Exploring the Robustness of Task-oriented Dialogue Systems for Colloquial German Varieties

Ekaterina Artemova [▲*] **Verena Blaschke** [▲🤖] **Barbara Plank** [▲🤖🧭]

▲ MaiNLP, Center for Information and Language Processing, LMU Munich, Germany

🤖 Munich Center for Machine Learning (MCML), Munich, Germany

🧭 Department of Computer Science, IT University of Copenhagen, Denmark

{verena.blaschke, b.plank}@lmu.de

Abstract

Mainstream cross-lingual task-oriented dialogue (ToD) systems leverage the transfer learning paradigm by training a joint model for intent recognition and slot-filling in English and applying it, zero-shot, to other languages. We address a gap in prior research, which often overlooked the transfer to lower-resource colloquial varieties due to limited test data. Inspired by prior work on English varieties, we craft and manually evaluate perturbation rules that transform German sentences into colloquial forms and use them to synthesize test sets in four ToD datasets. Our perturbation rules cover 18 distinct language phenomena, enabling us to explore the impact of each perturbation on slot and intent performance. Using these new datasets, we conduct an experimental evaluation across six different transformers. Here, we demonstrate that when applied to colloquial varieties, ToD systems maintain their intent recognition performance, losing 6% (4.62 percentage points) in accuracy on average. However, they exhibit a significant drop in slot detection, with a decrease of 31% (21 percentage points) in slot F_1 score. Our findings are further supported by a transfer experiment from Standard American English to synthetic Urban African American Vernacular English.

1 Introduction

The usability of dialog systems heavily relies on the ability to handle user inputs in multiple languages. Recent language models (LMs) have become state-of-the-art tools to carry out the primary task-oriented dialogue (ToD) problems, including intent recognition and slot filling. What is more, LMs leverage multilingual pre-training to facilitate transfer across languages. To achieve this, the mainstream approach involves fine-tuning the LM on a pivot language, commonly English, and subsequently employing the LM in a zero-shot manner to process target languages (Hu et al., 2020).

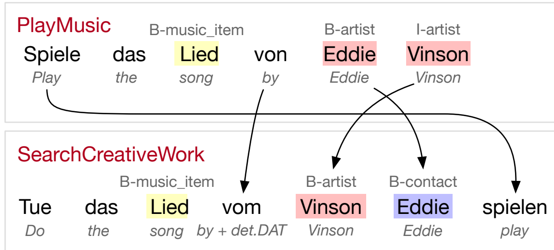


Figure 1: An illustrative example selected from xSID. The top part displays the intact sentence with gold labels, the bottom part shows the prediction for the perturbed sentence. The perturbations `tun_imperative`, `article_name`, `name_order` are applied. There are errors in predicting the intent and one of the two slots.

While this language transferring approach has achieved impressive results for many language pairs, its effectiveness is limited when it comes to processing low-resource language varieties and dialects (Hedderich et al., 2021). These varieties are often underrepresented in the LM's pre-training data and may not align well with the characteristics of the chosen pivot language. Our current understanding of how well modern LMs handle dialects and the extent of disparity between standard languages and dialects remains limited. Therefore, it is important to assess the performance gap in the first place, as highlighted by Kantharuban et al. (2023) to identify key directions for further development.

Processing (non-standardized) dialects brings unique challenges: large volumes of writing such as newspapers or fiction are rarely produced, and access to conversational data in social media is limited. Besides, dialects lack unified spelling rules (Millour and Fort, 2019) and exhibit a high degree of variation over space and time (Dunn and Wong, 2022). Finally, dialects may additionally show a significant rate of code-mixing compared to standard languages (Muysken et al., 2000).

[*] Now at Toloka.AI

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 445–468
March 17-22, 2024

To tackle these challenges, recent studies have introduced techniques that mimic dialectal morphosyntactic variation through rule-based translation systems which perturb sentences into respective dialect variants in German (Gerlach et al., 2022) and English (Ziems et al., 2022). This approach is highly practical as it avoids the expense of annotating new data while still effectively stress-testing applications like question answering and machine translation (Ziems et al., 2023). Building on this, we choose ToD as a task where we expect a high level of linguistic variation in real-life application settings (Trong et al., 2019; Aepli et al., 2023). We experimentally evaluate how well ToD systems handle dialectal data by simulating dialectal and colloquial variations in English and German to explore the following research questions (RQs). **RQ 1:** How does the LM performance in intent recognition and slot filling change when applied to synthetic dialectal data in both English and German? **RQ 2:** Considering that each perturbation isolates a specific dialectal phenomenon, which perturbations have the most significant effect? **RQ 3:** How do LMs differ in terms of robustness to dialectal perturbations? Figure 1 illustrates our approach.

To address these RQs, we contribute in the following ways: *(i)* We define and implement a set of hand-crafted perturbation rules for translating from Standard German to its spoken varieties (§3.2). *(ii)* We systematically test a range of perturbations, each representing distinct dialectal phenomena, in two languages, to quantify their individual effect on ToD performance (§5, RQ 1&2). *(iii)* We provide an extensive analysis of joint intent recognition and slot filling experiments using a diverse set of cross-lingual encoders in two languages (§5, RQ 3).

We release the code for the perturbation rules and the results of our experimental evaluation for further uptake: `github.com/mainlp/dialect-ToD-robustness`.

2 Related Work

Robustness of ToD systems. The evaluation of Task-oriented Dialogue (ToD) systems' robustness aims to investigate the generalization capabilities of LMs and their ability to adapt to domain shifts, with a specific focus on English (Chang et al., 2021). The robustness of ToD systems has been widely investigated using adversarial attacks, which involve manipulating the gradients and weights of LMs to alter their predictions (Cheng et al., 2019).

Nevertheless, white-box methods lack linguistic awareness, making them not easily interpretable (Zeng et al., 2021). In contrast, recent black-box methods have emerged that aim to mimic language variation and real-life noise, including speech artifacts and typos, with the primary objective of crafting instances that deceive LMs (Lee et al., 2022; Liu et al., 2021; Peng et al., 2021; Cho et al., 2022).

A related line of research focuses on developing defenses against adversarial attacks and enhancing the robustness of ToD systems by employing techniques such as data augmentations and incorporating regularization terms in the loss function (Einolghozati et al., 2019; Sengupta et al., 2021).

NLP for dialects and non-standard varieties. Previous efforts in processing dialects and non-standard varieties have primarily focused on differentiating between dialects and closely related languages. Notably, the VarDial initiative (Gaman et al., 2020; Chakravarthi et al., 2021; Aepli et al., 2022, 2023) has conducted a series of evaluation campaigns aimed at dialect identification and discrimination between similar languages. Additional research directions in the field include part-of-speech (POS) tagging (Hollenstein and Aepli, 2014; Zampieri et al., 2019), syntactic parsing (Blodgett et al., 2018), low-resource intent identification and slot filling (Aepli et al., 2023). Moreover, machine translation techniques have been applied to re-write sentences from dialect to standard language (Kchaou et al., 2022; Plüss et al., 2020; Lambrecht et al., 2022). To overcome the limited availability of parallel training data, rule-based perturbations simulating dialectal morphosyntactic phenomena have been developed to generate synthetic parallel sentence pairs (Gerlach et al., 2022).

The emergence of pre-trained LMs has shifted the focus towards investigating disparities in representation and downstream performance between non-standard and standard languages. To this end, LM diagnostic tools encompass a wide range of techniques, including cloze tests (Zhang et al., 2021) and contrastive evaluation via minimal pairs (Demszky et al., 2021). Ziems et al. (2022, 2023) have created a rule-based translation system that converts English into various dialects. They use this system to conduct stress tests on multiple downstream models and reveal performance disparities between English dialects.

Methods to improve LMs' robustness towards dialects include integrating morphological information into LMs' tokenizers through inflection perturbations (Tan et al., 2020a,b), manipulating the parse tree of the source sentence to align with the word order in the target dialect (Wang and Eisner, 2016; Wu et al., 2023), and character noise injection (Aepli and Sennrich, 2022). Using perturbed data during LM pre-training or adapter training has shown significant benefits for dialectal variants of the GLUE tasks (Wang et al., 2018) specifically designed to dialects (Held et al., 2023).

A related line of research concentrates on the processing of spoken dialects, with a specific emphasis on dialectal speech to standard language recognition (Samardžić et al., 2016; Plüss et al., 2022) and spoken dialect identification (Zampieri et al., 2019).

3 Perturbations Based on Dialect Variations

In this section, we introduce perturbations that are specifically motivated by dialectal variation. In English and German, these perturbations specifically focus on altering the morphosyntactic structure of the sentence to simulate dialects, while keeping the semantics unchanged.

3.1 English Perturbations

We re-use a set of perturbations obtained from the Multi-VALUE framework (Ziems et al., 2023),[1] which translate text from Standard American English (SAE) to Urban African American Vernacular English (UAAVE). This set comprises a total of 118 perturbations, covering morphosyntactic phenomena present in UAAVE. The quality of the perturbation-based translation system is evaluated through prior human evaluation. These patterns are documented in and sourced from the Electronic World Atlas of Varieties of English (eWAVE, Kortmann et al., 2020), which lists 235 features from 75 English varieties, collected by 87 professional linguists in 175 peer-reviewed publications.

3.2 German Perturbations

Aligned with the Multi-VALUE framework, we implement a set of perturbations designed to translate text from Standard German into non-standard varieties. Since there is *no resource* detailing syntactic variations in German varieties similar to those for other languages such as English (Kortmann et al., 2020), North Germanic languages (Lundquist et al., 2019), creole and pidgin languages (Michaelis et al., 2013) or South American languages (Muysken et al., 2016), we review over thirty linguistic works published in the last decades.[2] We select a set of morphosyntactic features that include different grammatical areas and features both regional and supraregional variation. Similarly to the work by Ziems et al. (2022), our feature set is meant to showcase different types of variation rather than being exhaustive.

Table 1 presents an overview of the perturbations, along with examples and pointers to relevant linguistic literature for further reference.[3] We group the perturbations according to eWAVE's category definitions and de-facto category assignments of similar English examples.[4] Several of our rules target grammatical areas that are not covered by eWAVE/Multi-VALUE, sometimes in ways relevant to the ToD context. For instance, we also include changes to adpositions (relevant for labeling slots in queries relating to flight itineraries) and personal names (pertinent for queries like calling a contact or checking a birthday).

We include features that are common and unmarked in colloquial German across all of the German-speaking area (such as eliding the word-final schwa in inflected verbs), as well as some that are specific only to certain non-standard dialects (such as the choice of directive or locative preposition). Some of these features cannot be easily placed on this scale of regional specificity, as they might be licensed in more construction types in some areas than in others (like the progressive tense constructed with the preposition *am*; Auer, 2003). In total, we developed 18 perturbations that cover a wide range of phenomena.

Implementation. Perturbation rules are implemented as rule-based functions that modify input sentences according to morphosyntax parses. For part-of-speech (POS) tagging and dependency parsing, we employ German SoTA models in spaCy (Honnibal et al., 2020) and Stanza (Qi et al., 2020). Noun inflection is handled using Derbi (Schmaltz, 2022), verb conjugation is conducted with Pattern-

[1] Usage terms at `https://value-nlp.org/`.

[2] While German dialectology has traditionally focused more on phonological/phonetic and lexical variation, we take advantage of the popularity that dialect syntax studies have gained in the past decades (cf. Glaser, 1997; Scheutz, 2005).

[3] For a general introduction to syntactic variation in colloquial varieties of German, see Fleischer (2019).

[4] For instance, our `comparative` feature resembles eWAVE features 82 and 85.

Category	Perturbation	Example: Before → After	Source
Noun Phrase			
	`possession_von` *von* construction instead of genitive	des Baums → von dem Baum *the.*GEN *tree's → of the.*DAT *tree*	Bülow et al. (2021); Eichhoff (2000, map 77)
	`possession_pron` Dative with poss. pron. instead of genitive	Kafkas Werke → Kafka seine Werke *Kafka's works → Kafka.*DAT *his works*	Bülow et al. (2021); Eichhoff (2000, map 77)
	`article_name` Article before personal names	Franz Kafka → der Franz Kafka *Franz Kafka → the Franz Kafka*	Fleischer (2019); Eichhoff (2000, map 76)
	`comparative` Comparitives with *wie* or *als wie*	größer als → größer wie *bigger than*	Jäger (2018)
	`double_det` Emphatic double article	ein so großer Baum → ein so ein großer Baum *a such big tree → a such a big tree*	Auer (2003)
Discourse & Word Order			
	`name_order` Swapped family and given names	Franz Kafka → Kafka Franz	Auer (2003)
	`denn` Obligatory particle *denn* in questions	Wie ist das Wetter? → Wie ist denn das Wetter? *How is the weather? → How is* PART *the weather?*	Fleischer (2019)
	`verb_clusters` Raised auxiliary/modal in 2-verb clusters	da sie das getan hat → da sie das hat getan *because she it done had → because she it had done* 'because she had done it'	Bader and Schmid (2009)
Tense & Aspect			
	`progressive` Progressive construction with *am*	ich koche Suppe → ich bin Suppe am kochen *I cook soup → I am soup* PREP *cooking*	Flick and Kuhmichel (2013); Fleischer (2019)
Adverbs & Prepositons			
	`pronominal_adverbs` Splitting of pronominal adverbs with *da-*	davon weiß ich nichts → da weiß ich nichts von *of.this know I nothing → there know I nothing of*	Fleischer (2002)
	`direction` Directive preposition *auf*	nach München → auf München *to Munich*	Merkle (1993, p. 185); Elspaß and Möller (2003–, entry 12/4g)
	`location` Locative preposition *zu*	in München → zu München *in Munich*	Merkle (1993, p. 186)
Negation			
	`negative_concord` Negative concord	ich sehe kein Haus → ich sehe kein Haus nicht *I see no house → I see no house not* 'I don't see any house'	Fleischer (2019); Auer (2003)
Relativization			
	`relative_pron` Relative marker *wo*	der Stern, der funkelt → der Stern, wo funkelt *the star* REL *sparkles*	Moser (2023)
Complementation			
	`es_hat` Existential clause *es hat*	es gibt noch Brot → es hat noch Brot *it gives still bread → it has still bread* 'there is still bread left'	König et al. (2015, p. 243)
Verb Morphology			
	`tun_imperative` Periphrastic imperatives with *tun* 'do'	räum auf → tu aufräumen *tidy.*2SG.IMP *up → do.*2SG.IMP *tidy.up.*INF	Merkle (1993, p. 66) *
	`schwa_elision` Schwa elision at the end of 1.SG.PRES verbs	ich habe → ich hab *I have*	Keel (1980)
Pronouns			
	`clitic_es` Enclitic form of *es* 'it' after inflected verbs	ist es → ist's *is it*	Abraham (1996)

Table 1: Our collection of syntactic perturbations, sorted according to eWAVE's categories (in **bold face**). We give examples in German, with glosses in *gray italics*. *This feature, `tun_imperative`, is also inspired by systematic variation we could observe between the Standard and Swiss German versions of one of the datasets we use, xSID (van der Goot et al., 2021a; Aepli et al., 2023).

de[5] (De Smedt and Daelemans, 2012). We incorporate the list of first names from Nett et al. (2019). Refer to Appendix A for examples of automatically perturbed sentences.

Human evaluation. We create a human evaluation dataset by manually labelling up to eight sentences per perturbation from each dataset. As certain rules can only be applied to fewer than eight sentences in some datasets, the human evaluation dataset comprises 200 sentences in total.

These sentences are assessed for fluency on a five-point Likert scale, where a score of 5 means that perturbed sentences are highly fluent and natural, while a score of 1 indicates the opposite. Appendix E presents the annotation guidelines. The annotations are carried out by two native German speakers with a background in computational linguistics and significant exposure to diverse dialects.[6]

When evaluating the inter-annotator agreement based on raw scores, the percentage of cases where both annotators assign the same score is 53.51% and the Pearson correlation coefficient is 0.51. Overall, the scores provided by both annotators average at 3.92 and 4.63. In 96 (48%) and 3 (1.5%) cases, both annotators give a score of 5 and 1 to the same sentence, respectively. Notably, the perturbations `verb_clusters` shows significant disparity, with the mean score assigned by one annotator being 1, while the other annotator assigned a mean score of 5.[7] Below is an example of a sentence pair that the annotators judged with opposite scores (1 vs. 5). *A* is for German, *B* is for the dialect rewrite. The fragment of the sentence affected with the `verb_clusters` perturbation is underlined.

A Frag ob Pauline zu meinem Thanksgiving -
Ask if Pauline to my Thanksgiving -
Treffen <u>kommen</u> <u>will</u> .
gathering come.INF *wants* .

B Frag ob Pauline zu meinem Thanksgiving -
Ask if Pauline to my Thanksgiving -
Treffen <u>will</u> <u>kommen</u> .
gathering wants come.INF .

[5] digiasset.org/pattern-de

[6] One annotator is one of the authors. The second annotator was hired and received fair compensation according to the local employment regulations.

[7] This feature is regionally very specific (Elspaß and Möller, 2003–, entry 3/13abc). The annotator providing high rankings is not from an area using this construction but was familiar with relevant literature and examples beforehand. The other annotator, unfamiliar until a pre-task explanation, gave lower rankings.

Similar discrepancies are observed in other perturbations such as `pronominal_adverbs`, `relative_pron`, and `name_order`.

Additionally, we map the score to a binary scale (where scores 1 and 2 were grouped as 0, and scores 3, 4, and 5 were grouped as 1). The exact match agreement becomes 91.89%. Cohen's kappa (McHugh, 2012) reaches a 0.61. Areas of disagreement include `verb_clusters` and `progressive`. These perturbations account for the majority of the discrepancies, with 7 items and 4 items respectively. The results indicate moderate to substantial levels of agreement between annotators and shed light on which perturbations tend to cause the most disagreement. Since linguistic acceptability in the context of language variation can be subjective, we chose to keep all perturbations, even if there were disagreements among annotators.

4 Methodology

We choose task-oriented dialogue systems as a task where we expect a high level of linguistic variation in real-life application settings. There is limited research on whether these systems commonly encounter inputs from dialect speakers in real-world applications (Bird, 2020; Nekoto et al., 2020). Nevertheless, several works encourage the localization of dialogue systems to dialect varieties. One common motivational aspect shared by these works is the aim to encourage the use of dialects, with the expectation of positively impacting the prestige of the language (Trong et al., 2019; Aepli et al., 2023).

Datasets. Table 2 provides a brief description of the ToD datasets for intent recognition and slot filling. All of the datasets considered support zero-shot cross-lingual setups by including English training and German development and test data. Except for xSID, all datasets are further equipped with German training data. In this study, we concentrate on German and English, leaving other languages for future work.

Method. We adopt a joint approach for intent detection and slot filling, leveraging the implementation of MaChAmp (van der Goot et al., 2021b). It uses an encoder and a separate decoder head for each task, one for intent classification and one for slot detection with a CRF layer on top. We use the default settings, which include a learning rate of 0.0001. We experiment with six encoder-based multilingual LMs (Table 3). Each LM undergoes training with five random seeds, and results are av-

Label	Source	# Langs.	Domain	# Intents	# Slots	Train / dev / test	DE tr?	License
xSID	van der Goot et al. (2021a) Aepli et al. (2023)	15	General	16	33	43k / 300 / 500		CC BY-SA 4.0
MultiATIS++	Xu et al. (2020)	9	Aviasales	18	84	3.7k / 1.2k / 893	✓	Apache 2.0
MASSIVE	Bastianelli et al. (2020) FitzGerald et al. (2023)	51	Virtual assistant, smart home	60	55	11k / 2k / 3k	✓	Apache 2.0
MTOP	Li et al. (2021)	6	Virtual assistant	117	78	16k / 1.8k/ 3.5k	✓	CC BY-SA 4.0

Table 2: The datasets, used for experiments. Key: **# langs.** is the number of languages included in the dataset. **# intents** and **# slots** stands for the the number of intents and slots in the dataset. **Train/dev/test** is the number of sentences in train, validation and test sets. **DE tr?** indicates whether training data in German is available.

Label	HuggingFace ID (Wolf et al., 2020)	Source	# Params.	Tr. data	Dialect?	License
mBERT	`bert-base-multilingual-cased`	Devlin et al. (2019)	177M	Wiki	✓	Apache 2.0
XLM-R	`xlm-roberta-base`	Conneau et al. (2020)	279M	CC		MIT
RemBERT	`google/rembert`	Chung et al. (2020)	575M	Wiki+CC		Apache 2.0
mDeBERTa	`microsoft/mdeberta-v3-base`	He et al. (2021a,b)	276M	CC		MIT
DistilmBERT	`distilbert-base-multilingual-cased`	Sanh et al. (2019)	134M	Wiki	✓	Apache 2.0
mMiniLM	`microsoft/Multilingual-MiniLM-L12-H384`	Wang et al. (2020)	117M	CC		MIT

Table 3: The cross-lingual LMs used in the study. Key: **Tr. data** denotes pre-training datasets, where Wiki stands for Wikipedia, CC stands for CommonCrawl (Wenzek et al., 2020). **Dialect?** indicates whether German dialect data was explicitly included in the LM's pre-training data. The dashed line separates the base-size LMs from the distilled LMs. DistilmBERT is distilled from mBERT, mMiniLM is distilled from XLM-R.

eraged across all runs. LMs are trained on a single NVIDIA A100 device.

Experimental setup. Evaluation metrics are accuracy for intent recognition and the span F_1 score for slot filling, where both span and label must match exactly. We explore three experimental setups: *(i)* zero-shot setup: models are trained on English training data; *(ii)* zero-shot setup with German development data; *(iii)* fully supervised setup (where available): models trained on German training data.

Model selection over epochs is based on its performance on development data in English *(i)* and German *(ii, iii)*, without any access to labeled UAAVE or German data during the training phase.

To assess the robustness of the ToD model, we apply perturbations to generate synthetic UAAVE and German dialect test data. We then use fine-tuned models to make predictions on this perturbed data. We evaluate the impact of these perturbations by measuring *the difference in performance before and after* the perturbation is applied. In addition, following the research on adversarial attacks (Tsai et al., 2019) we define the *success rate* of a perturbation as the number of instances that become misclassified after the perturbation was applied.

5 Results

RQ 1: What is the impact of perturbed data on performance? Table 4 and Table 6 (Appendix B) present the intent recognition and slot filling test results for zero-shot *(i)* German and English, respectively, with and without perturbations. Additionally, in Appendix B, Table 7 displays the results for setup *(ii)*, while Table 8 presents the fully-supervised German setup *(iii)*. The performance scores align with earlier results reported in the dataset papers and recent research (Aepli et al., 2023). The perturbations are used in two scenarios: *(a)* with 18 German and 118 English perturbations applied individually and average performance computed across them,[8] *(b)* with all perturbations applied simultaneously.

Table 4 shows the performance gap[9] in zero-shot evaluation on test sets before and after German perturbations are applied concerning the dataset and the LM. The decrease in performance is minimal for intent recognition accuracy, averaging at 0.33, when individual perturbations are applied. However, it drops further by an average of 4.62 when all perturbations are applied simultaneously. The drop is more pronounced for slot filling, where

[8]While some of the syntactic features tend to co-occur, e.g., the `name_order` swap is most commonly found in varieties that also exhibit the `article_name` feature (Elspaß and Möller, 2003–, entry 10/16ab). We nevertheless apply rules individually in scenario *(a)*, as the borders between feature areas do not form perfect isoglosses. In the given example, name swapping without any added article is attested in some locations near the Belgian and Dutch borders (ibid.).

[9]All of the performance changes detailed in the following are in percentage points.

		Intact		Individual Perturbations		All Perturbations	
		Intent Acc	Slot F_1	Δ Intent Acc	Δ Slot F_1	Δ Intent Acc	Δ Slot F_1
xSID	mBERT	76.36	70.57	0.40	2.32	5.60	20.70
	XLM-R	90.20	76.23	0.31	2.70	4.08	22.95
	RemBERT	91.08	79.44	0.34	2.78	4.16	23.59
	mDeBERTa	94.88	82.62	0.24	2.69	3.12	23.03
	DistilmBERT	71.04	66.62	0.43	2.17	4.88	19.94
	mMiniLM	72.16	69.29	0.34	2.25	3.56	22.36
MultiATIS++	mBERT	76.91	62.22	0.07	2.50	0.81	9.57
	XLM-R	78.75	76.18	0.02	3.72	0.18	11.13
	RemBERT	79.28	83.32	0.01	4.05	0.27	15.95
	mDeBERTa	79.17	80.10	0.01	3.89	0.27	10.93
	DistilmBERT	74.67	56.72	0.05	2.38	0.43	8.74
	mMiniLM	74.65	68.49	0.00	3.12	0.25	9.21
MASSIVE	mBERT	54.63	49.25	0.43	2.38	5.74	21.56
	XLM-R	74.86	65.75	0.42	2.80	6.70	26.47
	RemBERT	83.86	73.33	0.41	3.02	6.29	27.64
	mDeBERTa	83.91	73.86	0.39	3.02	6.29	28.08
	DistilmBERT	45.53	42.74	0.38	1.99	4.42	19.30
	mMiniLM	58.14	54.57	0.30	2.44	5.34	23.02
MTOP	mBERT	67.34	66.99	0.51	2.58	8.20	26.96
	XLM-R	88.76	77.53	0.60	2.96	8.88	30.44
	RemBERT	91.35	79.33	0.58	3.05	8.79	31.41
	mDeBERTa	90.66	79.26	0.60	2.95	8.24	30.50
	DistilmBERT	58.72	59.71	0.46	2.50	7.32	25.79
	mMiniLM	75.89	70.53	0.52	2.79	7.17	29.13
Mean		76.37	69.36	0.33	2.79	4.62	21.60

Table 4: The overall results for intent recognition and slot filling on test sets in German in zero-shot setup *(i)* and the gap in performance before and after dialect perturbations are applied (in percentage points). Intact (left): performance on intact test sets. Individual perturbations (middle): 18 individual perturbations are applied and average performance gap is computed across them. All perturbations (right): all perturbations applied simultaneously. Δ denotes the difference between performance on intact and perturbed data. Performance on intact data consistently surpasses that on perturbed data, leading to positive Δ values. The results are averaged across five runs with varying random initialization.

performance decreases by 2.79 Slot F_1 after individual perturbations and by 21.60 Slot F_1 after the simultaneous application of all perturbations.

In the evaluation for English (Table 6, Appendix B), we observe similar trends. The decline in intent recognition is minimal, with average drops of merely 0.10 up to 2.48 accuracy in the two considered scenarios. Conversely, the decline in slot filling is more pronounced, with 9.87 and 49.37 F_1 score on average for individual and combined perturbations, respectively. The simultaneous application of all perturbations affects the performance more than applying individual perturbations.

Further experiments with setup *(ii)* show that the choice between English or German development data has no significant impact on the performance on perturbed data (compare Table 4 with Table 7, Appendix B). In particular, while zero-shot downstream performance improves for all LMs (e.g. mDeBERTa and RemBERT, show gains of 0.61 accuracy and 0.18 F_1 score and 0.77 accuracy and 2.36 F_1 score, respectively), the impact of the perturbations remains similar with comparable results to the results discussed earlier in the setup *(ii)* (higher impact on slots than intents).

In the fully-supervised setup *(iii)* with fine-tuning on German data (Table 8, Appendix B), we observe an expected significant improvement in performance across all three datasets, due to the in-language training data. While the performance drop is almost identical to the zero-shot set-up for intent accuracy, the slot filling performance is considerably more robust. Here, the average drop is only 6.27 F_1 when all perturbations are applied

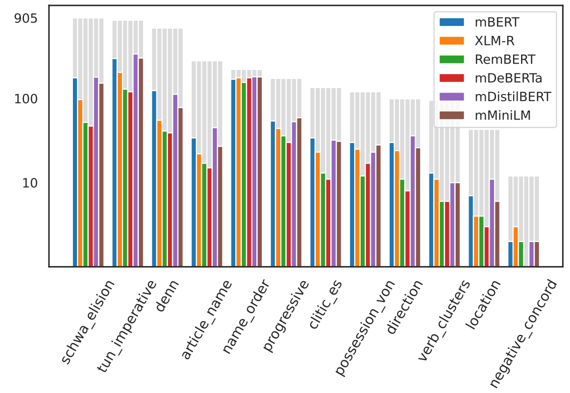


Figure 2: Intent prediction success rates on the perturbed German test set on MASSIVE with respect to most impactful individual perturbations. The grey bars denote the count of perturbed sentences, the colored bars show the success rate. A logarithmic scale is used.

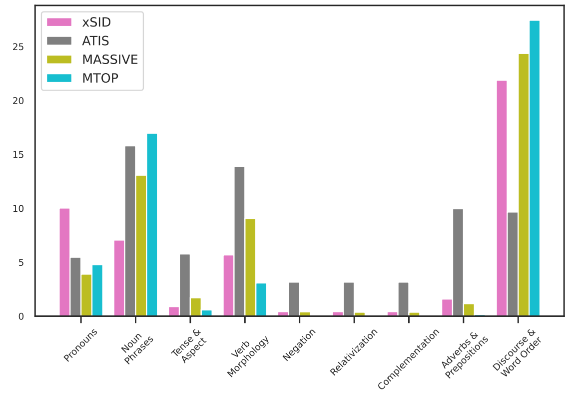


Figure 3: The Δ slot F_1 score of the best performing mDeBERTa with respect to perturbation category in perturbed German test set in four datasets. Δ denotes the difference in F_1 score between performance on intact and perturbed data.

(compared to 21.60 in the zero-shot set-up). This suggests that fine-tuning with in-language data improves performance on both intact and perturbed test sets.

To sum up, while LMs can still produce accurate predictions on the sentence level after the sentence is perturbed with dialectal variations (i.e., intent recognition), their performance suffers particularly on the word level (i.e., slot filling), and this becomes more pronounced as the sentence's perturbation increases. Fine-tuning with in-language data improves overall performance and enhances significantly the treatment of perturbed data. These findings remain consistent across all four datasets and the various LMs considered.

RQ 2: Which perturbations affect performance the most? This part focuses on the zero-shot scenario *(i)*. First, we examine perturbations that result in a non-zero perturbation success rate, indicating their ability to change the predicted intent. Figure 2 illustrates the success rate of 12 individual perturbations on the German test set of MASSIVE, compared with the count of perturbed sentences. The six remaining perturbations do not affect the performance and have zero success rate. While all perturbations preserve semantics, those with higher success rates induce a more substantial shift in the representation space and effectively fool LMs. The perturbations `schwa_elision` and `tun_imperative` impact a similar number of sentences, yet their success rates differ, with the latter exhibiting a higher success rate. This could be attributed to the alteration in the number of words in `tun_imperative` and the change in the position of the main content word, shifting from the first to the final position in the sentence (see the example in Figure 1). The `name_order` perturbation exhibits the highest success rate, while the `negative_concord` perturbation demonstrates the lowest non-zero success rate. The analysis of success rates in German and English across various datasets (Figures 4 and 5, Appendix C) confirms that the frequency of perturbations differs across datasets due to their design. However, the success rates remain consistent. There are frequent perturbations that have little impact, such as `location` and `direction` in German (except for MultiATIS++, see below), and `zero_plural` in English. Some perturbations demonstrate consistently stable success rates in all four datasets, as observed in the case of `progressive` in English and `word_order` in German. This could be linked to the frequency of respective dialect phenomena in the LM's pre-training data, where rarely seen dialect phenomena deceive it more effectively.

Figure 3 examines how the F_1 score declines after individual perturbations are applied. Here, the perturbations are grouped according to eWAVE categories, and mDeBERTa serves as the backbone LM. Across datasets, the F_1 score is mostly affected by the three perturbations falling under the **Discourse & Word Order** category, followed by perturbations affecting **Noun Phrases** and **Verb Morphology**. In turn, in English the **Tense & Aspect** category stands out, followed by **Pronouns** and **Noun Phrases** (Appendix D).

There are structural and domain-specific variations in performance across datasets. In xSID, the **Pronouns** category experiences a significant impact, indicating a higher frequency of the usage of *es* 'it' (shortened to *'s* by our perturbations) compared to other datasets. In MultiATIS++, the **Adverbs & Prepositions** category is notably affected. This category includes perturbations that modify directive and locative prepositions, which are commonly employed in MultiATIS++ due to its specific domain (with queries like "What are flights *to* X that also stop *in* Y?").

RQ 3: How does the performance of LMs vary? Table 4 shows that in the zero-shot setup mDeBERTa consistently outperforms other LMs, followed closely by RemBERT. XLM-R and mBERT also exhibit competitive performance, while DistilmBERT and mMiniLM tend to have lower scores. There is a consistent drop in performance when dialect perturbations are applied, indicating that all LMs are sensitive to dialectal variations. Figure 2 exhibits similar trends across all LMs, with mDeBERTa and RemBERT displaying comparatively lower success rates for individual perturbations. Conversely, distilled models, DistilmBERT and mMiniLM, show higher success rates.

Our results suggest that mDeBERTa and RemBERT are more robust to dialectal variations, outperforming other LMs in both tasks across four datasets. This aligns with previous cross-lingual studies (Adelani et al., 2022; Malmasi et al., 2022), where they outperformed other LMs and demonstrated superior results in lower-resource settings.

Error analysis. Next, we focus on German for error analysis. In intent recognition, LMs often confuse semantically similar intents (PLAYMUSIC and SEARCHCREATIVEWORK), or intents associated with the same service, (ALARM/CANCEL_ALARM and ALARM/SET_ALARM, xSID). These errors become apparent when the LMs are tested on intact data and become even more pronounced when dialect perturbations are applied. Lastly, LMs tested on perturbed data tend to misinterpret intents that commonly share homonymous words (BOOKRESTAURANT and RATEBOOK, xSID).

There are three primary errors in slot filling. Firstly, the LMs incorrectly identify slot boundaries when perturbations impact word order. In such cases, the LM tends to make errors in predicting slot boundaries, as observed in instances like "Merkel Angela" (B-PERSON I-PERSON) transformed from "Angela Merkel", where the LMs often predict B-PERSON B-PERSON, splitting the span inaccurately. Secondly, when the word order is maintained, the LMs exhibit more mistakes in predicting slot types. For instance, when the direction perturbation is applied, the LMs frequently assign incorrect slot types. Finally, when an extra auxiliary verb is introduced, as in the case of the `progressive` perturbation, LMs frequently assign it a slot label.

6 Conclusion and Future Work

This project tests the robustness of task-oriented dialogue systems (ToD) towards English and German dialects. Our methodology involves applying rule-based perturbations to translate ToD datasets from Standard American English to Urban African American Vernacular English, and from German to its non-standard variety. To the best of our knowledge, we are the first to design such perturbations for German. Subsequently, we train multiple joined ToD models, equipped with various Transformer-based backbones, assessing their performance on intact and perturbed data.

We conclude, that **Re RQ 1:** The impact of perturbed data on LM performance varies depending on the type of perturbation and the task. In general, we note a minor decrease in intent recognition but a notable drop in slot filling. Issues in slot filling involve inaccuracies in boundary identification, mistakes in predicting slot types with altered word order, and frequent misalignments of slot labels with an extra verb. **Re RQ 2:** Across languages, the performance drop varies by dataset and LM, indicating domain and language-specific patterns in response to phenomena-based perturbations. **Re RQ 3:** There is no clear winner, but mDeBERTa and RemBERT outperform other LMs by gaining higher performance scores and being more robust to dialectal variations.

Future work includes (i) extension to other languages with distinct dialectal variation; (ii) development of fair evaluation approaches, that do not favor standard languages but account for dialects; (iii) incorporating phonological phenomena for a deeper understanding of dialectal variations in written and spoken forms.

Conducting similar experiments with other languages and dialects can help in understanding how these models generalize across diverse linguistic landscapes.

Limitations

Focus on written text. Our study predominantly focuses on written text, and we do not account for phonological or lexical differences between the standard language and non-standard varieties. Our emphasis is primarily on syntactic differences, and as such, we acknowledge that our analysis may not fully capture the complete spectrum of linguistic nuances present in spoken language variation.

Choice of LMs. Our choice of LMs is inherently limited; we do not use auto-regressive or sequence-to-sequence language models for the sake of compute time.

German perturbations. The selected German perturbations do not perfectly capture any particular German dialect, but they are based on prevalent patterns found in a selection of dialects and colloquial varieties.

Design of perturbations. The perturbation rules, borrowed from Ziems et al. (2023) for English and developed by us for German, specifically target syntactic phenomena, excluding orthographic and lexical variations.

Focus on zero-shot settings. In our approach, the primary focus is on zero-shot settings, where dialect data is intentionally excluded from the training process to prevent any potential leakage. This choice allows us to follow a practical scenario where the model can handle diverse dialects without the need for collecting specific dialect data during training. However, deviating from the zero-shot setting could potentially yield models that are more robust to direct perturbation. In such cases, the upper bound for evaluating robustness would involve incorporating dialect training data, providing an alternative perspective to the zero-shot approach.

Ethical considerations

Human assessment. This work involves human assessment of synthetically generated data. Two annotators were involved. One annotator is one of the authors. The second annotator was hired and received fair compensation according to the local employment regulations.

Perturbation rules. Our software allows automatically applying changes to German sentences that simulate dialectal and colloquial variation. Our selection of perturbation rules is not exhaustive enough to simulate any one dialect and is taken to be representative of the breadth of variation in the German dialect landscape. Because of these restrictions, we find it unlikely that our system could be used for the mockery and parody of any dialects or registers. We release the code for perturbations for research purposes only and expressly forbid usage for mockery or parody of any dialects or registers.

Acknowledgements

We thank our colleagues from the MaiNLP research group and the anonymous reviewers for their feedback. This research is supported by the ERC Consolidator Grant DIALECT 101043235.

References

Werner Abraham. 1996. Personalpronomina, Klitiktypologie und die Struktur des 'Mittelfeldes'. In Ewald Lang and Gisela Zifonun, editors, *Deutsch – typologisch*, pages 428–470. de Gruyter.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*

(VarDial 2023), pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Peter Auer. 2003. Non-standard evidence in syntactic typology – methodological remarks on the use of dialect data vs. spoken language data. volume 153 of *Trends in Linguistics. Studies and Monographs*, pages 69–92. De Gruyter Mouton, Berlin, New York.

Markus Bader and Tanja Schmid. 2009. Verb clusters in colloquial German. *Journal of Computational Linguistics*, 12:175–228.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Lars Bülow, Philip C. Vergeiner, and Stephan Elspaß. 2021. Structures of adnominal possession in Austria's traditional dialects: Variation and change. *Journal of Linguistic Geography*, 9(2):69–85.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.

Kai-Wei Chang, He He, Robin Jia, and Sameer Singh. 2021. Robustness and adversarial examples in natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 22–26, Punta Cana, Dominican Republic & Online. Association for Computational Linguistics.

Minhao Cheng, Wei Wei, and Cho-Jui Hsieh. 2019. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3325–3335, Minneapolis, Minnesota. Association for Computational Linguistics.

Hyundong Cho, Chinnadhurai Sankar, Christopher Lin, Kaushik Sadagopan, Shahin Shayandeh, Asli Celikyilmaz, Jonathan May, and Ahmad Beirami. 2022. Know thy strengths: Comprehensive dialogue state tracking diagnostics. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5345–5359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Tom De Smedt and Walter Daelemans. 2012. Pattern for Python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Dunn and Sidney Wong. 2022. Stability of syntactic dialect classification over space and time. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 26–36, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jürgen Eichhoff. 2000. *Wortatlas der deutschen Umgangssprachen*, volume 4. De Gruyter Saur, Berlin, Boston.

Arash Einolghozati, Sonal Gupta, Mrinal Mohit, and Rushin Shah. 2019. Improving robustness of task oriented dialog systems. *arXiv preprint arXiv:1911.05153*.

Stephan Elspaß and Robert Möller. 2003–. Atlas zur deutschen Alltagssprache (AdA).

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Jürg Fleischer. 2002. *Die Syntax von Pronomialadverbien in den Dialekten des Deutschen: Eine Untersuchung zu Preposition Stranding und verwandten Phänomenen*, chapter 2.1 Spaltungskonstruktion bei 'da'. Franz Steiner Verlag.

Jürg Fleischer. 2019. Vergleichende Aspekte der deutschen Regionalsprachen: Syntax. volume 4 of *Sprache und Raum – Ein internationales Handbuch der Sprachvariation*, pages 635–664. De Gruyter Mouton, Berlin, Boston.

Johanna Flick and Katrin Kuhmichel. 2013. Der *am*-Progressiv in Dialekt und Standardsprache. *Jahrbuch für Germanistische Sprachgeschichte*, 4(1):52–76.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Johanna Gerlach, Jonathan Mutal, and Bouillon Pierrette. 2022. Producing Standard German subtitles for Swiss German TV content. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 37–43, Dublin, Ireland. Association for Computational Linguistics.

Elvira Glaser. 1997. Dialektsyntax: eine Forschungsaufgabe. In Peter Ott, Thomas A Hammer, Ruth Jörg, Niklaus Bigler, Hans-Peter Schifferle, and Andreas Burri, editors, *Bericht über das Jahr 1996*, pages 11–30. Schweizerischen Idiotikon, Zürich.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. DeBERTa: Decoding-enchanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

William Held, Caleb Ziems, and Diyi Yang. 2023. TADA : Task agnostic dialect adapters for English. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 813–824, Toronto, Canada. Association for Computational Linguistics.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 85–94, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Agnes Jäger. 2018. *Vergleichskonstruktionen im Deutschen: Diachroner Wandel und synchrone Variation*, chapter 6. Vergleichskonstruktionen in den heutigen Dialekten des Deutschen. De Gruyter, Berlin, Boston.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Saméh Kchaou, Rahma Boujelbane, Emna Fsih, and Lamia Hadrich-Belguith. 2022. Standardisation of dialect comments in social networks in view of sentiment analysis : Case of Tunisian dialect. In *Proceedings of the Thirteenth Language Resources and*

Evaluation Conference, pages 5436–5443, Marseille, France. European Language Resources Association.

William D. Keel. 1980. Apocope and syncope in modern German dialects. Papers of the Mid-America Linguistics Conference, University of Kansas.

Werner König, Stephan Elspaß, and Robert Möller. 2015. *dtv-Atlas deutsche Sprache*, 18 edition. dtv. Graphics by Hans-Joachim Paul.

Bernd Kortmann, Kerstin Lunkenheimer, and Katharina Ehret, editors. 2020. *eWAVE*.

Louisa Lambrecht, Felix Schneider, and Alexander Waibel. 2022. Machine translation from Standard German to alemannic dialects. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 129–136, Marseille, France. European Language Resources Association.

Harrison Lee, Raghav Gupta, Abhinav Rastogi, Yuan Cao, Bin Zhang, and Yonghui Wu. 2022. SGD-X: A benchmark for robust generalization in schema-guided dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10938–10946.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021. Robustness testing of language understanding in task-oriented dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2467–2480, Online. Association for Computational Linguistics.

Björn Lundquist, Ida Larsson, Maud Westendorp, Eirik Tengesdal, and Anders Nøklestad. 2019. Nordic Word Order Database: Motivations, methods, material and infrastructure. *Nordic Atlas of Language Structures (NALS) Journal*, 4(1).

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater Reliability: the Kappa Statistic. *Biochemia medica*, 22(3):276–282.

Ludwig Merkle. 1993. *Bairische Grammatik*, 5th edition. Heinrich Hugendubel Verlag, Munich.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber. 2013. APiCS online. Max Planck Institute for Evolutionary Anthropology.

Alice Millour and Karën Fort. 2019. Unsupervised data augmentation for less-resourced languages with no standardized spelling. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 776–784, Varna, Bulgaria. INCOMA Ltd.

Ann-Marie Moser. 2023. The ups and downs of relative particles in German diachrony: On loss, grammaticalization, and standardization. *Journal of Historical Linguistics*, 13(3).

Pieter Muysken, Harald Hammarström, Olga Krasnoukhova, Neele Müller, Joshua Birchall, Simon van de Kerke, Loretta O'Connor, Swintha Danielsen, Rik van Gijn, and George Saad Saad. 2016. South American Indigenous Language Structures (SAILS) online. Max Planck Institute for the Science of Human History.

Pieter Muysken et al. 2000. *Bilingual speech: A Typology of Code-mixing*. Cambridge University Press.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Tillmann Nett, Angela Dorrough, Marc Jekel, and Andreas Glöckner. 2019. Perceived biological and social characteristics of a representative set of German first names. *Social Psychology*.

Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021. RADDLE: An evaluation benchmark and analysis platform for robust task-oriented dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*

(Volume 1: Long Papers), pages 4418–4429, Online. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, and Manfred Vogel. 2020. GermEval 2020 Task 4: Low-Resource Speech-to-Text. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob - a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108*.

Hannes Scheutz. 2005. Perspektiven einer neuen Dialekt-Syntax. volume 130 of *ZDL-Beihefte*. Franz Steiner Verlag.

Max Schmaltz. 2022. DERBI: DEutscher Regel-Basierter Inflektor. `https://github.com/maxschmaltz/DERBI`. Accessed: September 20, 2023.

Sailik Sengupta, Jason Krone, and Saab Mansour. 2021. On the robustness of intent classification and slot labeling in goal-oriented dialog systems to real-world noise. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 68–79, Online. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020a. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020b. Mind your inflections! Improving NLP for non-standard Englishes with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663, Online. Association for Computational Linguistics.

Trung Ngo Trong, Kristiina Jokinen, and Ville Hautamäki. 2019. Enabling Spoken Dialogue Systems for Low-Resourced Languages—end-to-end Dialect Recognition for North Sami. In *9th International Workshop on Spoken Dialogue System Technology*, pages 221–235. Springer.

Yi-Ting Tsai, Min-Chu Yang, and Han-Yu Chen. 2019. Adversarial attack on sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 233–240, Florence, Italy. Association for Computational Linguistics.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021a. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021b. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages

4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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*, pages 38–45, Online. Association for Computational Linguistics.

Zhengxuan Wu, Alex Tamkin, and Isabel Papadimitriou. 2023. Oolong: Investigating what makes transfer learning hard with controlled studies. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3280–3289, Singapore. Association for Computational Linguistics.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16, Ann Arbor, Michigan. Association for Computational Linguistics.

Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. OpenAttack: An open-source textual adversarial attack toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 363–371, Online. Association for Computational Linguistics.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. 2021. Sociolectal analysis of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A Examples of perturbed sentences

Perturbation	Sentence →	Perturbed sentence
Noun Phrase		
`possession_von`	Welcher Ort steht in der Erinnerung für das Abendessen des Schachclubs ? 'What's the location of the Chess Club dinner reminder?'	Welcher Ort steht in der Erinnerung für das Abendessen vom Schachclub ?
`possesion_pron`	Wann ist Robin Williams Geburtstag ? 'What is Robin William's birthday?'	Wann ist Robin William sein Geburtstag ?
`article_name`	Email an Natalie zu ihrem Geburtstag . 'Email Natalie for her birthday.'	Email an die Natalie zu ihrem Geburtstag .
`comparative`	Wird es morgen heißer als 25 Grad Celsius ? 'Will it be hotter than 25°C?'	Wird es morgen heißer wie 25 Grad Celsius ?
`double_det`	Ich möchte noch ein so lustiges Lied hören 'I want to hear another song this funny'	Ich möchte noch ein so ein lustiges Lied hören
Discourse & Word Order		
`name_order`	Ruf stattdessen Gloria Burgess an 'Call Gloria Burgess instead'	Ruf stattdessen Burgess Gloria an
`denn`	Wie lange geht meine *[sic]* Timer noch ? 'How much time is left on my timer?'	Wie lange geht denn meine Timer noch ?
`verb_clusters`	Zeige alle Erinnerungen an , die mit Familie zu tun haben . 'Show all family reminders'	Zeige alle Erinnerungen an , die mit Familie haben zu tun .
Tense & Aspect		
`progressive`	Ich höre Jazz . 'I listen to jazz.'	Ich bin Jazz am hören .
Adverbs & Prepositions		
`pronominal_adverbs`	Stelle dafür einen Timer . 'Set a timer for this.'	Stelle da einen Timer für .
`direction`	Berechne eine Route nach Hamburg . 'Calculate the route to Hamburg.'	Berechne eine Route auf Hamburg .
`location`	Was kostet der Bodentransport in Denver ? 'How much is ground transportation in Denver?'	Was kostet der Bodentransport zu Denver ?
Negation		
`negative_concord`	Nimm heute keine Anrufe an . 'Don't take any calls today.'	Nimm heute keine Anrufe nicht an .
Relativization		
`relative_pron`	Freunde , die jetzt online sind 'Friends who are online right now'	Freunde , wo jetzt online sind
Complementation		
`es_hat`	Sende Andre die neuesten IT Themen die es gibt	Sende Andre die neuesten IT Themen die es hat .
Verb Morphology		
`tun_imperativ`	Erinnere mich an notwendige veranstaltungen . 'Remind me of necessary events.'	Tu mich an notwendige veranstaltungen erinnern .
`schwa_elision`	Welche Erinnerungen habe ich für meinen Chef ? 'What reminders do I have for my boss?'	Welche Erinnerungen hab' ich für meinen Chef ?
Pronouns		
`clitic_es`	Wird es für die Party am Samstag sonnig ? 'Will it be sunny for the party on Saturday?'	Wird's für die Party am Samstag sonnig ?

Table 5: Examples of automatically perturbed sentences from the task-orieneted datasets used in this study.

B Performance in joint intent recognition and slot filling

B.1 Performance on perturbed English test sets

		Intact		Individual Perturbations		All Perturbations	
		Intent Acc	Slot F_1	Δ Intent Acc	Δ Slot F_1	Δ Intent Acc	Δ Slot F_1
xSID	mBERT	99.04	95.28	0.09	10.34	2.32	57.14
	XLM-R	99.20	95.93	0.11	9.94	1.96	57.66
	RemBERT	99.12	96.11	0.04	9.90	1.32	57.68
	mDeBERTa	99.04	96.00	0.07	9.96	1.68	58.01
	DistilmBERT	99.00	94.52	0.09	10.19	2.36	56.61
	mMiniLM	99.24	95.20	0.06	9.74	1.84	58.11
MultiATIS++	mBERT	79.69	93.00	0.00	10.98	0.04	45.46
	XLM-R	79.75	92.99	0.00	10.82	0.18	45.30
	RemBERT	79.73	93.31	0.00	11.04	0.18	45.59
	mDeBERTa	79.84	92.93	0.00	10.76	-0.07	46.31
	DistilmBERT	79.78	92.98	0.00	10.99	0.07	45.43
	mMiniLM	75.39	90.76	0.00	10.79	-0.04	44.24
MASSIVE	mBERT	87.95	81.92	0.14	14.08	4.93	44.50
	XLM-R	89.11	82.79	0.17	13.78	3.88	44.42
	RemBERT	89.25	83.10	0.08	14.21	3.57	44.82
	mDeBERTa	89.59	82.99	0.24	14.32	2.68	44.64
	DistilmBERT	87.11	80.65	0.27	14.17	5.20	43.61
	mMiniLM	84.77	79.91	0.15	13.65	4.21	43.14
MTOP	mBERT	96.40	89.14	0.18	4.54	4.83	50.58
	XLM-R	96.65	89.78	0.10	4.58	3.16	50.46
	RemBERT	97.15	89.83	0.13	4.52	4.17	50.44
	mDeBERTa	96.71	89.24	0.11	4.43	3.07	49.98
	DistilmBERT	96.01	88.53	0.19	4.56	4.52	50.19
	mMiniLM	93.17	88.84	0.18	4.53	3.45	50.47
Mean		90.53	89.82	0.10	9.87	2.48	49.37

Table 6: The overall results for intent recognition and slot filling on test sets in English and the gap in performance before and after UAAVE perturbations are applied. Intact (left): performance on intact test sets. Individual perturbations (middle): 118 individual perturbations are applied and average performance gap is computed across them. All perturbations (right): all perturbations applied simultaneously. Δ denotes the difference between performance on intact and perturbed data. Performance on intact data consistently surpasses that on perturbed data, leading to positive Δ values. The results are averaged across five runs with varying random initialization.

B.2 Performance with German development set on perturbed German test sets

		Intact		Individual Perturbations		All Perturbations	
		Intent Acc	Slot F_1	Δ Intent Acc	Δ Slot F_1	Δ Intent Acc	Δ Slot F_1
xSID	mBERT	78.72	71.81	0.43	2.47	6.12	21.28
	XLM-R	91.08	78.19	0.26	2.76	3.44	23.46
	RemBERT	94.88	83.12	0.39	2.78	5.20	23.76
	mDeBERTa	96.88	83.08	0.27	2.73	3.28	23.15
	DistilmBERT	75.88	66.33	0.58	2.23	6.24	19.78
	mMiniLM	72.32	70.51	0.31	2.39	2.84	22.64
MultiATIS++	mBERT	76.89	62.73	0.05	2.57	0.56	9.46
	XLM-R	79.08	78.49	0.02	3.88	0.27	11.52
	RemBERT	79.24	83.82	0.02	4.11	0.36	11.03
	mDeBERTa	78.84	80.12	-0.01	3.89	-0.02	10.90
	DistilmBERT	74.98	57.41	0.05	2.45	0.31	8.92
	mMiniLM	74.42	69.17	0.00	3.17	0.16	9.50
MASSIVE	mBERT	55.65	50.41	0.43	2.37	5.76	21.63
	XLM-R	75.10	65.55	0.42	2.80	6.75	26.42
	RemBERT	83.83	73.29	0.40	3.00	6.17	27.51
	mDeBERTa	84.05	73.83	0.40	3.04	6.43	28.03
	DistilmBERT	47.20	42.68	0.32	2.01	4.38	19.10
	mMiniLM	57.91	54.72	0.29	2.44	5.11	23.13
MTOP	mBERT	69.17	67.59	0.60	2.57	9.30	26.89
	XLM-R	88.40	77.84	0.59	2.95	8.70	30.57
	RemBERT	91.73	79.69	0.58	3.05	8.82	31.35
	mDeBERTa	91.24	79.78	0.59	2.97	8.19	30.70
	DistilmBERT	60.21	59.73	0.45	2.46	7.22	25.13
	mMiniLM	76.14	70.60	0.52	2.77	7.13	29.12
Mean		77.24	70.02	0.33	2.83	4.70	21.46

Table 7: The overall results for intent recognition and slot filling on test sets in German and the gap in performance before and after dialect perturbations are applied. Setup *(ii)*: English train set is used for training, German development set is used for model selection. Intact (left): performance on intact test sets. Individual perturbations (middle): 18 individual perturbations are applied and average performance gap is computed across them. All perturbations (right): all perturbations applied simultaneously. Δ denotes the difference between performance on intact and perturbed data. Performance on intact data consistently surpasses that on perturbed data, leading to positive Δ values. The results are averaged across five runs with varying random initialization.

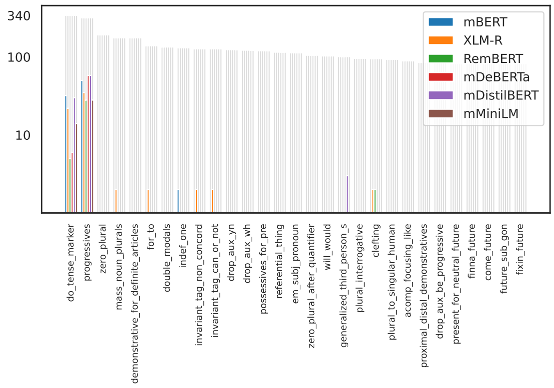
B.3 Performance on perturbed German test sets in fine-tuning setup

		Intact		Individual Perturbations		All Perturbations	
		Intent Acc	Slot F_1	Δ Intent Acc	Δ Slot F_1	Δ Intent Acc	Δ Slot F_1
MultiATIS++	mBERT	79.17	92.66	0.01	5.05	0.33	4.24
	XLM-R	79.17	92.56	0.01	4.97	0.33	3.09
	RemBERT	79.22	92.33	0.01	4.97	0.33	3.52
	mDeBERTa	79.46	92.55	0.01	4.99	0.33	3.08
	DistilmBERT	79.28	92.12	0.01	5.09	0.33	5.44
	mMiniLM	75.70	89.00	0.00	4.71	0.11	8.27
MASSIVE	mBERT	84.79	77.38	0.37	3.01	6.22	6.98
	XLM-R	86.73	78.83	0.34	3.06	5.78	6.95
	RemBERT	87.22	80.05	0.33	3.17	5.71	6.98
	mDeBERTa	87.19	79.62	0.35	3.12	5.95	6.96
	DistilmBERT	83.25	76.60	0.35	2.98	5.68	7.13
	mMiniLM	80.09	76.43	0.33	2.99	5.48	6.99
MTOP	mBERT	94.64	83.74	0.49	3.04	7.46	7.36
	XLM-R	95.71	84.37	0.48	3.07	7.43	6.95
	RemBERT	95.98	84.19	0.49	3.10	7.86	7.27
	mDeBERTa	95.62	84.55	0.48	3.07	7.86	6.96
	DistilmBERT	93.94	82.05	0.50	3.02	7.63	7.35
	mMiniLM	89.78	82.77	0.55	3.03	8.48	7.36
Mean		85.94	84.54	0.28	3.69	4.63	6.27

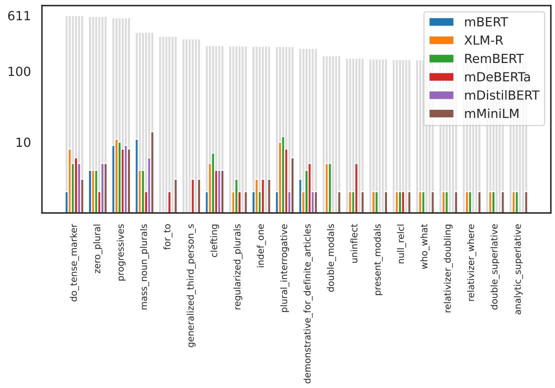
Table 8: The overall results for intent recognition and slot filling on test sets in German and the gap in performance before and after dialect perturbations are applied. Setup *(iii)*: German train set is used for training; German development set is used for model selection. Intact (left): performance on intact test sets. Individual perturbations (middle): 18 individual perturbations are applied and average performance gap is computed across them. All perturbations (right): all perturbations applied simultaneously. Δ denotes the difference between performance on intact and perturbed data. Performance on intact data consistently surpasses that on perturbed data, leading to positive Δ values. The results are averaged across five runs with varying random initialization.

C Success rate

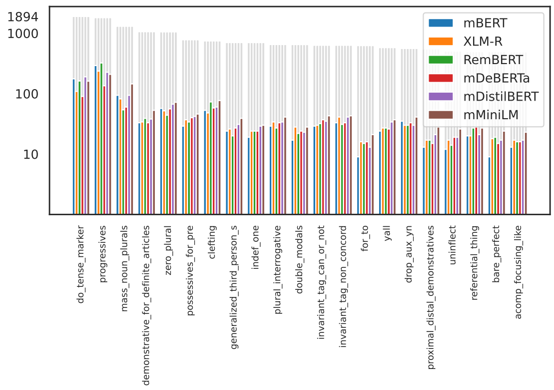
C.1 Success rate of English perturbations



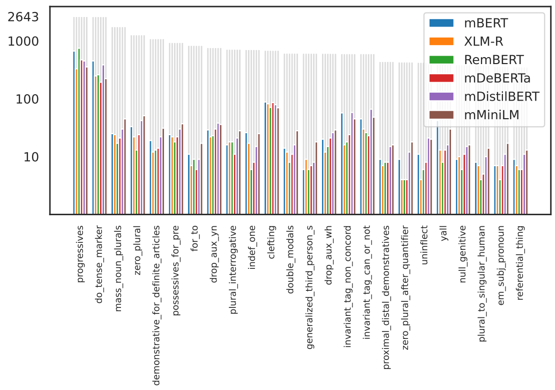
(a) xSID



(c) MultiATIS++



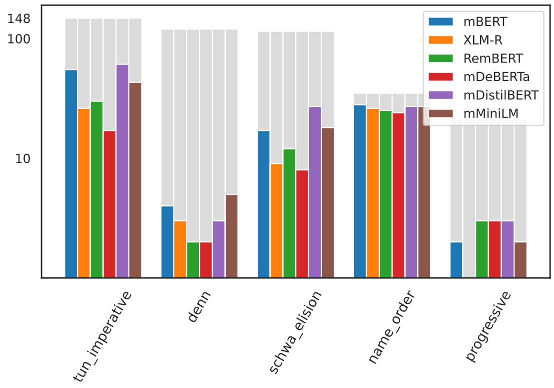
(b) MASSIVE



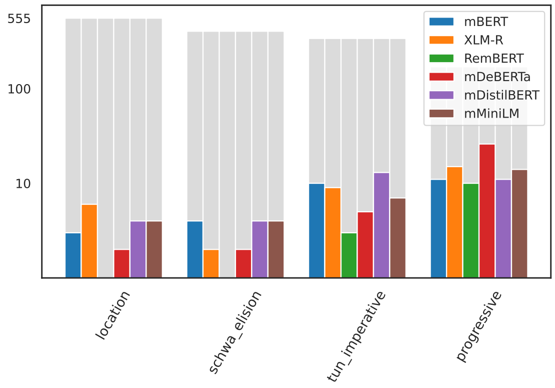
(d) MTOP

Figure 4: The success rates in intent prediction on the perturbed English tests sets with respect to individual perturbations. The grey bars represent the perturbation frequency (i.e., the count of altered sentences), while the colored bars indicate the success rate (i.e., the number of misclassified sentences after applying the perturbation). A logarithmic scale is utilized for improved clarity.

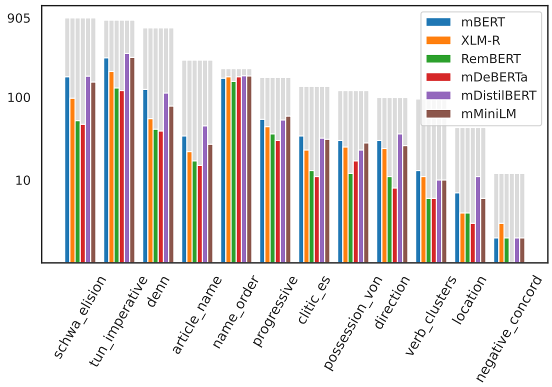
C.2 Success rate of German perturbations



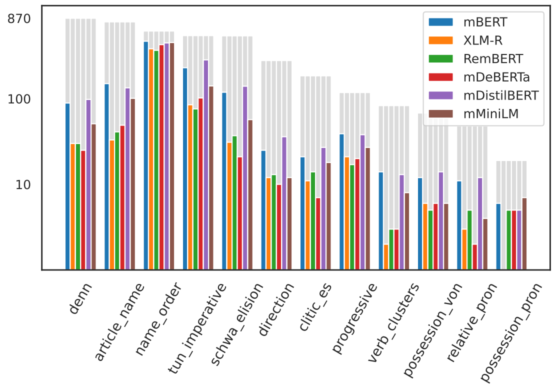
(a) xSID



(c) MultiATIS++



(b) MASSIVE

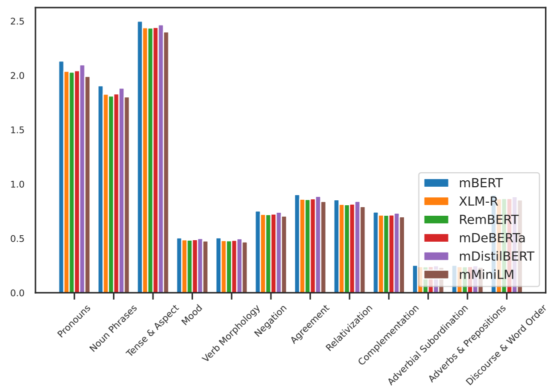


(d) MTOP

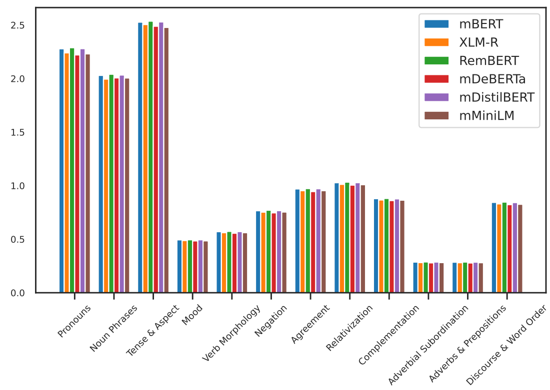
Figure 5: The success rates in intent prediction on the perturbed German tests sets with respect to individual perturbations. The grey bars represent the perturbation frequency (i.e., the count of altered sentences), while the colored bars indicate the success rate (i.e., the number of misclassified sentences after applying the perturbation). A logarithmic scale is utilized for improved clarity.

D Evaluation of performance drop

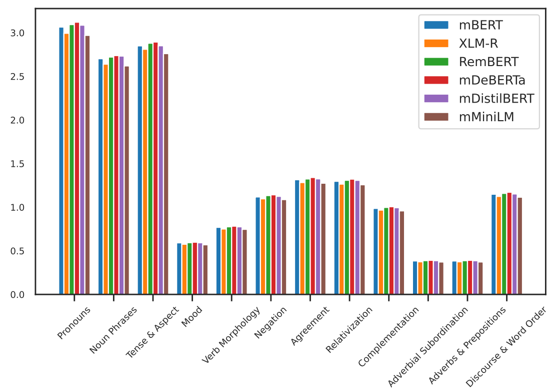
D.1 Performance drop in English



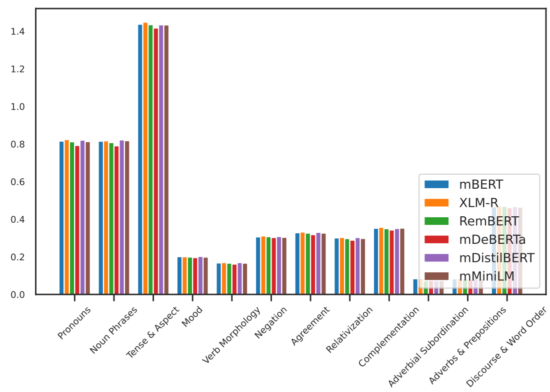
(a) xSID



(c) MultiATIS++



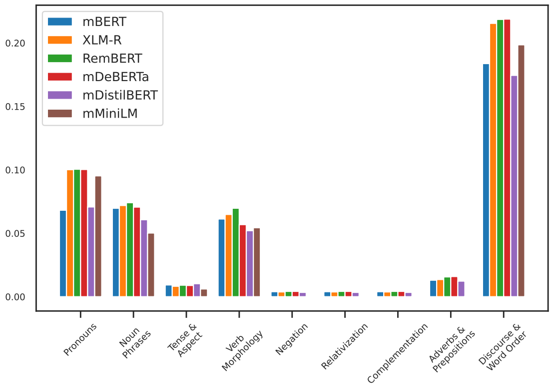
(b) MASSIVE



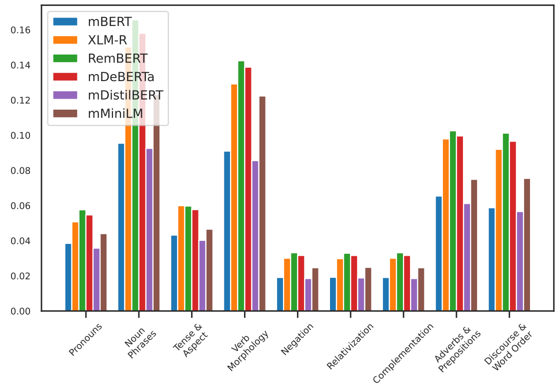
(d) MTOP

Figure 6: The ΔF_1 with respect to perturbation category in perturbed English test sets.

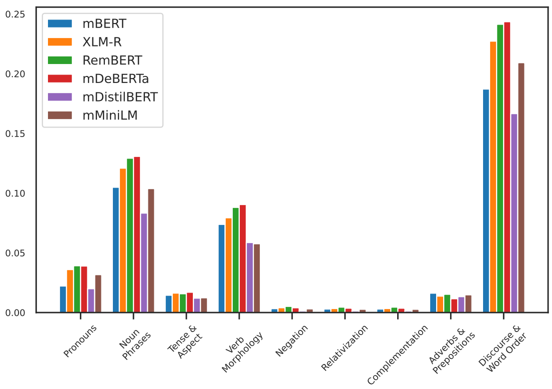
D.2 Performance drop in German



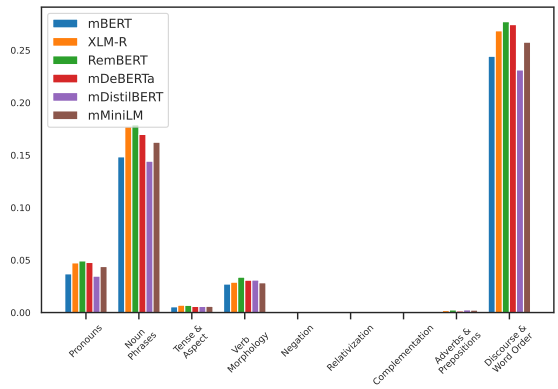
(a) xSID



(c) MultiATIS++



(b) MASSIVE



(d) MTOP

Figure 7: The ΔF_1 with respect to perturbation category in perturbed German test sets.

E Human evaluation guidelines

Sentence Pair Assessment

You'll be given a pair of sentences. One is in standard German, and the other is a re-write in dialect or colloquial German. A is for German sentences, B is for dialect re-writes.

Your job is to rate the naturalness and fluency of the re-write on a scale of one to five. Does the re-write sound like something you could say? A score of one indicates that the re-write sounds unnatural, while a score of five means that the re-write is fluent and completely acceptable. Trust your gut feeling and don't overthink it. If you're unsure about the score, choose the "idk" option (I don't know). Feel free to add comments if necessary.

Example

A Ich muss Papa jetzt anrufen .

B Ich muss den Papa jetzt anrufen .

1 - bad	2	3	4	5 - great
Comments (free form):				

The information from your evaluation will only be used for research.

Thank you for your time and effort!