimmer and Stewart, 2013). This is partly due to the arbitrariness in the selection of the dictionary entries or keywords, where (potentially biased) choices made in the creation of the dictionary can impact the analysis. Another reason for the often low content validity is that dictionary-based methods are not context-sensitive. For instance, Rooduijn and Pauwels (2011) have tried to capture notions of *people-centrism* and *anti-elitism* in text using a dictionary-based approach, and found a reduced content validity compared to manual coding, especially for people-centrism.

The fourth approach uses **supervised machine learning (ML)** for populism detection. First steps in this direction have been taken by Dai (2018); di Cocco and Monechi (2021) and Huguet Cabot et al. (2021). Dai (2018) presents an approach based on document embeddings and SVMs to predict whether a text is populist or not. The reported performance is quite high (95% acc.), but merely due to the choice of evaluation metric and the highly skewed class distribution (i.e., only 4% of the instances in the dataset are labeled as populist).

In contrast, di Cocco and Monechi (2021) do not rely on manual annotations but approximate populism by party affiliation. They consider all sentences uttered by members of a populist party as populist and show that their measure of populism, based on the predictions of a classifier trained on the weakly supervised data, correlates with party membership and, thus, with the experts' ratings of populism. However, the approach does not capture the defining features of the construct, and it is unclear what has been learned by the classifier.

Huguet Cabot et al. (2021) present a dataset of Reddit comments annotated for stance (Discriminatory, Critical, Neutral, Supportive) and emotions towards six social groups (Conservatives, Liberals, Immigrants, Refugees, Jews, Muslims). While they also aim at detecting *Us vs. Them* rhetoric, in their work, the groups are given. In contrast, we explicitly model the building blocks of populism, i.e., references to *The People* and *The Elite*, and detect all mentions of either group in text. The advantage of our approach is threefold. First, our representations are contextualized, thus overcoming the shortcomings of dictionary-based approaches. Second, by manually coding all mentions to *The People* and *The Elite* in text, we can overcome the problem of incomplete or biased keyword lists, which is

party	speeches	speakers	tokens
CDU/CSU	76	57	72,113
SPD	58	44	48,988
AfD	39	30	29,301
FDP	34	25	22,736
Left	29	21	20,266
Greens	27	18	18,756
cross-bencher	3	1	1,457
total	267	196	213,617

Table 2: Some statistics for our new data set (CDU/CSU: Christian Democratic Union and Christian Social Union; SPD: Social Democratic Party; AfD: Alternative for Germany; FDP: Free Democratic Party; Left: The Left; Greens: The Greens).

another weakness of dictionary-based approaches (Grimmer and Stewart, 2013). Finally, our approach yields more fine-grained results that allow us to study differences in populist rhetoric, e.g., for actors from different ideological backgrounds.

3 MoPE: Annotating *M*entions *o*f the *P*eople and the *E*lite

We now present MoPE, our new data set with annotated mentions of *The People* and *The Elite*.

***The People* versus *The Elite*.** According to Mudde (2017), the difference between the two camps in populist rhetoric is not based on issues of class or nationality, but rather on *morality*. *The People* are an artificial construct of a (non-existing) homogeneous community whose defining criteria are self-ascribed and depend on the specific ideology that serves as the carrier for the *thin-centered ideology*, i.e., populism (see §2.1). *The Elite*, on the other hand, can be seen as the anti-thesis of the *The People* and also obtains its defining features based on the situational context.

To operationalize the two concepts, we use a hierarchical schema where we encode instances of the two classes on the first level (Table 1). Level 2 then distinguishes individuals and groups of persons from elite organizations, while Level 3 encodes fine-grained information about the individual actors. Our schema builds upon and extends the categories in the codebook of Wirth et al. (2019, p.12)[1]. Additionally, Level 3 encodes geo-political entities (GPE) as they provide important information for many applications. Following Jagers and Walgrave (2007) and Wirth et al. (2019), we use the

[1]https://osf.io/2z3dk/

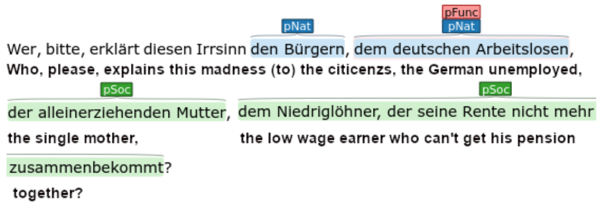

Figure 1: Annotations of references to *The People* (PNAT: people by nationality; PFUNC: people by function; PSOC: social variables like gender, class).

term *Elite* in a broad sense as referring to persons, groups, organizations or institutions with a disproportionate amount of power, wealth, privilege or skills through which they can have an impact on politics and society. As instances of *The People*, we consider (a) unspecified groups of people and (b) individuals that denote common members of the public, such as John Q. Public.

German Bundestag data. We extracted a dataset of German parliamentary debates for the 19th legislative term (2017–2021), controlled for topic and party membership of the speakers.[2] The time frame was selected because of its relevance for the rise and consolidation of populist rhetoric in German politics. Our data set includes 267 speeches by 196 different speakers from 6 German parties (Table 2). Figure 1 shows an example annotation from our data, with references to different mentions of *The People*. Please note that while our task has some similarities to Named Entity Recognition (NER), there are also crucial differences. Most importantly, only some of our mentions are proper names, while many of them are noun phrases that include subordinated clauses like relative clauses (e.g., "the low wage earner who can't get his pension together" in Figure 1). This means that the average span length of our mentions is considerably longer than for NER, which introduces additional ambiguity for annotation and prediction.[3] We will come back to this issue in §3.2. Annotations can (and often do) include embedded mentions. Entities can belong to more than one class (see, e.g., *the German unemployed* in Figure 1, which belongs to the classes "People by Nation" and "People by Function").

[2]We follow best practices and provide a datasheet (Bender and Friedman, 2018; Gebru et al., 2021) with details on corpus creation and sampling in the supplementary materials.

[3]For example, some of the ambiguities arise from PP attachment ambiguities for longer mention spans.

	Label Domain	*exact* F1	*overlap* F1	*mentions* avg. #
Elite (Person)	Politics	0.73	0.84	2,017.5
	Science	0.37	0.37	40.5
	Culture	0.59	0.65	17.0
	Economy	0.11	0.11	9.5
	Finance	0.11	0.11	9.0
	Movements	0	0	7.5
	NGO	0.18	0.18	5.5
	Media	0.22	0.55	4.5
	Military	0	0.25	4.0
	Religion	1.00	1.00	1.0
	avg.	**70.6**	**81.3**	**2,116.0**
Elite (Organisation)	Politics	0.76	0.84	2,443.0
	Finance	0.64	0.79	147.0
	Military	0.72	0.77	132.0
	Economy	0.32	0.56	97.5
	NGO	0.40	0.42	42.5
	Media	0.54	0.77	26.0
	Science	0.46	0.57	17.5
	Movements	0.59	0.59	8.5
	Culture	0	0	2.5
	Religion	0	0	2.0
	avg.	**72.8**	**81.2**	**2,918.5**
People	Function	0.58	0.76	1,572.0
	Age	0.73	0.87	487.5
	Social	0.49	0.61	426.5
	Nation	0.56	0.70	258.5
	Generic	0.42	0.42	187.0
	Ethnicity	0.41	0.51	128.0
	avg.	**57.2**	**71.9**	**3,059.5**

Table 3: Average F1 (micro) for exact match and span overlap for the two coders on the full German data.

English Europarl-UdS data. We additionally compile an English data set to enable testing for the generalization capabilities of our models not only across languages but also beyond recent debates and topics. The English data was extracted from the EuroParl-UdS corpus (Karakanta et al., 2018), a multilingual (En, De, Es) parallel corpus of parliamentary debates from the European parliament, with speeches from 1999–2018. We randomly selected speeches from three different years (1999, 2014, 2015), with 70 different speakers from 18 countries (for details, see Appendix, Tables 12, 10).

Annotation process. The data was double annotated by two student assistants with background in political/social science. During the annotation process, we had weekly meetings to discuss ambiguous cases. The final version was adjudicated by one of the authors (a linguist by training), who also corrected inconsistent span annotations: it includes 9,297 annotated mentions (German subcorpus). In our experiments, we ignore all mentions where the speakers refer to themselves (Label EPOWN) using the pronouns *I/me*, since this label can be assigned based on a simple string match. This results in a set of 7,422 mentions with 22,479 annotated tokens that we divide into training, dev and test set (see Appendix B, Table 11 for more details on the size and distribution of the different splits).

The English data set includes 29,584 tokens with 1,423 annotated mentions (1,074 w/o EPOWN) and 3,567 annotated tokens (3,218 w/o EPOWN).

	Label Domain	*exact* F1	*overlap* F1	*mentions* avg. #
Elite (Person)	Politics	0.76	0.83	241.0
	Movements	0.29	0.57	3.5
	Science	0	0	1.0
	avg.	**0.75**	**0.82**	**245.5**
Elite (Organisation)	Politics	0.75	0.82	410.0
	Movements	0.15	0.15	6.5
	Economy	0.65	0.69	24.5
	NGO	0.55	0.73	5.5
	Science	0.67	0.67	1.5
	Media	0.86	0.86	3.5
	Finance	0	0	1.0
	Military	0	0	0.5
	avg.	**0.73**	**0.80**	**453.0**
People	Social	0.71	0.87	151.5
	Function	0.28	0.38	29.0
	Nation	0.67	0.78	18.0
	Generic	0	0	5.0
	Age	0.67	0.67	7.5
	Ethnicity	0	0	1.5
	avg.	**0.62**	**0.76**	**212.5**

Table 4: Average F1 (micro) for exact match and span overlap for the two coders on the English data.

3.1 Inter-annotator agreement (IAA)

Since our data includes multi-label annotations, we cannot report Cohen's κ. We follow Hripcsak and Rothschild (2005) and compute F1, treating the annotations of one annotator as the ground truth and the other as the predicted annotations. We then switch roles and report averaged micro F1 on the mention level for the fine-grained labels (level 3).[4] Table 3 reports micro F1 on the mention level for German, using a strict measure that only considers a mention as correct when all tokens that belong to that mention have been identified correctly. The last column shows the average number of tokens annotated by our two coders (i.e., the number of in-

[4] Also see the discussion in Hripcsak and Rothschild (2005) why chance-corrected measures are not optimal for NER and other sequence-level tasks where the number of negative entities is unknown.

stances *before* adjudication). As the exact mention metric is rather strict and punishes spans that have been identified correctly by both coders but where the span boundaries slightly disagree, we also report a measure based on token overlap that has been introduced for the evaluation of opinion role spans (Katiyar and Cardie, 2016). Here we consider a mention as correct if the annotations overlap and both annotators have assigned the same label. Micro F1 for exact match is 0.69, while the overlap measure is much higher with an F1 of 0.80.

Table 4 shows IAA for the English data from the EU parliament. As for German, references to the people seem to be the most difficult class.

3.2 Error analysis

We notice a high variance in F1 for the different classes. In particular, we can see that F1 for the frequent label types is much higher than IAA for the low-frequency labels. Looking at the data, we see that our domain expert annotators often disagree on the exact span of the mentions. In particular, one annotator often failed to include complement clauses which strongly impacts exact IAA.

The F1 scores for overlapping annotation spans (Table 3) show a substantial increase for many classes, confirming our assumption that the annotators did not so much disagree on the *class labels* but on the *span boundaries* of the mentions. As mentioned above, at times, the domain experts also struggled with PP attachment decisions, as illustrated in Example 3.1 where "at age 63" should not be included in the mention span.

Ex. 3.1 *So why should* **professional soldiers at age 63** *no longer be able to meet the physical demands of service [...] (E. Brecht, SPD, 9/6/2021)*

In addition, the confusion matrix (Appendix B, Table 7) suggests that recall is a problem, showing a considerable number of instances that have been coded by one annotator only. We confirm this problem by looking at individual classes. Especially generic mentions of *The People* have been annotated mostly by one of the two annotators (263 instances have been identified as PGEN by A1 while A2 annotated 111 instances only). This recall problem has been discussed by Beigman Klebanov et al. (2008) for the metaphor detection task where the authors distinguish between *genuine disagreements* and *slips of attention*, which is a common phenomenon, especially for rare classes where the units of analysis are not given, and the annotators first have to detect them in longer texts before they can assign the labels.

We also notice some systematic disagreements for the classes in our schema. Examples are, for instance, the classes PEOPLE BY NATION and PEOPLE BY ETHNICITY, where A1 shows a bias for the first label while A2 preferred the second. This happened for mentions like *the population of X*, which can be interpreted as 'citizens of X' (PNAT) or as referring to all people who live in the country and thus share the same cultural background (PETH). Another systematic disagreement concerns PEOPLE BY FUNCTION and GENERIC mentions, illustrated in Example 3.2. Here, A1 interpreted the mention ("the people who...") as a generic reference (PGEN) while A2 focused on the function of the people (rebuilding the country) and assigned the label PFUNC.

Ex. 3.2 *I am proud of our country and of* **[the people who, through the economic miracle, have made it a country that is treated with respect and appreciation** $_{pFunc/pGen}$**]**.
(J. Juratovic, SPD, 28/5/2020)

In general, we notice that IAA for mentions of *The Elite* is higher than for references to *The People*. We suggest that this is due to two reasons. First, mentions to *The People* are, per definition, more abstract and vague, and second, the average mention length for instances of *The People* is longer than for *The Elite* (elite person: 2.3, elite organization: 2.7, people: 3.1 tokens).

4 Experiments

We use our data set from §3 to benchmark the task of predicting mentions of *The Elite* and *The People* from text sentences. Our task can be decomposed into two separate sub-tasks: (i) mention *detection* (`MD`) and (ii) mention *classification* (`MC`). We present experiments where we compare different transformer-based model architectures (Vaswani et al., 2017; Devlin et al., 2019) for those tasks. Specifically, we compare (i) a pipeline approach (`MD`→`MC`) with (ii) an end-to-end token classification model (`E2E-Tok`) and (iii) semi-supervised tri-training (`TRI`) (Zhou and Li, 2005).

Mention detection. Our `MD` model is a token classification model, similar to the NER model of Devlin et al. (2019), and predicts the span boundaries for mentions of *The People* and *The Elite* on the token level. We use the BIO schema to encode

	Task & model architecture		dev set Prec	dev set Rec	dev set F1	test set Prec	test set Rec	test set F1
	span detect.	MD	82.0 ± 1.00	83.0 ± 0.80	82.4 ± 0.86	79.5 ± 1.21	80.4 ± 1.91	80.0 ± 1.34
Level 1	*label predict.*	MC	97.6 ± 0.10	97.5 ± 0.10	97.6 ± 0.10	96.8 ± 0.03	96.8 ± 0.03	96.8 ± 0.03
Level 2	*upper bound*	MC	96.5 ± 0.10	96.4 ± 0.10	96.4 ± 0.10	95.9 ± 0.37	95.9 ± 0.37	95.9 ± 0.37
Level 3	*on gold spans*	MC	92.5 ± 0.46	92.4 ± 0.47	92.4 ± 0.47	88.1 ± 1.76	88.1 ± 1.76	88.1 ± 1.76
Level 1	Pipeline	MD→MC	74.5± 1.0	81.1± 1.07	77.7± 1.03	72.6± 1.13	79.6± 1.24	75.9± 1.18
Level 1	End-to-end	E2E-Tok	82.6± 1.09	83.1± 1.41	82.8± 0.20	77.1± 2.84	79.6± 1.29	78.3± 1.63
Level 2	Pipeline	MD→MC	72.7 ± 0.2	78.9 ± 0.22	75.7 ± 0.21	70.9 ± 0.22	77.6 ± 0.24	74.1 ± 0.23
Level 2	End-to-end	E2E-Tok	83.0± 0.31	80.7± 0.80	81.9± 0.55	79.2± 0.89	78.3± 0.74	78.7± 0.39
Level 3	Pipeline	MD→MC	68.7 ± 3.0	72.3 ± 3.16	70.4 ± 3.08	63.8 ± 3.85	67.9 ± 4.10	65.8 ± 3.97
Level 3	End-to-end	E2E-Tok	80.6 ± 1.38	79.6 ± 0.88	80.1 ± 0.49	73.6 ± 2.00	74.8 ± 1.21	74.2 ± 0.48

Table 5: F1 (micro), precision and recall for the different models on the German dev and test sets. **Bold** indicates the best performing end-to-end scores for each annotation level and ± shows stdev over the three runs.

the span boundaries and, for each token, predict whether it belongs to a specific mention.

Mention classification. Our next model architecture tries to predict the label for a given mention using sequence classification. For this, we concatenate the input sentence with the respective mention, separated by a [SEP] token, and input the sequence to the model, which then predicts a label for the entire sequence. Please note that this model relies on gold spans as input and provides an upper bound for determining the correct class of a mention.

Pipeline. When performing mention classification, the span-based `MC` model needs to know the span boundaries to predict a mention's label. Therefore, we test a pipeline approach where we first use the `MD` model to detect the spans of the mentions and then predict the label, using the `MD` output as input to the `MC` model.

End-to-end token classification. We compare the pipeline results to an end-to-end token classification model. The architecture is similar to the `MD` model, but in addition to span boundary detection, we also predict the labels of the mentions on the token level. We use the BIO schema as prefixes to the class labels to encode the span boundaries *and* class for each mention and, for each token, predict whether it belongs to a specific span *and* class (including the None class).

Cross-lingual tri-training with disagreement. Semi-supervised approaches have successfully improved model performance, especially in low-resource scenarios. We, therefore, test the potential of *tri-training* (Zhou and Li, 2005) in a cross-lingual setting to improve results for knowledge transfer from German to English. Tri-training is an iterative process where we use the predictions of two classifiers c_1, c_2 to assign labels to unlabeled instances and expand the training set of a third classifier. Previous work has shown that *tri-training with disagreement*, i.e., adding only those instances to the training data of c_3 where c_1 and c_2 agree with each other's predictions but disagree with the prediction of c_3, can filter out uninformative instances and improve the efficiency of the training process (Chen et al., 2006; Zhou, 2008; Søgaard, 2010).

Specifically, we use the end-to-end architecture (`E2E-Tok`) to train three multilingual classifiers based on `bert-base-multilingual-cased` with different seeds on the German train set. For each seed, we select the model that performed best on the dev set. We then use the three classifiers to predict labels for new, unlabeled data points from the English part of the EuroParl-UdS corpus and, for each classifier c_i, select new instances based on *disagreement* and add them to c_i's training set. Please note that this results in different training sets for each classifier. We then continue fine-tuning the classifiers on the expanded training data for m iterations, followed by n iterations of supervised training on gold data. We repeat this process until the results on the dev set stop improving. Then we use the three semi-supervised classifiers to predict labels for the test set based on majority voting.

In contrast to previous work (Ruder and Plank, 2018), we do not share parameters between learners but encourage the diversity of the models by keeping them separate. For efficiency, we do not fully retrain the models on the expanded data but

		German test_{de}				English test_{en}		
Level	**Model**	**prec**	**rec**	**F1**	**Model**	**prec**	**rec**	**F1**
Level 1	mBERT	**78.7**$\pm$ 1.59	76.3$\pm$ 0.68	**77.5**$\pm$ 0.96	ZERO	**71.9**$\pm$ 2.33	74.7$\pm$ 1.00	73.3$\pm$ 0.75
	TRI	77.2$\pm$ 0.22	**77.7**$\pm$ 0.28	77.4$\pm$ 0.25	TRI	70.6$\pm$ 1.14	**79.6**$\pm$ 1.06	**74.8**$\pm$ 1.11
Level 2	mBERT	77.0$\pm$ 1.07	75.0$\pm$ 0.15	76.0$\pm$ 0.60	ZERO	69.6 $\pm$ 2.00	74.0$\pm$ 1.63	71.7$\pm$ 1.81
	TRI	**78.2**$\pm$ 0.84	**77.2**$\pm$ 0.44	**77.7**$\pm$ 0.19	TRI	**70.1**$\pm$ 1.62	**79.4**$\pm$ 1.20	**74.4**$\pm$ 0.41
Level 3	mBERT	70.9$\pm$ 0.92	72.6$\pm$ 0.40	71.7$\pm$ 0.42	ZERO	68.3$\pm$ 1.20	74.8$\pm$ 0.66	71.4$\pm$ 0.96
	TRI	**75.3**$\pm$ 0.03	**72.7**$\pm$ 1.34	**74.0**$\pm$ 0.70	TRI	**69.8**$\pm$ 1.50	**75.5**$\pm$ 0.42	**72.5**$\pm$ 0.87

Table 6: Results for zero-shot learning and tri-training for the mBERT E2E-Tok model on the German test set and on the English benchmark data.

simply add $m + n$ epochs of fine-tuning in each iteration. For details on model setup and parameter settings, see Appendix B.1, B.4 and B.2.

4.1 Results for German

In all experiments, we report results averaged over three runs with different initializations. All models are implemented with the Huggingface transformers library (Wolf et al., 2020) and PyTorch (Paszke et al., 2017). For evaluation, we use seqeval (Nakayama, 2018), a python implementation of the well-known CoNLL 2000 evaluation script for sequence tagging tasks (Tjong Kim Sang and Buchholz, 2000), and report precision, recall and F1 (micro) in *strict* mode on the mention level for the different levels of our hierarchical annotations (see Appendix B.3 for details).

We first report results for the token-based **mention detection** task (Table 5). F1 on the development and test set are close with around 80%. The upper bound for **mention classification** of gold mention spans is very high for the coarse-grained levels where we distinguish between mentions of *The People* and *The Elite* (Level1/2), with an F1 of around 96%. For the fine-grained classes, the upper bound is around 92% for dev and 88% for test (Table 5, MC, Level3).

We now turn to the end-to-end architectures (MD→MC, E2E-Tok) where we predict the span boundaries *and* the class labels. While the MC model performs well on gold mentions, it visibly struggles to predict labels for automatically determined spans, and F1 decreases by around 20% for all levels (Table 5). On the other hand, our end-to-end token-based model is much better suited for this task, with an F1 over 74% for L3 and around 80% for the coarse-grained prediction of mentions of *The People* and *The Elite*.

4.2 Cross-lingual transfer to English

Zero-shot transfer. Lauscher et al. (2020) have shown that results for *zero-shot cross-lingual transfer* do not decrease much for lower-level tasks like PoS and NER if source and target language are typologically close. This observation encourages us to try zero-shot transfer learning for our task, which is closely related to NER. We use the E2E-Tok architecture from our previous experiments and initialize it with a pretrained multilingual transformer (mBERT). We then train mBERT on the German data and use it to detect instances of *The People* and *The Elite* in the English debates. The experiments are meant to investigate how well we can transfer information from German to English without annotating *any* English data.

Table 6 shows results for the mBERT model on the German test set and zero-shot learning, using the same model to predict labels for the English benchmark data. We can see that F1 for the fine-grained Level-3 predictions on the English test set is only slightly lower than for German (71.7% vs. 71.4% F1). However, the gap between precision and recall is more substantial than in the monolingual setting, and the trend is reversed, showing higher recall with much lower precision. Not surprisingly, results for mBERT on the German test set are lower than the ones for the German BERT model (cf. Table 5).

Looking at the tri-training results, we observe another increase of around 1% for the English data. Interestingly, training the classifier on unlabeled English data also yields an improvement of >2% F1 on the German test set (L3) for mBERT, closing the gap between the mBERT and German BERT results. Overall, the results indicate a successful transfer, considering that the model did not see *any* hand-labeled English data during training.

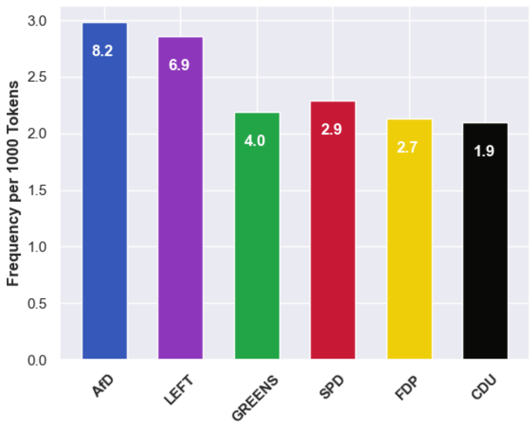

Figure 2: Distribution of references to *The People* in the German Bundestag (2017-2021). Numbers in the bar show POPPA scores for people-centrism.

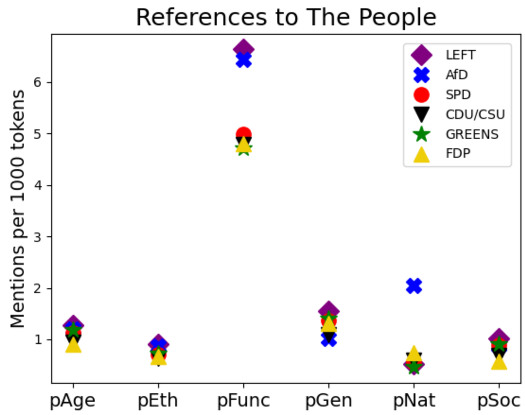

Figure 3: Distribution of group mentions in the 19th legislative term of the German Bundestag (2017-2021).

5 Measuring *thin populism* from text

We are now able to investigate Jagers and Walgrave (2007)'s concept of *thin populism* by looking at how often political actors refer to different subsets of *The People*. For that, we use our three monolingual classifiers described in §4 and predict labels for all debates from the German Bundestag from the 19th legislative term (2017–2021) ($>$ 16 million tokens). We take the majority vote of the three classifiers to determine the final predictions. Figure 2 shows the distribution of the aggregated counts for all references to *The People* for each party. [5]

We can now validate how well our operationalization of *thin populism* in text correlates with expert ratings. For that, we compute Spearman's rank correlation between the normalized counts for each party and the party's score for people-centrism in the Populism and Political Parties Expert Survey (POPPA) (Meijers and Zaslove, 2021b) (also see Table 9 in the Appendix, C). We observe a very strong positive correlation ($\rho = .94$, $p = .005$) between the expert ratings for people-centrism and our predicted counts (Level 1), where both left and right-wing populist parties show a substantially higher amount of people mentions.

However, when looking at the fine-grained predictions for different subgroups of *The People* (Level 3, Figure 3), we also notice interesting differences. For example, both populist parties use a higher amount of references to PEOPLE BY FUNCTION than the mainstream parties. At the same time, only the far-right AfD shows excessive use of PEOPLE BY NATION, often as a dog-whistle to send the message that some people are not "our kind of people".[6]

Overall, our approach of predicting references to *The People* is able to successfully identify populist rhetoric in large amounts of text and agrees well with expert ratings. However, our results also highlight the importance of a more fine-grained operationalization of *thin populism* that distinguishes between different subgroups of *The People*.

6 Conclusions

In this paper, we presented MoPE, a novel data set for detecting mentions of *The People* and *The Elite* in political text. Our data set includes more than 9,000 annotated mentions for German and an English benchmark set with around 1,600 mentions for cross-lingual transfer learning. We evaluated different transformer-based model architectures on our new data set and explored zero-shot cross-lingual transfer and cross-lingual tri-training.

In future work, we will combine references to *The Elite* with stance detection, which will allow us to model and quantify the different dimensions of populism separately, i.e., *people-centrism* and *anti-elitism*, thus enabling large-scale studies of populism from left- and right-wing political actors in different contextual settings.

[5] We excluded the CSU from the analysis. While the party is forming a joint parliamentary group with the CDU in the Bundestag, it is only running for election in a single German province, Bavaria. This results in a conflict between the party's "Bavaria first!" policy on the province level and the need to accommodate their sister party's policies on the federal level (Frymark, 2018, pp.2-3). We, therefore, expect that the governing faction is not representative of the party as a whole.

[6] This observation is consistent with the AfD's high POPPA score for nativism (9.7 of 10).

7 Limitations

We would like to point out some limitations of our work. First, in this paper, we do not yet provide measures of populist rhetoric but release a data set and method for detecting instances of *The People* and *The Elite* in text, which we see as a prerequisite for a theoretically grounded, multi-dimensional model of populism that captures the core features of the construct, i.e., *anti-elitism* and *people-centrism*. While our results correlate with expert ratings from survey tools for German, the validity of the English annotations still needs to be tested, and the accuracy for infrequent classes needs to be improved. In addition, further work needs to investigate the robustness of our models on data from different domains and text types.

Acknowledgements

We are grateful to Laura Schmitt and Marlene App for their annotation work. This work is supported by a research grant from the Ministry of Science, Research and the Arts (MWK) Baden-Württemberg.

References

Koen Abts and Stefan Rummens. 2007. Populism versus democracy. *Political Studies*, 55(2):405–424.

Paris Aslanidis. 2016. Is Populism an Ideology? A Refutation and a New Perspective. *Political Studies*, 64(1_suppl):88–104.

Paris Aslanidis. 2018. Measuring populist discourse with semantic text analysis: an application on grassroots populist mobilization. *Quality & Quantity*, 52:1241–1263.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 2–7, Manchester, UK. Coling 2008 Organizing Committee.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bart Bonikowski and Noam Gidron. 2016. The populist style in American politics: presidential Campaign discourse, 1952–1996. *Social Forces*, 94(4):1593–1621.

Bart Bonikowski and Yueran Zhang. 2023. Populism as Dog-Whistle Politics: Anti-Elite Discourse and Sentiments Toward Minority Groups. *Social Forces*. Soac147.

Manuela Caiani and Donatella della Porta. 2011. The Elitist Populism of the Extreme Right: A Frame Analysis of Extreme Right Wing Discourses in Italy and Germany. *Acta Politica*, 46(2):180–202.

Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. 2006. Chinese chunking with tri-training learning. In *Proceedings of the 21st International Conference on Computer Processing of Oriental Languages: Beyond the Orient: The Research Challenges Ahead*, ICCPOL'06, page 466–473, Berlin, Heidelberg. Springer-Verlag.

Yaoyao Dai. 2018. Measuring Populism in Contexts: A Supervised Approach with Word Embedding Models. Working paper.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jessica di Cocco and Bernarddo Monechi. 2021. How populist are parties?: Measuring degrees of populism in party manifestos using supervised machine learning. *Political analysis*, pages 1–17.

Nicole Ernst, Frank Esser, Sina Blassnig, and Sven Engesser. 2019. Favorable opportunity structures for populist communication: Comparing different types of politicians and issues in social media, television and the press. *The International Journal of Press/Politics*, 24(2):165–188.

Kamil Frymark. 2018. The Free State of Bavaria and its party: the CSU faces an electoral test. OSW Commentary NUMBER 288 | 10.10.201.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297.

Kirk A. Hawkins. 2009. Is Chávez Populist? Measuring Populist Discourse in Comparative Perspective. *Comparative Political Studies*, 42(8):1040–1067.

Kirk A. Hawkins, R. Aguilar, B. C. Silva, E. K. Jenne, B. Kocijan, and Cristóbal Rovira Kaltwasser. 2019. Measuring populist discourse: The global populism database. In *The EPSA Annual Conference*, pages 20–22.

Kirk A. Hawkins and B. Castanho Silva. 2018. Text analysis: Big data approaches. In *The ideational approach to populism: Theory, method & analysis*. Routledge.

Kirk A. Hawkins and Cristóbal Rovira Kaltwasser. 2017. What the (Ideational) Study of Populism Can Teach Us, and What It Can't. *Swiss Political Science Review*, 23(4):526–542.

G. Hripcsak and A.S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 3(12):296–298.

Pere-Lluís Huguet Cabot, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2021. Us vs. them: A dataset of populist attitudes, news bias and emotions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1921–1945, Online. Association for Computational Linguistics.

Jan Jagers and Stefaan Walgrave. 2007. Populism as Political Communication Style: An Empirical Study of Political Parties' Discourse in Belgium. *European Journal of Political Research*, 46(3):319–345.

Alina Karakanta, Mihaela Vela, and Elke Teich. 2018. EuroParl-UdS: Preserving and extending metadata in parliamentary debates. In *ParlaCLARIN workshop, 11th Language Resources and Evaluation Conference (LREC2018)*, Miyazaki, Japan.

Arzoo Katiyar and Claire Cardie. 2016. Investigating LSTMs for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany. Association for Computational Linguistics.

Ernesto Laclau. 1977. Towards a theory of populism. In E. Laclau, editor, *Politics and Ideology in Marxist Theory*, page 143. London: New Left Books.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Constantine Lignos and Marjan Kamyab. 2020. If you build your own NER scorer, non-replicable results will come. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 94–99, Online. Association for Computational Linguistics.

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

Luke March. 2017. Left and right populism compared: The british case. *The British Journal of Politics and International Relations*, 19(2):282–303.

Donald McRae. 1969. Populism as an ideology. In Ghiţa Ionescu and Ernest Gellner, editors, *Populism: Its Meanings and National Characteristics*, pages 153–65. London: Weidenfeld and Nicolson.

Maurits Meijers and Andrej Zaslove. 2020. Populism and Political Parties Expert Survey 2018 (POPPA).

Maurits J. Meijers and Andrej Zaslove. 2021a. Measuring populism in political parties: Appraisal of a new approach. *Comparative Political Studies*, 54(2):372–407.

Maurits J. Meijers and Andrej Zaslove. 2021b. Measuring populism in political parties: Appraisal of a new approach. *Comparative Political Studies*, 54(2):372–407.

Benjamin Moffitt. 2016. *The global rise of populism: Performance, political style, and representation*. Stanford University Press, Stanford.

Cas Mudde. 2004. The Populist Zeitgeist. *Government and Opposition*, 39(4):541–563.

Cas Mudde. 2007. *Populist Radical Right Parties in Europe*. Cambridge University Press, Cambridge.

Cas Mudde. 2017. Populism: An Ideational Approach. In C. Rovira Kaltwasser, P. Taggart, and et al. Ochoa Espejo, P., editors, *The Oxford Handbook of Populism*, pages 27–47. Oxford: Oxford University Press.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Pierre Ostiguy. 2017. Populism: A Socio-cultural Approach. In C. Rovira Kaltwasser, P. Taggart, and et al. Ochoa Espejo, P., editors, *The Oxford Handbook of Populism*, pages 74–97. Oxford: Oxford University Press.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Teun Pauwels. 2011. Measuring populism: a quantitative text analysis of party literature in Belgium. *Journal of Elections, Public Opinion, and Parties*, 21(1):97–119.

Jonathan Polk, Jan Rovny, Ryan Bakker, Erica Edwards, Liesbet Hooghe, Seth Jolly, Jelle Koedam, Filip Kostelka, Gary Marks, Gijs Schumacher, Marco Steenbergen, Milada Vachudova, and Marko Zilovic. 2017. Explaining the salience of anti-elitism and reducing political corruption for political parties in europe with the 2014 chapel hill expert survey data. *Research & Politics*, 4(1):2053168016686915.

M. Rooduijn, S. Van Kessel, C. Froio, A. Pirro, S. De Lange, D. Halikiopoulou, P. Lewis, C. Mudde, and P. Taggart. 2019a. The populist: An overview of populist, far right, far left and eurosceptic parties in europe.

Matthijs Rooduijn and Taun Pauwels. 2011. Measuring populism: Comparing two methods of content analysis. *West European Politics*, 34:1272–1283.

Matthijs Rooduijn, Stijn van Kessel, Caterina Froio, Sarah de Lange, Daphne Halikiopoulou, Paul Lewis, Cas Mudde, and Paul Taggart. 2019b. The popuList: An overview of populist, far right, far left and Eurosceptic parties in Europe.

Cristóbal Rovira Kaltwasser, Paul Taggart, Paulina Ochoa Espejo, and Pierre Ostiguy. 2017. *The Oxford Handbook of Populism*. Oxford: Oxford University Press.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054, Melbourne, Australia. Association for Computational Linguistics.

Anne Schulz, Philipp Müller, Christian Schemer, Dominique Stefanie Wirz, Martin Wettstein, and Werner Wirth. 2017. Measuring Populist Attitudes on Three Dimensions. *International Journal of Public Opinion Research*, 30(2):316–326.

Anders Søgaard. 2010. Simple semi-supervised training of part-of-speech taggers. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 205–208, Uppsala, Sweden. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Sofia Vasilopoulou, Daphne Halikiopoulou, and Theofanis Exadaktylos. 2014. Greece in crisis: Austerity, populism and the politics of blame. *Journal of Common Market Studies*, 52:388–402.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Michael Vaughan and Annett Heft. 2022. Anti-elitism in the european radical right in comparative perspective. *JCMS: Journal of Common Market Studies*, 61(1):76–94.

Kurt Weyland. 2001. Clarifying a Contested Concept: Populism in the Study of Latin American Politics. *Comparative Politics*, 34(1):1–22.

Kurt Weyland. 2021. Populism as a political strategy: An approach's enduring – and increasing – advantages. *Political Studies*, 69(2):185–189.

Werner Wirth, Martin Wettstein, Dominique Wirz, Nicole Ernst, Florin Büchel, Anne Schulz, Frank Esser, and et al. 2019. *Codebook: NCCR democracy Module II: The Appeal of populist Ideas and Messages*. Unpublished paper.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Zhi-Hua Zhou. 2008. Semi-supervised learning by disagreement. In *2008 IEEE International Conference on Granular Computing*, pages 93–93.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541.

Supplementary Material

A Inter-annotator agreement (IAA)

Table 7 shows the confusion matrix for our two human annotators (A1, A2) for the fine-grained classes (Level 3) in the German Bundestag debates. Due to space limitations, only the most frequent classes are shown. The **prefixes** of the labels are EP: Elite-Person, EO: Elite-Organisation, P: People. The **domains of the labels** are FIN: finance, MIL: military; POL: politics; ECO: economy; AGE: people by age; ETH: people by ethnicity; FUN: people by professon/function; GEN: Generic mentions; NAT: people by nation; SOC: social variables (gender, class); GPE: geo-political entities.

B Training details

B.1 Setup and parameters

For all experiments, we report results averaged over three runs. In each run, we initialise the model with a different seed: $\{18, 23, 44\}$. As optimizer, we use AdamW (Loshchilov and Hutter, 2019). The initial learning rate was set to 2.69^{-05}, with a weight decay of 0.0198. We did not freeze any layers but fine-tuned the whole model in all experiments. For tri-training, we experimented with $m = \{3, 5\}$ and $n = \{1, 5\}$ and found that n=3 and m=1 were robust across different levels. A more principled hyperparameter search might further improve results.

B.2 Training/dev/test splits

Table 11 shows the distribution of labels in the different data splits (train/development/test) for each level in our hierarchical annotation schema. We ensure that none of the agenda items in the test set are included in the training set. This results in a much more challenging and realistic setting compared to distributing speeches from the same agenda item into training and test sets.

B.3 Sequence tagging evaluation

As noted by Lignos and Kamyab (2020), many evaluation scripts for sequence tagging tasks will produce non-replicable results due to inconsistent handling of “improper label sequences”, i.e., mentions that have been labeled with the correct class but have been assigned an incorrect prefix. This results in an inconsistent number of entities in the gold standard and thus produces results that are not comparable. To avoid this problem, we report results for the *strict* mode where prefixes are included in the evaluation.

For illustration, consider the following two sequences:

- GOLD: [’B-ELI’, ’O’, ’B-ELI’, ’I-ELI’, ’O’, ’B-ELI’]

- PRED: [’B-ELI’, ’O’, ’O’, ’I-ELI’, ’O’, ’B-ELI’]

In *strict* mode, the seqeval evaluation script would consider only proper mentions starting with ’B’ for calculation (precision $\frac{2}{2} = 1.00$):

- GOLD: [**’B-ELI’**, ’O’, ’B-ELI’, ’I-ELI’, ’O’, **’B-ELI’**]

- PRED: [**’B-ELI’**, ’O’, ’O’, ’I-ELI’, ’O’, **’B-ELI’**]

However, in *default* mode, the seqeval evaluation first "repairs" the improper label sequences:

- PRED: [’B-ELI’, ’O’, ’O’, **’B-ELI’**, ’O’, ’B-ELI’]

After that, in *default* mode, all three mentions are used for calculation, even if they do not start in the original sequence with a starting token (precision $\frac{2}{3} = 0.67$):

- GOLD: [**’B-ELI’**, ’O’, **’B-ELI’, ’I-ELI’**, ’O’, **’B-ELI’**]

- PRED: [**’B-ELI’**, ’O’, ’O’, **’B-ELI’**, ’O’, **’B-ELI’**]

B.4 Tri-training with disagreement

We use a sample of 20,000 instances (sentences) from the EuroParl-UdS corpus as unlabelled data for tri-training. The data size was determined to extract a sufficient number of data points for tri-training while keeping the additional time for training and prediction low. From the 20,000 instances, between 950 to 1,500 instances have been selected for each classifier during tri-training (see Table 8 for exact numbers).

We loaded the checkpoints for the three best baseline classifiers (E2E) and continued training for 5 epochs on the newly extracted instances. Finally, we trained each classifier for another 5 epochs on the original training set. Then we used the three classifiers to predict labels for the test instances based on a majority vote.

A1 / A2	eoFin	eoMil	eoPol	eoEco	epPol	pAge	pEth	pFun	pGen	pNat	pSoc	GPE	None
eoFin	93	0	6	7	0	0	0	1	0	0	0	0	44
eoMil	0	100	0	0	0	0	0	2	0	1	0	0	42
eoPol	5	8	1,641	1	46	0	1	1	0	1	0	17	583
eoEco	1	0	1	33	0	0	0	11	0	0	0	0	59
epPol	1	0	43	0	1,273	0	3	54	1	26	3	2	293
pAge	0	0	0	0	2	330	0	5	1	1	32	0	50
pEth	0	0	0	0	1	3	54	5	7	6	7	0	25
pFun	0	1	0	0	1	10	2	912	40	15	124	5	314
pGen	0	0	1	0	0	0	8	0	78	1	0	0	23
pNat	0	0	0	0	0	0	30	3	12	144	2	0	26
pSoc	0	0	0	0	1	2	2	12	3	1	194	0	35
GPE	0	0	13	0	0	0	2	0	1	0	1	1,008	188
None	16	5	203	18	93	62	33	341	121	43	110	102	198,211

Table 7: Confusion matrix for two human annotators A1, A2 for the fine-grained classes (Level 3) in the German Bundestag debates (most frequent classes only).

	Level1	Level2	Level3
Clf 1	1,142	1192	947
Clf 2	969	946	1024
Clf 3	1,066	1236	1518

Table 8: Unlabelled training instances extracted for each level and classifier during tri-training.

C Populism and Political Parties Expert Survey (POPPA)

Table 9 shows expert ratings from the 2018 Populism and Political Parties Expert Survey (POPPA) (Meijers and Zaslove, 2021b) for all six German parties that participated in government in the 19th legislative term (2017–2021). The first column lists scores for people-centrism, a core feature of populism strongly related to Jagers and Walgrave (2007)'s concept of *thin populism*, and the second column shows the mean populism score for each party, aggregated over all relevant dimensions of populism in the survey. The ratings were collected between April 2018 and July 2018 from 294 country experts and include survey items for populism, political style, party ideology, and party organization in 28 European countries.[7]

party	people-centrism	populism
AfD	8.2	9.4
LEFT	6.9	5.6
GREENS	4.0	1.4
CSU	3.9	3.2
SPD	2.9	1.5
FDP	2.7	2.5
CDU	1.9	0.8

Table 9: POPPA-2018 expert ratings for people-centrism and populism for the parties in the German Bundestag.

D Dataset details

[7] http://poppa-data.eu/

Id	Country	# toks
AT	Austria	260
BE	Belgium	2,161
BG	Bulgaria	114
CZ	Czech Republic	31
DE	Germany	358
DK	Denmark	757
EE	Estonia	655
ES	Spain	1,188
FR	France	2,111
GB	United Kingdom	6,918
IE	Ireland	1,063
IT	Italy	2,166
LV	Latvia	256
MT	Malta	214
NA	no information available	7,235
NL	Netherlands	1,492
PL	Poland	474
RO	Romania	895
SE	Sweden	1,525

Table 10: No. of tokens per country for the English data set from the EU parliament (1999-2015). NA indicates that no country information was specified in the meta-data.

		Dataset distribution							
		train		**dev**		**test**		**total**	
	Label	**#ment.**	**#token**	**#ment.**	**#token**	**#ment.**	**#token**	**#ment.**	**#token**
Level 1									
Elite	ELITE	2603	8028	438	1342	1049	3302	4090	12672
People	PEOPLE	1510	5093	134	501	656	2503	2300	8097
Level 2									
Person	ELITE-PERSON	1033	3607	172	573	402	1408	1607	5588
Organisation	ELITE-ORGAN	1571	4421	267	769	656	2503	2488	7084
People	PEOPLE	1510	5093	134	501	650	1894	2300	8097
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	epCULT	8	50	2	3	8	17	15	77
military	epMIL	4	44	6	37	67	149	5	46
finance	epFIN	2	5	None	None	1	8	7	41
economy	epECON	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	epMED	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	eoFIN	76	279	1	3	1	2	117	441
military	eoMIL	70	192	6	30	21	156	148	414
economy	eoECON	50	148	11	48	68	319	90	346
NGOs	eoNGO	25	82	4	13	74	209	40	124
media	eoMED	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	pFUN	736	2771	202	491	4	18	1125	4354
age	pAGE	252	720	16	43	9	23	388	1136
social	pSOC	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 11: Label distribution (per annotated token and per mention) for the train/dev/test splits for different levels of annotation.

Id	Name	Party	# toks
1	Mauro NOBILIA	Union for Europe of the Nations Group	562
2	Ole KRARUP	Group for a Europe of Democracies and Diversities	327
3	Carl LANG	Technical Group of Independent Members	360
4	Philip BUSHILL-MATTHEWS	Europ. People's Party (Christian Democrats) and Europ. Democrats	336
5	Alejandro CERCAS	Party of Europ. Socialists	583
6	Daniel DUCARME	Europ. Liberal, Democrat and Reform Party	235
7	Maj Britt THEORIN	Party of Europ. Socialists	412
8	Bartho PRONK	Europ. People's Party (Christian Democrats) and Europ. Democrats	508
9	Anne VAN LANCKER	Party of Europ. Socialists	866
10	Anne E. JENSEN	Europ. Liberal, Democrat and Reform Party	430
11	Hélène FLAUTRE	Greens/Europ. Free Alliance	1,141
12	Herman SCHMID	Confederal Europ. United Left/Nordic Green Left	507
13	Liam HYLAND	Union for Europe of the Nations Group	556
14	Rijk van DAM	Group for a Europe of Democracies and Diversities	375
15	Marco CAPPATO	Technical Group of Independent Members	472
16	Renzo IMBENI	Party of Europ. Socialists	309
17	Maurizio TURCO	Technical Group of Independent Members	74
18	Vytenis Povilas ANDRIUKAITIS	Party of Europ. Socialists	1,362
19	Julie GIRLING	Europ. Conservatives and Reformists Group	519
20	Lynn BOYLAN	Confederal Europ. United Left	268
21	Pavel POC	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	31
22	Anthea McINTYRE	Europ. Conservatives and Reformists Group	185
23	Nessa CHILDERS	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	224
24	Štefan FÜLE	Party of Europ. Socialists	3,017
25	Jacek SARYUSZ-WOLSKI	Europ. People's Party (Christian Democrats)	254
26	Johannes Cornelis van BAALEN	Alliance of Liberals and Democrats for Europe	317
27	Sandra KALNIETE	Europ. People's Party (Christian Democrats)	71
28	Marju LAURISTIN	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	127
29	Victor BOŞTINARU	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	152
30	Paul NUTTALL	Europe of Freedom and Direct Democracy Group	103
31	Mike HOOKEM	Europe of Freedom and Direct Democracy Group	394
32	Ioan Mircea PAŞCU	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	216
33	Richard HOWITT	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	244
34	Georgi PIRINSKI	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	114
35	Andrus ANSIP	Alliance of Liberals and Democrats for Europe	83
36	Tatjana ŽDANOKA	Greens/Europ. Free Alliance	185
37	Jean-Claude JUNCKER	Europ. People's Party (Christian Democrats)	551
38	Syed KAMALL	Europ. Conservatives and Reformists Group	1,011
39	Guy VERHOFSTADT	Alliance of Liberals and Democrats for Europe	1,060
40	Nigel FARAGE	Europe of Freedom and Direct Democracy Group	1,042
41	Gerard BATTEN	Europe of Freedom and Direct Democracy Group	208
42	Theodor Dumitru STOLOJAN	Europ. People's Party (Christian Democrats)	123
43	Věra JOUROVÁ	Alliance of Liberals and Democrats for Europe	1,046
44	Janice ATKINSON	Europe of Freedom and Direct Democracy Group	197
45	Louise BOURS	Europe of Freedom and Direct Democracy Group	249
46	Mairead McGUINNESS	Europ. People's Party (Christian Democrats)	15
47	Terry REINTKE	Greens/Europ. Free Alliance	358
48	Sophia in 't VELD	Alliance of Liberals and Democrats for Europe	292
49	Mary HONEYBALL	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	239
50	Ulrike LUNACEK	Greens/Europ. Free Alliance	260
51	Jonathan ARNOTT	Europe of Freedom and Direct Democracy Group	104
52	Julie WARD	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	193
53	Clare MOODY	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	223
54	Theresa GRIFFIN	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	375
55	Bill ETHERIDGE	Europe of Freedom and Direct Democracy Group	225
56	Diane DODDS	Non-attached Members	196
57	Doru-Claudian FRUNZULICĂ	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	404
58	Julia PITERA	Europ. People's Party (Christian Democrats)	220
59	Yana TOOM	Alliance of Liberals and Democrats for Europe	158
60	Luigi COCILOVO	Europ. People's Party (Christian Democrats) and Europ. Democrats	515
61	Jan ANDERSSON	Party of Europ. Socialists	606
62	Luciana SBARBATI	Europ. Liberal, Democrat and Reform Party	234
63	Alain LIPIETZ	Greens/Europ. Free Alliance	245
64	Sylviane H. AINARDI	Confederal Europ. United Left/Nordic Green Left	365
65	Margrethe Vestager	Alliance of Liberals and Democrats for Europe	1,259
66	Kaja KALLAS	Alliance of Liberals and Democrats for Europe	287
67	Ramon TREMOSA i BALCELLS	Alliance of Liberals and Democrats for Europe	605
68	Steven WOOLFE	Europe of Freedom and Direct Democracy Group	430
69	Anneliese DODDS	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	445
70	Alfred SANT	Progressive Alliance of Socialists and Democrats in the Europ. Parliament	214
	total		**29,584**

Table 12: Speakers and party affiliation for the English data set from the EU parliament (1999-2015).