DIALECTBENCH: A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages

Fahim Faisal[α,*] **Orevaoghene Ahia**[β,*] **Aarohi Srivastava**[γ] **Kabir Ahuja**[β]
David Chiang[γ] **Yulia Tsvetkov**[β] **Antonios Anastasopoulos**[α,δ]

[α]George Mason University [β]University of Washington [γ]University of Notre Dame
[δ]Archimedes Research Unit, RC Athena, Greece

{ffaisal,antonis}@gmu.edu {oahia,kahuja,yuliats}@cs.washington.edu
{asrivas2,dchiang}@nd.edu

Abstract

Language technologies should be judged on their usefulness in real-world use cases. An often overlooked aspect in natural language processing (NLP) research and evaluation is language variation in the form of non-standard dialects or language varieties (hereafter, varieties). Most NLP benchmarks are limited to standard language varieties. To fill this gap, we propose DIALECTBENCH, the first-ever large-scale benchmark for NLP on varieties, which aggregates an extensive set of task-varied variety datasets (10 text-level tasks covering 281 varieties). This allows for a comprehensive evaluation of NLP system performance on different language varieties. We provide substantial evidence of performance disparities between standard and non-standard language varieties, and we also identify language clusters with larger performance divergence across tasks. We believe DIALECTBENCH provides a comprehensive view of the current state of NLP for language varieties and one step towards advancing it further. [1]

1 Introduction

Benchmarking is important for tracking the progress the field of natural language processing (NLP) has made in various tasks. In the past few years, large-scale multilingual benchmarks like XTREME (Hu et al., 2020), XTREME-R (Ruder et al., 2021), and XGLUE (Liang et al., 2020) have played a pivotal role in evaluating the multilingual capabilities of NLP models. These efforts have sought to make model evaluation more accessible to researchers and representative of a variety of languages (Song et al., 2023). However, most benchmarks have focused on the standard varieties of languages, largely neglecting non-standard dialects and language varieties (Blasi et al., 2022).

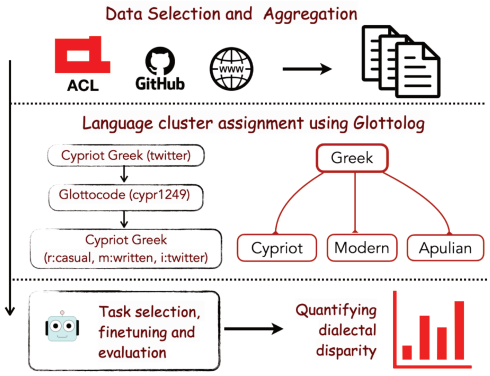


Figure 1: DIALECTBENCH Evaluation Suite.

We refer to non-standard dialects and language varieties simply as *varieties*, and sometimes include low-resource related languages, writing system variants, and other kinds of variation. Varieties contain subtle but notable variations in vocabulary, pronunciation, orthography and grammar, reflecting regional, social, and cultural differences (Chambers and Trudgill, 1998). The non-standard nature of these language varieties oftentimes contributes to the scarcity of substantial datasets that accurately capture these variations (Hedderich et al., 2021). As a result, they have often been absent from widely adopted benchmarks, even from admirable efforts like XTREME-up (Ruder et al., 2023), GLUECoS (Khanuja et al., 2020) and CreoleVal (Lent et al., 2023), which focus on under-resourced, code-switched, and creole languages, respectively. It is currently challenging to accurately test the robustness of multilingual models on a large suite of varieties without establishing an NLP evaluation framework covering multiple language clusters (each of which contains standard languages alongside its closely related varieties).

To this end, we create DIALECTBENCH, a large-scale benchmark covering 40 language clusters with 281 varieties, spanning 10 NLP tasks. We observe that the performance disparity between different varieties of the same language cluster becomes

[*]Equal contribution.
[1]More information can be found at the following:
Code/data: https://github.com/ffaisal93/DialectBench
Website: https://fahimfaisal.info/DialectBench.io

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14412–14454

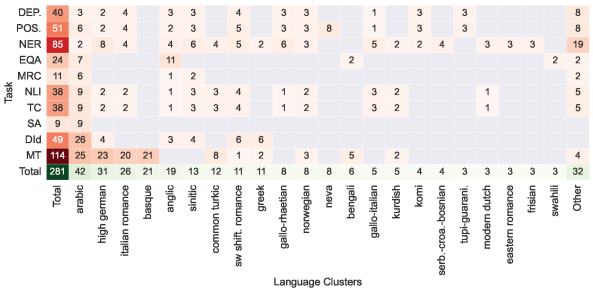


Figure 2: DIALECTBENCH language clusters with their variety counts per task. "Other" encompasses 18 clusters (full cluster list in Appendix Table 8).

more pronounced when we shift from zero-shot evaluation to fine-tuning on variety data, because of uneven data availability across varieties. Certain language clusters exhibit varying performance across downstream tasks within the same category, due to low-resource limitations. Additionally, we improve the dialectal task coverage for natural language inference by constructing a *translate-test* evaluation dataset. Putting these all together, DIALECTBENCH serves as a comprehensive suite that attains a multifaceted purpose: identifying broader limitations in dialectal NLP, while reflecting on potential areas for improvement.

2 DIALECTBENCH

DIALECTBENCH is a benchmark created to unify dialectal datasets across different tasks to foster research on language varieties and non-standard dialects. Below we describe the design choices we undertook to achieve this goal. This includes our language variety and task selection procedures, data collation methods, and evaluation principles.

Variety Selection We first looked through papers published in the ACL Anthology[2] from the last 10 years to find usable language resources, as well as commonly used online data repositories (Littauer, n.d.). We selected languages that have well-established, high-resourced varieties. Varieties may vary by location, ethnicity, or other factors. We also found instances where varieties are classified by writing system or even by genre (e.g., Twitter). When varying by location, varieties may be classified by different datasets at different levels of *granularity*, sometimes country, region, or city. In some cases, we found resources with two or more varieties within one dataset (e.g., the `UD_Portuguese-Bosque` dependency treebank (Rademaker et al., 2017) includes examples from both European and Brazilian Portuguese variants. To incorporate all these cases under one paradigm, we formulate a cluster-variety mapping procedure.

Cluster-Variety Mapping We construct several language clusters comprising of both high-resourced varieties and their low-resourced counterparts. We use the Glottolog language database (Nordhoff, 2012) to define clusters and assign varieties as outlined in Figure 1. This design choice enables us to keep varieties that are closely related in terms of either mutual intelligibility, phylogenetic similarity or geographic proximity within the same cluster. Hence, all cluster varieties always root back to the closest common linguistic ancestor and the whole cluster maps to an established phylogenetic subtree. For example, Fiji Hindi and Hindi, with Hindustani[3] as their closest common ancestor, are placed in the Hindustani cluster.

We primarily use the Glottocode language identification scheme (Hammarström and Forkel, 2022), ensuring a standardized naming scheme across all varieties. For instance, AAE variety from TwitterAAE (Blodgett et al., 2018) dependency parsing dataset is renamed as `African American Vernacular English` with a corresponding Glottocode `afri1276`. In cases where Glottocodes are unavailable, like for spoken English from South India, we substitute with the immediate ancestor Glottocode (`indi1255`) and further categorize the varieties using the following metadata identifiers:

1. Area (a): the region where the variety is spoken or where its dataset was collected.
2. Language register (r): frozen, formal, consultative, casual, and intimate.
3. Language mode (m): written, spoken, and signed language.
4. Orthography (o): In DIALECTBENCH this is only specific to Sinitic varieties. This could be either traditional or simplified.
5. Identifier (i): Dataset-specific metadata, could be domain (eg. twitter).

We encapsulate all this information, into a naming convention, and use the template: {Glottocode name}-(a:{},r:{},m:{},o:{},i:{}). [4]

[2] `https://aclanthology.org`

[3] `https://glottolog.org/resource/languoid/id/hind1270`

[4] For example, *mandarin chinese (a:mainland, o:simplified* refers to Mandarin Chinese (mand1415) spoken in Mainland China and written in simplified characters.

Category	Task	Metric	Source Dataset
Structured Prediction	DEP parsing	UAS	Universal Dependency (Zeman et al., 2021), TwitterAAE (Blodgett et al., 2018), Singlish (Wang et al., 2017)
	POS tagging	F1	Universal Dependency (Zeman et al., 2021), Singlish (Wang et al., 2017), Noisy Dialects (Blaschke et al., 2023)
	NER	F1	Wikiann (Pan et al., 2017; Rahimi et al., 2019), Norwegian NER (Johansen, 2019)
Classification	DId	F1	MADAR (Bouamor et al., 2018), DMT (Jauhiainen et al., 2019), Greek (Sababa and Stassopoulou, 2018), DSL-TL (Zampieri et al., 2023), Swiss Germans (Scherrer et al., 2019)
	SA	F1	TSAC (Medhaffar et al., 2017), TUNIZI (Fourati et al., 2021), DzSentiA (Abdelli et al., 2019), SaudiBank (Alqahtani et al., 2022), MAC (Garouani and Kharroubi, 2022), ASTD (Nabil et al., 2015), AJGT (Alomari et al., 2017), OCLAR (Al Omari et al., 2019)
	TC	F1	SIB-200 (Adelani et al., 2023)
	NLI	F1	XNLI (Conneau et al., 2018) translate-test
Question Answering	MRC	F1	Belebele (Bandarkar et al., 2023)
	EQA	Span F1	SDQA (Faisal et al., 2021)
Generation	MT	BLEU	CODET (Alam et al., 2023), TIL-MT (Mirzakhalov, 2021)

Table 1: The tasks and data sources of DIALECTBENCH (Detailed discussion: Appendix B).

Cluster Representative Each cluster will often have a high-resourced variety usually with the largest speaker population. We choose this high-resourced variety as the *cluster representative*. This selection might vary across downstream tasks depending on the data availability. We primarily utilize this representative variety to evaluate the performance gap across cluster varieties, and also rely on it for transfer-learning in resource-scarce settings. Sometimes, the members of a cluster are considered closely related languages, and sometimes dialects; to avoid making this distinction, we refer to all the members of a cluster simply as *varieties* of the cluster representative.

Task and Dataset Selection In selecting tasks, we maintain a balanced approach, promoting task diversity while also including tasks that require diverse levels of textual understanding. In the end, our complete list of tasks are as follows:

1. Dependency parsing (DEP parsing)
2. Parts-of-speech tagging (POS tagging)
3. Named entity recognition (NER)
4. Dialect identification (DId)
5. Sentiment analysis (SA)
6. Topic classification (TC)
7. Natural language inference (NLI)
8. Multiple-choice machine reading comprehension (MRC)
9. Extractive question answering (EQA)
10. Machine translation (MT)

In Table 1, we present the task and dataset details. We mostly keep the datasets in their originally published form (except for varieties renaming). For NLI, we use the existing English test set of XNLI (Conneau et al., 2018) and construct a multilingual dialect-focused translated evaluation dataset. We refer to this as translate-test NLI.

Evaluation Principles On the ground level, we evaluate existing NLP systems on text-based tasks using standard evaluation metrics (e.g., UAS for parsing, F1 for classification tasks, BLEU for translation). At a global level, we believe a sustainable NLP system should be user-focused while providing substantial (i) *linguistic utility* and (ii) *demographic utility* (Song et al., 2023; Blasi et al., 2022). Blasi et al. (2022) defined the utility of a task and language, as the corresponding performance normalized by the best possible performance (usually human-level performance). *Demographic utility* considers the demand for a language technology within a specific language, where the demand is proportional to the number of speakers of that language. *Linguistic utility*, on the other hand, asserts that "all languages are created equal" regardless of the number of speakers, and hence all languages in the world should receive identical weights.

Overall, we want to capture the performance gap between language clusters (e.g., Anglic vs. Italian Romance) as well as within language clusters (e.g., Norwegian Bokmål vs. Nynorsk). To attain this, we define the performance gap metrics in §3.3 . We also vary the experimental settings in §3.2, including zero-shot and few-shot cross-lingual transfer, as well as fine-tuning with similar high-resource languages. This is essential given that we lack clean annotated data in many varieties.

3 Experiments

Here, we report the entire process involved in creating and evaluating baselines for all of the tasks and varieties in DIALECTBENCH. Additionally, we define a *dialectal gap* metric to analyze performance disparities within and across clusters.

3.1 Models

We evaluate using two multilingual models: mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) for all tasks except MT. For MT, we do zero-shot evaluation with NLLB (NLLB

Team et al., 2022), using both the 600M and 1.3B variants. In addition, we use Mistral 7B (Jiang et al., 2023) to evaluate the current capability of LLMs on multilingual and dialectal understanding tasks. Our main goal is collating dialectal data across different languages and tasks under a single platform, hence we do not optimize for the best model performance. Rather, we focus on understanding and reporting the current state of performance on all DIALECTBENCH varieties.

3.2 Training and Evaluation

Training and evaluation procedures are largely determined by the availability of training or evaluation data for each task.

For any cluster C, let $\bar{C}$ be the highest-resourced variety (which is usually the cluster representative) of C. In addition, for any variety $v \in C$, we write $\bar{v}$ for the highest-resourced variety, that is, $\bar{v} = \bar{C}$. For any varieties t and v, let $\mathcal{S}_t(v)$ be the raw evaluation score of a system fine-tuned on t and tested on v (higher is better).

We use five general approaches for task-specific model training:

1. **In-variety fine-tuning** ($\mathcal{S}_v(v)$): In cases where there is available training data for a variety v, we fine-tune the base model on v. This primarily applies to tasks such as POS tagging and dependency parsing.
2. **In-cluster fine-tuning** ($\mathcal{S}_{\bar{v}}(v)$): In-variety fine-tuning is quite resource-intensive when we have a large number of varieties within a cluster C. In such cases, we fine-tune the base model on $\bar{C}$. Then we evaluate this model on each variety $v \in C$. This is the most common setting in our experiments, as it allows us to evaluate the dialectal performance gap without increasing the computation cost. For the DId task, we typically use a dataset of sentences annotated with variety labels to fine-tune one dialect-identification model for each language cluster.
3. **Combined fine-tuning** ($\mathcal{S}_{\mathcal{L}}(v)$): Unlike in the previous two methods, where each training set contains data from a single language cluster, we fine-tune our baseline question-answering (EQA and MRC) models using the SD-QA (Faisal et al., 2021) and Belebele (Bandarkar et al., 2023) datasets respectively, both of which contain training data in multiple standard varieties only and test data in other varieties. The SD-QA training data ($\mathcal{L}$ in the notation) contains questions in 9 standard varieties ($\mathcal{L} = \{\text{eng, ara, ben, fin, ind, swa, kor, rus, tel}\}$), while Belebele assembles data from 6 distinct multiple-choice QA datasets in standard English ($\mathcal{L} = \{\text{eng}\}$).
4. **Zero-shot evaluation** ($\mathcal{S}_{\text{eng}}(v)$): For certain varieties, obtaining training data even for in-cluster fine-tuning can be a challenge. Fortunately, English training data is always available for the datasets we study, so we use English to fill the gaps when we lack in-variety, in-cluster, or combined training data. At the same time, we aim to assess the feasibility of using this zero-shot cross-lingual transfer in reducing any existing performance gap across varieties. So we perform zero-shot cross-lingual transfer from English to each variety for 6 tasks in total. We only leave out those tasks such as dialect identification that explicitly require in-cluster training data.
5. **In-context learning** ($\mathcal{S}_{\text{icl}}(v)$): When evaluating large language models, we do not fine-tune them but instead rely on prompting and in-context learning. For this, we provide instructions and 5 examples in English as exemplars, followed by a prompt for predicting the test examples. Employing Mistral 7B (Jiang et al., 2023), we assess the present effectiveness of a close-to-state-of-the-art LLM on language varieties. The task-specific example prompts are reported in Appendix I.

Task	In-variety FT	In-cluster FT	Combined FT	Zero-shot	No reference	ref-In-context learning
DEP	✓			✓		
POS	✓			✓		
NER		✓		✓		
EQA			✓	✓		✓
MRC			✓			
NLI				✓		
TC		✓		✓		
SA		✓				✓
DId		✓				
MT				✓	✓	

Table 2: Task-specific training and evaluation procedure.

Table 2 summarizes the task-specific training procedures that we employ based on data availability. Note that, for MT, we perform zero-shot evaluation specifically in the translation direction, (*standard variety to English*) tested on (*dialectal variety to English*). But evaluation is a challenge because human-created reference translations into or out of non-standard varieties are usually very limited.

Therefore, we adopt an evaluation protocol from previous work (Alam et al., 2023) that uses pseudo-references. Given x, a sentence in a variety and $\bar{x}$, the translation of x into the standard variety, let y be the output of the MT system on input x and $\bar{y}$ be the output on input $\bar{x}$. Then we measure the quality of y (using, e.g., BLEU) compared against $\bar{y}$ as a pseudo-reference.

3.3 Quantifying the Dialectal Gap

To quantify the performance disparity across various resource-specific settings, language clusters and varieties, we introduce a *dialect performance gap* metric $\mathcal{G}_t(u, v)$: the relative decrease in performance of a system fine-tuned on variety t, tested on variety v compared to a baseline variety u:

$$\mathcal{G}_t(u, v) = \frac{\mathcal{S}_t(u) - \mathcal{S}_t(v)}{\mathcal{S}_t(u)}$$

with a special case for in-variety fine-tuning:

$$\mathcal{G}_{\text{in-variety}}(u, v) = \frac{\mathcal{S}_u(u) - \mathcal{S}_v(v)}{\mathcal{S}_u(u)}.$$

For the baseline score, we use either the score on the standard variety for each cluster ($u = \bar{v}$), or, in the zero-shot setting, the score on the language used for fine-tuning, namely English ($u = t =$ eng). Rather than computing an absolute gap, we opt for a relative gap (i.e., dividing by the baseline score). We also indicate whether the training setting is zero-shot ($t =$ eng) or fine-tuned on in-variety, in-cluster, or assembled data. Putting all these together, we compute the following three variations of dialectal inequality.

1. $\mathcal{G}_{\text{eng}}(\text{eng}, v)$: We calculate this metric to get a comprehensive measurement of global disparity across all varieties in a resource-limited environment (zero-shot transfer from English).
2. $\mathcal{G}_{\text{eng}}(\bar{v}, v)$: Using this variation, we keep the setting fixed as zero-shot and calculate the gap between the representative variety and any other variety.
3. $\mathcal{G}_t(\bar{v}, v)$: The two aforementioned metrics shed light on the extent of the variety performance gap in a resource-limited setting. To gain a more comprehensive perspective, we additionally compute another metric, this time utilizing the availability of resources. The computation approach remains as straightforward as before. We just use fine-tuning on a variety t instead of zero-shot transfer from Standard English: $t =$ in-variety for in-variety fine-tuning, $t = \bar{v}$ for in-cluster fine-tuning, or t is some set of varieties for combined fine-tuning.

For all three $\mathcal{G}$ metrics, we compute them at the variety level and then average them at the cluster level:

$$\mathcal{G}_t(u, C) = \frac{1}{|C|} \sum_{v \in C} \mathcal{G}_t(u, v)$$

$$\mathcal{G}_{\text{in-variety}}(u, C) = \frac{1}{|C|} \sum_{v \in C} \mathcal{G}_{\text{in-variety}}(u, v).$$

4 Results

We, first of all, discuss results by highlighting the highest possible score per variety, aka the *maximum obtainable evaluation scores* regardless of evaluation method or training data. Next, we extend our discussion further by reporting the existing dialectal disparity across clusters and varieties.

4.1 Maximum Obtainable Scores

Here we provide key findings from our evaluation on each task. A task-specific summary is reported in Table 3. Detailed results comprising all tasks, models, language clusters and varieties are reported in Tables 10 to 20 in Appendix E.

Structured prediction We present visualizations for the task-specific maximum scores. We show the one for Dependency Parsing in Fig. 3, where we observe that low-resource varieties from language clusters such as Tupi-Guarani (indigenous South American cluster), Saami and Komi (low-resource Uralic language clusters) have the lowest performance compared to Standard English and other closely related Germanic and Romance clusters. These low-resource varieties are also not included in the pretraining stage of our base language models (eg. mBERT). Furthermore, this trend is evident across all three structured prediction tasks. On the other hand, high-resource Indo-European languages such as Portuguese, French, and Norwegian usually perform better.

Sequence classification For DId and SA, we generally collate different datasets for each language cluster and therefore, report the comparative classification results together. As a result, the locality level (e.g. city/region/country) also varies across clusters. For example, we report city-level DId results for Arabic and High German but country-level

Category	Task	num. cl.	num. var.	avg. score	Max-score cluster/variety		Min-score cluster/variety	
Structured prediction	DEP parsing	16	40	64.3	sw. shifted romance/brazilian portuguese	94.4	tupi-guarani sg./mbyá guaraní (a:brazil)	9.0
	POS tagging	17	51	72.1	norwegian/norwegian bokmål (m:written)	98.7	tupi-guarani sg./mbyá guaraní (a:brazil)	1.9
	NER	27	85	70.1	eastern romance/romanian	94.2	anglic/jamaican creole english	0.0
Sequence classification	NLI	15	38	64.2	anglic/english	83.4	sotho-tswana (s.30)/southern sotho	34.6
	TC	15	38	77.7	sinitic/cmm. sinitic (o:traditional)	89.8	kurdish/central kurdish	19.4
	DId	6	49	67.0	sinitic/mandarin chinese (a:taiwan, o:simp.)	98.6	sw. shifted romance/portuguese (m:written)	17.4
	SA	1	9	80.3	arabic/tunisian arabic	94.6	arabic/south levantine arabic	58.9
Question Answering	MRC	4	11	40.9	anglic/english	53.4	sotho-tswana (s.30)/southern sotho	29.0
	EQA	5	24	74.2	arabic/arabic (a:saudi-arabia)	77.9	swahili/swahili (a:tanzania)	63.5
Generation	MT-dialect	12	73	25.2	arabic/gulf arabic (a:riy)	43.1	common turkic/sakha	2.5
	MT-region	2	41	33.0	high german/central alemannic (a:ur)	44.1	italian romance/italian (a:sardegna)	13.0

Table 3: Task specific result summary using Maximum Obtainable Score. The varieties with the minimum scores exhibit a noticeable lag in performance across various tasks when compared to the average task performance.

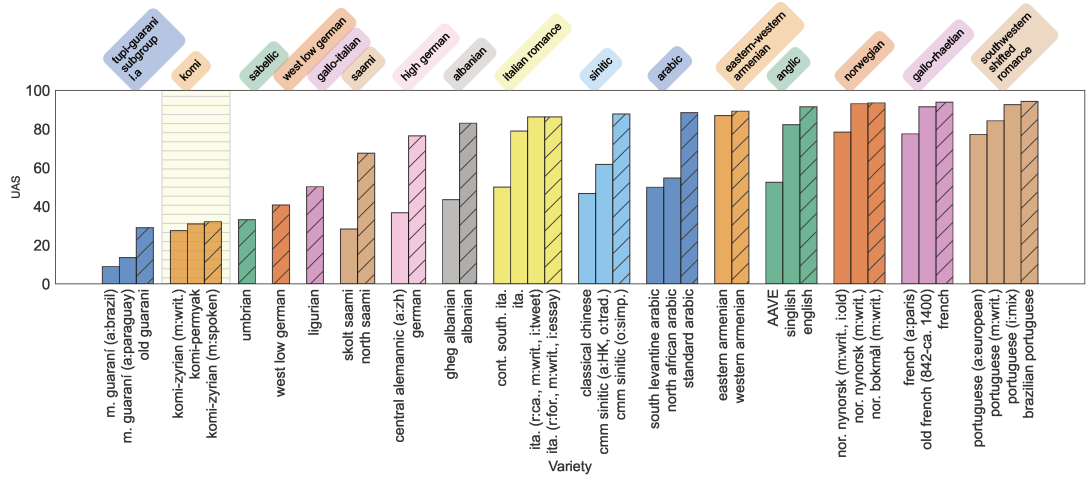


Figure 3: Maximum scores (*max. UAS*) in Dep. Parsing task. *Yellow-shaded region:* Komi is the only cluster having no varieties seen during mBERT pertaining. *Colored bars with diagonal stripes:* the cluster representative variety. Low-resourced cluster varieties score lower compared to high-resource Germanic clusters.

results for Portuguese, Spanish and English. In the case of SA, we have region/country-level results for Arabic varieties. For TC and NLI, we have the same set of clusters and varieties. However, we only report the zero-shot transfer performance from Standard English for NLI using the newly created translate-test NLI dataset.

For TC and NLI, we observe the largest in-cluster disparity in the Kurdish cluster, with Northern Kurdish outperforming all others. The Sotho varieties consistently perform significantly lower compared to other clusters. For all three sequence classification tasks, we generally find the Chinese cluster performing on par with high-resource Latin counterparts.

Question answering We generally do not see large gaps in performance within varieties in each language cluster. In EQA zero-shot experiments, English and its varieties have the highest performance overall and Korean varieties score the lowest. Combined fine-tuning boosts performance on all language clusters except in English. It's important to note that these EQA scores primarily indicate the model's robustness to accent-level differences and transcription noise, rather than broad dialectal robustness. However, further investigation is needed to determine whether this robustness specifically applies to both accent-level differences and transcription noise, or to any character-level variation up to a certain threshold.

For the MRC task, the performance peaked at 53.4 for Standard English, while the lowest score was 29 for Southern Sotho. More detailed results are presented in Tables 15 and 19.

Machine translation The performance gap here varies widely across and within language varieties. Performance is similar within the Swiss-German cluster, with higher performance (see Figure 5) across regions in Northern Switzerland, which is geographically closer to Germany. The performance gap for Norwegian dialects (Figure 10b) is surprising as we perform zero-shot transfer from

Task	Gap metric	Avg. val	cluster (max)	cluster (min)
DEP parsing	$\mathcal{G}_{\text{eng}}(\text{eng}, C)$	34.0	arabic, 50.8	sw. shift. romance, 19.4
	$\mathcal{G}_{\text{eng}}(\bar{v}, C)$	15.7	anglic, 34.2	sw. shift. romance, -0.9
	$\mathcal{G}_{\text{in-variety}}(\bar{v}, C)$	26.4	arabic, 93.8	italian romance, 0.6
POS tagging	$\mathcal{G}_{\text{eng}}(\text{eng}, C)$	27.4	arabic, 58.5	norwegian, 14.7
	$\mathcal{G}_{\text{eng}}(\bar{v}, C)$	6.7	anglic, 20.9	ew. armenian, -2.1
	$\mathcal{G}_{\text{in-variety}}(\bar{v}, C)$	6.2	arabic, 29.1	neva, -0.5
NER	$\mathcal{G}_{\text{eng}}(\text{eng}, C)$	31.7	kurdish, 77.1	modern dutch, 12.3
	$\mathcal{G}_{\text{eng}}(\bar{v}, C)$	22.3	kurdish, 78.2	hindustani, -23.6
	$\mathcal{G}_{\bar{v}}(\bar{v}, C)$	-28.6	kurdish, 91.5	sorbian, -1162.7
TC	$\mathcal{G}_{\text{eng}}(\text{eng}, C)$	22.4	kurdish, 74.2	sinitic, 0.4
	$\mathcal{G}_{\text{eng}}(\bar{v}, C)$	12.5	kurdish, 60.6	norwegian, -2.2
	$\mathcal{G}_{\bar{v}}(\bar{v}, C)$	-1.9	latvian, 21.0	kurdish, -61.3

Table 4: Comparative cluster-level dialectal gap across tasks. In general, the average disparity is larger for zero-shot transfer $\mathcal{G}_{\text{eng}}(\text{eng}, C)$. However, when we move from zeroshot to finetune (i.e. $\mathcal{G}_{\text{eng}} \rightarrow \mathcal{G}_{\bar{v}/\text{in-variety}}$) and compute the distance from a cluster representative $\bar{v}$, we observe increased dialectal disparity $|\mathcal{G}(\bar{v}, C)|$.

Norwegian Nynorsk (a Western dialect) but obtain better performance on the Eastern dialect. Within Arabic, Riyadh is the highest performer while Sfax performs the worst. For the Bengali cluster, Jessore has the highest performance –this is not surprising since it is one of the dialects from which standard Bengali originated (Alam et al., 2023). The Ethiopian variety of Tigrinya exhibits a higher performance than the Eritrean one, even though Tigrinya is more commonly spoken in Eritrea [5]. Amongst the clusters within the Basque cluster, Barkoxe and Maule have the lowest score while Azkaine scores the highest.

4.2 Dialectal Gap Across Language Clusters

In Fig. 4, we plot the zero-shot dialectal gap for three tasks. In the x-axis, we report the aggregated cluster-level gap $\mathcal{G}_{\text{eng}}(\text{eng}, v)$, compared against the fine-tuning variety (standard English) while in y-axis we report $\mathcal{G}_{\text{eng}}(\bar{v}, v)$, the gap compared against the representative variety of a cluster. In an ideal scenario, we would want both of these gap values to be close to zero. However, this is certainly not the case. The general observed trend is that the low-resource clusters have higher gaps of both $\mathcal{G}_{\text{eng}}(\text{eng}, C)$ and $\mathcal{G}_{\text{eng}}(\bar{v}, C)$, whereas high-resource Germanic and Sinitic language clusters consistently exhibit low dialectal gaps. That said, certain specific high-resource varieties, such as Standard German and its dialectal counterparts like Swiss German, showcase significant within-cluster dialectal gaps (Fig. 4a).

We primarily report dialectal gaps using zero-shot transfer because the finetuning data available across task and cluster is very disproportionate. Often the in-cluster/variety data is not good enough in terms of data quality and quantity. For example, we have 37 varieties in 13 clusters for dependency parsing but out of these, only 20 varieties have data available for in-variety fine-tuning. This lacking becomes more apparent when we compare the statistics of two types of within-cluster dialectal gaps: zero-shot $\mathcal{G}_{\text{eng}}(\bar{v}, C)$ against fine-tuning $\mathcal{G}_{\bar{v}/\text{in-variety}}(\bar{v}, C)$ in Table 4. In general, the within-cluster dialectal disparity is smaller for zero-shot transfer (i.e. $\mathcal{G}_{\text{eng}}(\bar{v}, C) \leq |\mathcal{G}_{\bar{v}/\text{in-variety}}(\bar{v}, C)|$). Here, the in-cluster/variety fine-tuning results in a higher performance deviation primarily due to the inconsistent variety-specific finetuning data quality.

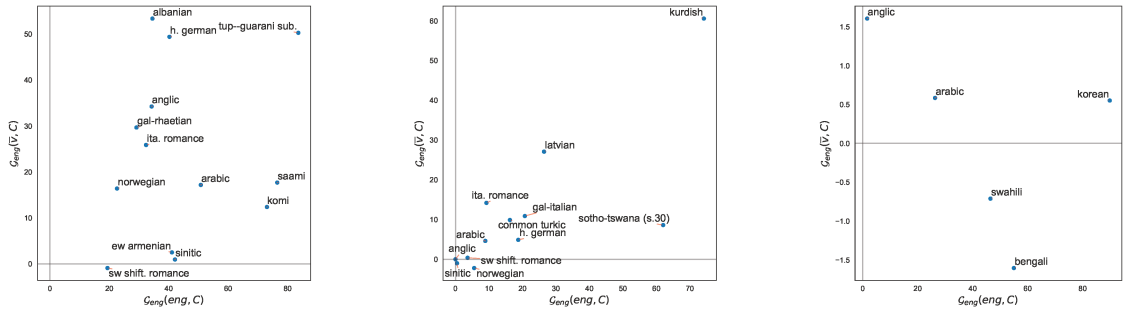
Task	zero-shot mBERT	zero-shot XLM-R	few-shot / FT mBERT	few-shot / FT XLM-R
DEP. Parsing	61.6	61.3	76.2	64.3
POS Tagging	69.5	69.7	89.8	89.1
NER	59.7	57.8	65.8	61.4
NLI	56.9	62.5	—	—
TC	72.3	71.4	73.1	68.9
MRC	—	—	39.4	40.3
EQA	53.9	51.9	69.2	67.2
SA	—	—	78.8	80.1
DId	—	—	65.8	59.3
win	4/6	2/6	6/8	2/8

Table 5: Base model comparison. We found mBERT was easier to fine-tune using the default hyperparameter setting thus, resulting in a higher winning rate.

5 Discussion

High resource vs. low resource varieties The highest-performing varieties are mostly standard high-resource languages and a few high-resource dialects (Norwegian dialects) whereas, the majority of the lowest-performing language variants are low-resourced varieties. This clear distinction of language varieties points towards the large existence of in-cluster dialectal gaps. Furthermore, this finding correlates with language script differences. We observe that 77.2% of top-10 varieties in terms of maximum obtainable score are written with Latin script. Another finding is the performance instability of low-resource varieties across tasks. For instance, Old Guarani performs better in DEP parsing whereas, Mbyá Guarani (Paraguay) surpasses it in POS tagging even though the dataset remains the same (i.e. UD). For more detailed comparisons, in Appendix Table 21, we report the top-10 highest and lowest-scoring varieties across different tasks.

[5] https://en.wikipedia.org/wiki/Tigrinya_language



(a) Dependency parsing (b) Topic classification (c) Extractive question answering

Figure 4: Dialectal gap visualization for language clusters utilizing zero-shot cross-lingual transfer from Standard English. In the x-axis, values far from zero have a larger performance gap from English whereas, in the y-axis, values far from zero have a larger within cluster gap. Ideally, we want both of them to be close to zero.

Model hyperparameter tuning In Table 5, we compare the baseline multilingual models. mBERT was comparatively easier to train than XLM-R using the default learning rates reported in the earlier task experiments. Often for tasks such as MRC, we needed to tune hyperparameters (e.g. learning rate, max. sequence length) in case of XLM-R. However, once we identify a hyperparameter configuration that converges in a zero-shot setting, we use the same to train all the language-variant training for that specific task. As a result, for some low-resource languages, XLM-R does not converge compared to mBERT. For example, XLM-R dependency parsing UAS score for Norwegian-NynorskLIA is 56.08 (zeroshot) and 8.25 (in-variety FT) whereas, we get 78.39 for in-variety mBERT fine-tuning. We suspect this hyperparameter tuning issue is one of the contributing factors toward a lower winning rate of XLM-R in few-shot / fine-tuning settings. However, this could be improved further with an extensive parameter grid search and settings specifically tailored for each language cluster.

Positive zero-shot transfer for Latin varieties Low-resource varieties written in Latin script receive greater benefit in zero-shot because of effective transfer from high-resource Standard English. With the presence of In-cluster/variety finetuning data, we effectively diminish this script effect to some extent. For example in the Hindi cluster, Fiji performs better than its Latin non-standard counterparts with in-cluster finetuning for NER (Table 12). In summary, if the standard variety of a cluster is non-latin but high-resource, then the success rate of in-cluster/variety fine-tuning tends to be higher. However, where all dialects are low-resource, Latin script varieties utilizing zero-shot transfer, eventually surpass others in the performance hierarchy. As an example, we report the zero-shot NER instances where the low-resource varieties perform better than the representative ones in Appendix Table 22 (most of these use Latin script).

LLM evaluation via In-context learning For SA and EQA tasks, we have in-context learning results using the Mistral7B LLM. Comparing the performance against zero-shot and fine-tuning using our encoder-based models, we find dialect performance of LLM is better than zero-shot transfer but falls behind the finetuned results. On top of that, data contamination (Ahuja et al., 2023) during LLM evaluation is another existing issue while considering these few available dialectal resources. Creating translation-based comparable data might be a solution to perform a fair benchmarking of LLMs on low-resource varieties.

Misinterpreting evaluation metrics We also report the cluster-level population-weighted average (i.e. *demographic utility*) which rewards a system more when it provides increased *linguistic utility* (eg. raw F1 score) for varieties, spoken by a larger population compared to varieties spoken by a smaller population. Alone, this metric could be misleading if we consider the fact that the performance gap among all varieties from a particular cluster should be minimal. On the other hand, solely looking into the *linguistic utility* average does not give a clear picture either (e.g. often overshadows the larger performance deviation of certain varieties having extreme scores). So for all clusters and tasks, we report the *linguistic utility* average as well as the *demographic utility* average, the minimum score of a cluster, and the standard deviation in Tables 23 to 30.

6 Conclusion

We propose DIALECTBENCH, the first-ever inclusive Dialectal NLP benchmark reporting performance evaluation and disparity across standard and non-standard varieties. This is one step towards the effort of bringing more and more language under the paradigm of NLP technology. We would like to further improve the benchmark, constructing high-quality comparable data and expanding the task coverage to speech-based NLP technologies.

Limitations

The data quality and quantity, variety coverage vary significantly across tasks because of data scarcity issues. We avoid full-scale LLM-evaluation consciously because of the uncertain data-contamination issue and their well-known lower performance threshold compared to smaller masked-language-modeling-based fine-tuned models. In addition, we focus on text-based NLP tasks for this current iteration. Moreover, we do not claim the representative varieties of each language clusters to be any kind of superior or standardized forms over the other varieties. These varieties are chosen to perform a well-informed comparison among the perceived well-resourced linguistic variety and its counterparts having lesser data availability. At the same time, the mutual intelligibility and phylogenetic similarity of the similar cluster varieties also vary across cluster and this was not selected in a numerically quantifiable manner.

Evaluation Limitations To further improve the evaluation fairness of the current version of DIALECTBENCH, we need (i) Parallel corpus utilization to prepare task-specific data (ii) Translation-based task data generation to perform comparable analysis (iii) Quantifying the resource-supply and demand as well as population-coverage (Song et al., 2023) to identify where a variety stands in the global landscape of *linguistic utility*. Here, we have accumulated data for diverse varieties across tasks that vary significantly in terms of quality, example count, and domain. However, to perform a perfectly fair comparison of dialectal inequality, we should consider high-quality comparable data (e.g. parallel corpus, similar varieties across tasks) which is not available at this point.

Continuity of DIALECTBENCH Despite our best effort, this benchmark does not include every one of the already published task-specific dialectal datasets. So, our next steps on this project involve hosting the benchmark on the website that displays the current statistics of the datasets in DIALECTBENCH. We will also encourage researchers to add new and existing datasets for tasks, language clusters that might be currently missing alongside the respective baselines.

Space Limitations Our study encompasses a large set of evaluation result tables, their corresponding visualizations and findings analysis. Due to space limitations, we have to move the detailed reports (Appendix E) and the rest of the visualizations (Appendix D) in the Appendix. To better assist, we include an Appendix Table of Content (Table 6) at the introductory section of Appendix (Section 6).

Ethics Statement

This work is a compilation of existent dialectal datasets across different tasks, including structured prediction and generative tasks. Our experiments do not particularly optimize for the best model performance of these tasks. Therefore we acknowledge that for some of the tasks, the baseline models might not be robust enough to handle dialectal text hence resulting in wrong predictions and generations. We believe that this underscores the need for building models robust to different language variations and future work should focus on this.

Acknowledgements

This material is based upon work supported by the US National Science Foundation under Grants No. IIS-2125466, IIS-2125948, CAREER Grant No. IIS2142739, as well as NSF Grants No. IIS-2203097 and IIS-2125201. We gratefully acknowledge support from Alfred P. Sloan Foundation Fellowship. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Adel Abdelli, Fayçal Guerrouf, Okba Tibermacine, and Belkacem Abdelli. 2019. Sentiment analysis of Arabic Algerian dialect using a supervised method. In *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*, pages 1–6.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. 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 7088–7105, Online. Association for Computational Linguistics.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Megaverse: Benchmarking large language models across languages, modalities, models and tasks.

Marwan Al Omari, Moustafa Al-Hajj, Nacereddine Hammami, and Amani Sabra. 2019. Sentiment classifier: Logistic regression for Arabic services' reviews in Lebanon. In *2019 International Conference on Computer and Information Sciences (ICCIS)*, pages 1–5.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2023. CODET: A benchmark for contrastive dialectal evaluation of machine translation.

Khaled Mohammad Alomari, Hatem M. ElSherif, and Khaled Shaalan. 2017. Arabic tweets sentimental analysis using machine learning. In *Advances in Artificial Intelligence: From Theory to Practice*, pages 602–610, Cham. Springer International Publishing.

Dhuha Alqahtani, Lama Alzahrani, Maram Bahareth, Nora Alshameri, Hend Al-Khalifa, and Luluh Aldhubayi. 2022. Customer sentiments toward Saudi banks during the Covid-19 pandemic. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 251–257, Trento, Italy. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages. In *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for 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.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jack K. Chambers and Peter Trudgill. 1998. *Dialectology*. Cambridge University Press.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chayma Fourati, Hatem Haddad, Abir Messaoudi, Moez BenHajhmida, Aymen Ben Elhaj Mabrouk, and Malek Naski. 2021. Introducing a large Tunisian Arabizi dialectal dataset for sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 226–230, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Moncef Garouani and Jamal Kharroubi. 2022. MAC: An open and free Moroccan Arabic corpus for sentiment analysis. In *Innovations in Smart Cities Applications Volume 5*, pages 849–858, Cham. Springer International Publishing.

Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Finnish dialect identification: The effect of audio and text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8777–8783, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Harald Hammarström and Robert Forkel. 2022. Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web Journal*, 13(6):917–924.

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.

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.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. Italian language and dialect identification and regional French variety detection using adaptive naive Bayes. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 119–129, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019. Discriminating between Mandarin Chinese and Swiss-German varieties using adaptive language models. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–187, Ann Arbor, Michigan. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825.

Bjarte Johansen. 2019. Named-entity recognition for Norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics, NoDaLiDa*.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Hans Erik Heje, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Søgaard, and Johannes Bjerva. 2023. CreoleVal: Multilingual multitask benchmarks for creoles.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Richard Littauer. n.d. Low resource languages: Resources for conservation, development, and documentation of low resource (human) languages. https://github.com/RichardLitt/low-resource-languages. [Accessed 15-11-2023].

Salima Medhaffar, Fethi Bougares, Yannick Estève, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 55–61, Valencia, Spain. Association for Computational Linguistics.

Jamshidbek Mirzakhalov. 2021. *Turkic Interlingua: A Case Study of Machine Translation in Low-resource Languages*. Ph.D. thesis, University of South Florida.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021a. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr, et al. 2021b. A large-scale study of machine translation in turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. ASTD: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling human-centered machine translation. arXiv:2207.04672.

Sebastian Nordhoff. 2012. Linked data for linguistic diversity research: Glottolog/langdoc and asjp. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 191–200. Springer, Heidelberg.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Lucy Park, Alice Oh, Jung-Woo Ha (NAVER AI Lab), Kyunghyun Cho, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Anthony Rios. 2020. FuzzE: Fuzzy fairness evaluation of offensive language classifiers on African-American English. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):881–889.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging

and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hanna Sababa and Athena Stassopoulou. 2018. A classifier to distinguish between Cypriot Greek and Standard Modern Greek. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 251–255.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Yueqi Song, Catherine Cui, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos, and Graham Neubig. 2023. GlobalBench: A benchmark for global progress in natural language processing.

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal Dependencies parsing for colloquial Singaporean English. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1732–1744, Vancouver, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language variety identification with true labels.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Ajede, and et al. 2021. Universal Dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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.

Appendix

In this supplementary material, we provide the following: (i) Relevant literature review (ii) Overall results and dataset details that we could not fit into the main body of the paper.

Section	Desctiption
Appendix A	Related Work
Appendix B	Tasks of DIALECTBENCH • Translate-Test NLI Dataset Statistics (Table 7)
Appendix C	Varieties and Clusters of DIALECTBENCH • DIALECTBENCH Variety list (Table 8) • Language Clusters and Representative Varieties (Table 9)
Appendix D	Result Visualizations Regional maps with aggregated Machine Translation scores (Figs. 5 to 6) • Map of Switzerland with aggregated BLEU scores of Swiss-German variety per region (Fig. 5) • Map of Italy with aggregated BLEU scores of Italian variety per region (Fig. 6) Task Specific Plot for Maximum Scores (Figs. 7 to 10) • Parts-of-Speech Tagging (Fig. 7a) • Named entity recognition (Fig. 7b) • Topic classification (Fig. 7c) • Natural language inference (Fig. 8a) • Extractive question answering (Fig. 8b) • Multiple-choice machine reading comprehension (Fig. 8c) • Sentiment analysis (Fig. 9a) • Dialect identification (Fig. 9b) • Machine translation (MT-region) (Fig. 10a) • Machine translation (MT-dialect) (Fig. 10b) Dialectal Gap visualization utilizing zero-shot cross-lingual transfer from Standard English. • Parts-of-Speech Tagging (Fig. 11a) • Named entity recognition (Fig. 11b) • Dialect identification (Fig. 11c)
Appendix E	Task Specific Evaluation Result Tables • Dependency parsing (Table 10) • Parts-of-Speech tagging (Table 11) • Named entity recognition (Table 12) • Natural language inference (Table 13) • Extractive question answering (Tables 14 to 15) • Dialect identification (Table 16) • Topic classification (Table 17) • Sentiment analysis (Table 18) • Multiple-choice machine reading comprehension (Table 19) • Machine translation (Table 20)
Appendix F	Highest performing and lowest performing varieties (Table 21)
Appendix G	Low-resource Variety performing better in zero-shot NER (Table 22)
Appendix H	Cluster level Result Summaries with Demographic Utility and Standard Deviation report (Tables 23 to 31)
Appendix I	In-Context Learning Details

Table 6: Table of contents for supplementary material.

A Related Work

The majority of multilingual benchmarks (Hu et al., 2020; Ruder et al., 2023, 2021; Liang et al., 2020; Wilie et al., 2020; Park et al., 2021) have heavily focused on dominant language varieties. However, the performance disparity between standard languages and their dialectal counterparts has been studied (Kantharuban et al., 2023; Ziems et al., 2022; Rios, 2020; Ziems et al., 2023) and shown to be significantly large across different tasks. Still, the large-scale generalization of this finding comprising numerous task

categories is yet to be done. There have been previous efforts to bridge this gap though. Some work has focused on curating dialectal datasets across several tasks within one language cluster, while others have focused on single tasks across many language clusters. For instance, ORCA (Elmadany et al., 2023), ARLUE (Abdul-Mageed et al., 2021) and ALUE (Seelawi et al., 2021) are dedicated natural language understanding benchmarks focusing on Arabic varieties alone. Multi-VALUE (Ziems et al., 2023) was developed for benchmarking NLP tasks in English varieties. (Mirzakhalov et al., 2021a) is a suite of resources for benchmarking MT in several Turkic languages. There has also been a large body of work on dialect identification across several languages (Jauhiainen et al., 2022; Hämäläinen et al., 2021). Recently CODET (Alam et al., 2023) was released as a contrastive dialectal MT benchmark covering 882 different variations from nine different languages. To the best of our knowledge, we are the first to do a large scale aggregation of dialectal data across several language clusters and tasks.

B Tasks of DIALECTBENCH

DIALECTBENCH includes 10 NLP tasks falling into four broader categories: structured prediction, sentence classification, question answering, and text generation. In Table 1, we present statistics for the datasets for each task and briefly discuss each task below.

Dependency parsing For the dependency parsing task, we include only those Universal Dependencies (UD) (Zeman et al., 2021) languages that have dialectal data. Beyond the data available in UD 2.12, we incorporate two additional datasets for African-American English (AAVE) Twitter data (Blodgett et al., 2018) and Singlish (Wang et al., 2017). To make these two datasets compatible with the UD processing pipeline, we replace the original dependency labels with the labels corresponding to the official UD formalism.

Part of speech (POS) tagging We use the same UD languages for POS tagging that we used for dependency parsing. At the same time, we use the POS data instances from Singlish (Wang et al., 2017). Moreover, we include six Finnish dialects, four Arabic dialects and Occitan through the UPOS label standardized pipeline proposed by Blaschke et al. (2023).

Named entity recognition (NER) We use data from the 176-language version of the Wikiann dataset processed by Rahimi et al. (2019). All these languages provide both training and test data. In addition, we include dialectal data from the original Wikiann dataset (282 languages) (Pan et al., 2017) for evaluation. Moreover, we include three Norwegian dialects (Johansen, 2019) with train, test and validation datasets that use a slightly different set of NER tags (GEO, ORG, OTH, PER) compared to the one we use in Wikiann (LOC, ORG, PER). We leave these levels as it is and do not convert to the Wikiann tags.

Dialect identification We include dialect identification experiments on Arabic, Greek, Portuguese, English, Spanish, and Swiss German dialectal datasets. In these datasets, we find large variations in the level of granularity with which dialects are classified. For instance, the MADAR corpus differentiates Arabic varieties at the city level (Bouamor et al., 2018), whereas our English and Spanish datasets are labeled with country names (Zampieri et al., 2023).

Sentiment classification Here we include several different Arabic varieties. Like other dialectal datasets, these datasets do not follow one standard labeling process. However, all datasets contain two main sentiment types: positive and negative. A number of datasets contain additional labels such as "objective" or "neutral." In our setting, we perform a further split of data to provide validation data for each dialect. However, we do not remove these extra labeled data for information preservation.

Topic classification We use the SIB-200 dataset (Adelani et al., 2023) for topic classification task SIB-200 was constructed from the FLORES-200 translation datasets. The authors annotated the English dataset of FLORES-200 with 6 topic labels and then further propagated these labels to the translated instances for all other languages. For our case of benchmarking dialectal segments, we use the dialectal and regional varieties from SIB-200.

cluster	Language code	Variety	# Sentences
anglic	eng_Latn	english	5010
arabic	acm_Arab	north mesopotamian arabic	5010
	acq_Arab	ta'izzi-adeni arabic	5010
	aeb_Arab	tunisian arabic	5010
	ajp_Arab	south levantine arabic	5010
	apc_Arab	levantine arabic (a:north)	5010
	arb_Arab	standard arabic	5010
	ars_Arab	najdi arabic	5010
	ary_Arab	moroccan arabic	5010
	arz_Arab	egyptian arabic	5010
common turkic	azb_Arab	south azerbaijani	5010
	azj_Latn	north azerbaijani	5010
	tur_Latn	central oghuz (m:spoken)	5010
gallo-italian	lij_Latn	ligurian	5010
	lmo_Latn	lombard	5010
	vec_Latn	venetian	5010
gallo-rhaetian	fur_Latn	friulian	5010
high german	lim_Latn	limburgan	5010
	ltz_Latn	luxemburgish	5010
italian romance	ita_Latn	italian	5010
	scn_Latn	sicilian	5010
kurdish	ckb_Arab	central kurdish	5010
	kmr_Latn	northern kurdish	5010
latvian	ltg_Latn	east latvian	5010
	lvs_Latn	latvian	5010
modern dutch	nld_Latn	dutch	5010
norwegian	nno_Latn	norwegian nynorsk (m:written)	5010
	nob_Latn	norwegian bokmål (m:written)	5010
sardo-corsican	srd_Latn	sardinian	5010
sinitic	yue_Hant	cantonese	5010
	zho_Hans	classical-middle-modern sinitic (o:simplified)	5010
	zho_Hant	classical-middle-modern sinitic (o:traditional)	5010
sotho-tswana (s.30)	nso_Latn	northern sotho	5010
	sot_Latn	southern sotho	5010
southwestern shifted romance	glg_Latn	galician	5010
	oci_Latn	occitan	5010
	por_Latn	portuguese (a:european)	5010
	spa_Latn	spanish	5010

Table 7: Data statistics for newly created translate-test natural language inference (NLI) dataset. We prepare this translate-test NLI dataset by translating XNLI (Conneau et al., 2018) english evaluation dataset.

Natural language inference For the natural language inference task, there is no existing dataset with varieties. So we use the existing English test set of XNLI (Conneau et al., 2018) and construct a multilingual dialect-focused translated evaluation dataset. We use a state-of-the-art machine translation model (NLLB-200 3B) to translate the English test set to 12 language clusters encompassing 40 varieties (Complete data statistics are reported in Table 7). After that, we perform zero-shot cross-lingual transfer from the English finetuned NLI model. We refer to this setting as *translate-test*.

Multiple-choice machine reading comprehension This task aims to evaluate the capability of multiple-choice question answering given a context passage. The question could be answered by understanding the context passage while the right answer is given at one of the multiple choices. We use the Chinese, Sotho and Arabic clusters data from the recently released Belebele MRC dataset (Bandarkar et al., 2023). This is an evaluation-only dataset.

Extractive question answering This task predicts the answer span given a question and context passage. We use the SD-QA dialectal question-answering dataset (Faisal et al., 2021). SD-QA is an evaluation dataset built on top of TyDiQA (Clark et al., 2020), another well-known typologically diverse question-answering dataset. SD-QA contains the spoken utterances and transcription of the original TyDiQA question from speakers of English, Bengali, Arabic, Korean, and Swahili. We only use the textual part that contains the transcription of the dialectal spoken question matching the original TyDiQA question text. Note that since the transcriptions of the questions are obtained through automatic speech recognition, they may include both dialectal variations *and* noise due to ASR transcription errors.

Machine translation We evaluate variety translation using the CODET benchmark (Alam et al., 2023) and the TIL MT corpus (Mirzakhalov et al., 2021b). CODET contains a contrastive dataset of 882 different varieties from nine different languages. We evaluate dialects here at city level for all languages except Italian and Swiss-German and aggregate dialects at the region level for them. The TIL corpus contains parallel translations across 22 Turkic languages, but in our evaluations we only include 8 turkic languages (Turkic, Sakha, Kazakh, Karakalpak, Bashkir, Azerbaijani, Kyrgyz) that have parallel English translations.

C Varieties and Clusters of DIALECTBENCH

C.1 DIALECTBENCH Variety list

Table 8: Varieties represented in DialectBench.

Language cluster	Variety name	Glottocode	mBERT seen	UDP	POS	NER	SDQA	RCMC	NLI	TC	SC	DI	MT
albanian	albanian	alba1267	✓	✓	✓	-	-	-	-	-	-	-	-
	gheg albanian	gheg1238	-	✓	✓	-	-	-	-	-	-	-	-
anglic	african american vernacular english	afri1276	-	✓	-	-	-	-	-	-	-	-	-
	australian english	aust1314	-	-	-	-	✓	-	-	-	-	-	-
	english	stan1293	✓	✓	✓	✓	-	✓	✓	✓	-	✓	-
	english (a:scotland)	stan1293	-	-	-	-	✓	-	-	-	-	-	-
	english (a:uk)	stan1293	-	-	-	-	-	-	-	-	-	✓	-
	english (o:controlled)	stan1293	-	-	-	✓	-	-	-	-	-	-	-
	indian english (a:north)	indi1255	-	-	-	-	✓	-	-	-	-	-	-
	indian english (a:south)	indi1255	-	-	-	-	✓	-	-	-	-	-	-
	irish english	iris1255	-	-	-	-	✓	-	-	-	-	-	-
	jamaican creole english	jama1262	-	-	-	✓	-	-	-	-	-	-	-
	kenyan english	keny1281	-	-	-	-	✓	-	-	-	-	-	-
	new zealand english	newz1240	-	-	-	-	✓	-	-	-	-	-	-
	nigerian english	nige1260	-	-	-	-	✓	-	-	-	-	-	-
	north american english	nort3314	-	-	-	-	-	-	-	-	-	✓	-
	old english (ca. 450-1100)	olde1238	-	-	-	✓	-	-	-	-	-	-	-
	philippine english	phil1246	-	-	-	-	✓	-	-	-	-	-	-
	singlish	sing1272	-	✓	✓	-	-	-	-	-	-	-	-
	southeast american english	sout3300	-	-	-	-	✓	-	-	-	-	-	-
	southern african english	sout3331	-	-	-	-	✓	-	-	-	-	-	-
arabic	aleppo	alep1241	-	-	-	-	-	-	-	-	-	✓	✓
	algerian arabic	alge1239	-	-	-	-	✓	-	-	-	✓	✓	✓
	arabian peninsula arabic (a:yemen)	arab1393	-	-	-	-	-	-	-	-	-	✓	✓
	arabic (a:bahrain)	-	-	-	-	-	✓	-	-	-	-	-	-
	arabic (a:jordan)	-	-	-	-	-	✓	-	-	-	✓	-	-
	arabic (a:saudi-arabia)	-	-	-	-	-	✓	-	-	-	✓	-	-

Continued on next page

Table 8: Varieties represented in DialectBench.

Language clusters	Variety name	Glottocode	mBERT seen	UDP	POS	NER	SDQA	RCMC	NLI	TC	SC	DI	MT
	egyptian arabic	egyp1253	-	-	✓	✓	✓	✓	✓	✓	✓	-	-
	egyptian arabic (a:alx)	egyp1253	-	-	-	-	-	-	-	-	-	✓	✓
	egyptian arabic (a:asw)	egyp1253	-	-	-	-	-	-	-	-	-	✓	✓
	egyptian arabic (a:cai)	egyp1253	-	-	-	-	-	-	-	-	-	✓	✓
	egyptian arabic (a:kha)	egyp1253	-	-	-	-	-	-	-	-	-	✓	✓
	fez. meknes	fezm1238	-	-	-	-	-	-	-	-	-	✓	✓
	gilit mesopotamian arabic	meso1252	-	-	-	-	-	-	-	-	-	✓	✓
	gulf arabic	gulf1241	-	-	✓	-	-	-	-	-	-	-	-
	gulf arabic (a:doh)	gulf1241	-	-	-	-	-	-	-	-	-	✓	✓
	gulf arabic (a:jed)	gulf1241	-	-	-	-	-	-	-	-	-	✓	✓
	gulf arabic (a:mus)	gulf1241	-	-	-	-	-	-	-	-	-	✓	✓
	gulf arabic (a:riy)	gulf1241	-	-	-	-	-	-	-	-	-	✓	✓
	levantine arabic	nort3139	-	-	✓	-	-	-	-	-	-	-	-
	levantine arabic (a:north)	nort3139	-	-	-	-	-	✓	✓	✓	-	-	-
	levantine arabic (a:north-dam)	nort3139	-	-	-	-	-	-	-	-	-	✓	✓
	libyan arabic	liby1240	-	-	-	-	-	-	-	-	-	-	-
	libyan arabic (a:ben)	liby1240	-	-	-	-	-	-	-	-	-	✓	✓
	moroccan arabic	moro1292	-	-	-	-	✓	✓	✓	✓	✓	-	-
	najdi arabic	najd1235	-	-	-	-	-	✓	✓	✓	-	-	-
	north african arabic	nort3191	-	✓	✓	-	-	-	-	-	-	-	-
	north mesopotamian arabic	nort3142	-	-	-	-	-	✓	✓	✓	-	-	-
	north mesopotamian arabic (a:bas)	nort3142	-	-	-	-	-	-	-	-	-	✓	✓
	north mesopotamian arabic (a:mos)	nort3142	-	-	-	-	-	-	-	-	-	✓	✓
	rabat-casablanca arabic	raba1252	-	-	-	-	-	-	-	-	-	✓	✓
	sfax	sfax1238	-	-	-	-	-	-	-	-	-	✓	✓
	south levantine arabic	sout3123	-	✓	✓	-	-	-	✓	✓	✓	-	-
	south levantine arabic (a:south-amm)	sout3123	-	-	-	-	-	-	-	-	-	✓	✓
	south levantine arabic (a:south-jer)	sout3123	-	-	-	-	-	-	-	-	-	✓	✓
	south levantine arabic (a:south-sal)	sout3123	-	-	-	-	-	-	-	-	-	✓	✓
	standard arabic	stan1318	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	-
	sunni beiruti arabic	sunn1238	-	-	-	-	-	-	-	-	-	✓	✓
	ta'izzi-adeni arabic	taiz1242	-	-	-	-	-	-	✓	✓	-	-	-
	tripolitanian arabic	trip1239	-	-	-	-	-	-	-	-	-	✓	✓
	tunisian arabic	tuni1259	-	-	-	-	✓	-	✓	✓	✓	-	-
	tunisian arabic (a:tun)	tuni1259	-	-	-	-	-	-	-	-	-	✓	✓
	tunisian arabic (r:casual)	tuni1259	-	-	-	-	-	-	-	-	✓	-	-
basque	basque (a:amenduze)	basq1248	✓	-	-	-	-	-	-	-	-	-	✓
	basque (a:azkaine)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:baigorri)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:barkoxe)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:donibane)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:garruze)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:iholdi)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:jatsu)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:jutsi)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:larzabale)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:luhuso)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:sara)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:senpere)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:suhuskune)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	basque (a:uharte)	basq1248	-	-	-	-	-	-	-	-	-	-	✓
	navarro-labourdin basque (a:behorlegi)	basq1249	-	-	-	-	-	-	-	-	-	-	✓
	navarro-labourdin basque (a:bidarrai)	basq1249	-	-	-	-	-	-	-	-	-	-	✓
	navarro-labourdin basque (a:helete)	basq1249	-	-	-	-	-	-	-	-	-	-	✓
	navarro-labourdin basque (a:mugerre)	basq1249	-	-	-	-	-	-	-	-	-	-	✓
	navarro-labourdin basque (a:urruna)	basq1249	-	-	-	-	-	-	-	-	-	-	✓
	souletin (a:maule)	basq1250	-	-	-	-	-	-	-	-	-	-	✓
bengali	vanga (a:barisal)	vang1242	-	-	-	-	-	-	-	-	-	-	✓
	vanga (a:dhaka)	vang1242	✓	-	-	-	✓	-	-	-	-	-	✓
	vanga (a:jessore)	vang1242	-	-	-	-	-	-	-	-	-	-	✓
	vanga (a:khulna)	vang1242	-	-	-	-	-	-	-	-	-	-	✓
	vanga (a:kushtia)	vang1242	-	-	-	-	-	-	-	-	-	-	✓
	vanga (a:west bengal)	vang1242	✓	-	-	-	✓	-	-	-	-	-	-
circassian	adyghe	adyg1241	-	-	-	✓	-	-	-	-	-	-	-
	kabardian	kaba1278	-	-	-	✓	-	-	-	-	-	-	-
common turkic	bashkir	bash1264	✓	-	-	-	-	-	-	-	-	-	✓
	central oghuz	azer1255	-	-	-	-	-	-	-	-	-	-	✓
	central oghuz (m:spoken)	azer1255	-	-	-	-	-	-	✓	✓	-	-	-
	crimean tatar	crim1257	-	-	-	✓	-	-	-	-	-	-	-
	kara-kalpak	kara1467	-	-	-	-	-	-	-	-	-	-	✓
	kazakh	kaza1248	✓	-	-	-	-	-	-	-	-	-	✓
	kirghiz	kirg1245	✓	-	-	-	-	-	-	-	-	-	✓
	north azerbaijani	nort2697	✓	-	-	✓	-	-	✓	✓	-	-	-
	sakha	yaku1245	-	-	-	-	-	-	-	-	-	-	✓
	south azerbaijani	sout2697	✓	-	-	✓	-	-	✓	✓	-	-	-
	turkish	nucl1301	✓	-	-	✓	-	-	-	-	-	-	✓
	uzbek	uzbe1247	✓	-	-	-	-	-	-	-	-	-	✓
eastern romance	aromanian	arom1237	-	-	-	✓	-	-	-	-	-	-	-
	moldavian	mold1248	-	-	-	✓	-	-	-	-	-	-	-

Continued on next page

Table 8: Varieties represented in DialectBench.

Language clusters	Variety name	Glottocode	mBERT seen	UDP	POS	NER	SDQA	RCMC	NLI	TC	SC	DI	MT
	romanian	roma1327	✓	-	-	✓	-	-	-	-	-	-	-
eastern-western armenian	eastern armenian	nucl1235	✓	✓	✓	-	-	-	-	-	-	-	-
	western armenian	homs1234	-	✓	✓	-	-	-	-	-	-	-	-
farsic	dari	dari1249	-	-	-	-	-	-	-	-	-	-	✓
frisian	ems-weser frisian	sate1242	-	-	-	✓	-	-	-	-	-	-	-
	northern frisian	nort2626	-	-	-	✓	-	-	-	-	-	-	-
	western frisian	west2354	✓	-	-	✓	-	-	-	-	-	-	-
gallo-italian	emiliano-romagnolo	emil1243	-	-	-	✓	-	-	-	-	-	-	-
	ligurian	ligu1248	-	✓	✓	✓	-	-	✓	✓	-	-	-
	lombard	lomb1257	✓	-	-	✓	-	-	✓	✓	-	-	-
	piemontese	piem1238	✓	-	-	✓	-	-	-	-	-	-	-
	venetian	vene1258	-	-	-	✓	-	-	✓	✓	-	-	-
gallo-rhaetian	anglo-norman	angl1258	-	-	-	✓	-	-	-	-	-	-	-
	arpitan	fran1260	-	-	-	✓	-	-	-	-	-	-	-
	french	stan1290	✓	✓	✓	✓	-	-	-	-	-	-	-
	french (a:paris)	stan1290	-	✓	✓	-	-	-	-	-	-	-	-
	friulian	friu1240	-	-	-	✓	-	-	✓	✓	-	-	-
	old french (842-ca. 1400)	oldf1239	-	✓	✓	-	-	-	-	-	-	-	-
	romansh	roma1326	-	-	-	✓	-	-	-	-	-	-	-
	walloon	wall1255	-	-	-	✓	-	-	-	-	-	-	-
greater panjabic	eastern panjabi	panj1256	✓	-	-	✓	-	-	-	-	-	-	-
	western panjabi	west2386	✓	-	-	✓	-	-	-	-	-	-	-
greek	apulian greek	apul1237	-	-	-	-	-	-	-	-	-	-	✓
	cretan	cret1244	-	-	-	-	-	-	-	-	-	-	✓
	cypriot greek	cypr1249	-	-	-	-	-	-	-	-	-	-	-
	cypriot greek (r:casual, m:written, i:fb)	cypr1249	-	-	-	-	-	-	-	-	-	✓	-
	cypriot greek (r:casual, m:written, i:other)	cypr1249	-	-	-	-	-	-	-	-	-	✓	-
	cypriot greek (r:casual, m:written, i:twitter)	cypr1249	-	-	-	-	-	-	-	-	-	✓	-
	modern greek	mode1248	✓	-	-	✓	-	-	-	-	-	-	-
	modern greek (r:casual, m:written, i:fb)	mode1248	-	-	-	-	-	-	-	-	-	✓	-
	modern greek (r:casual, m:written, i:other)	mode1248	-	-	-	-	-	-	-	-	-	✓	-
	modern greek (r:casual, m:written, i:twitter)	mode1248	-	-	-	-	-	-	-	-	-	✓	-
	pontic	pont1253	-	-	-	✓	-	-	-	-	-	-	-
high german	bavarian	bava1246	✓	-	-	✓	-	-	-	-	-	-	-
	central alemannic	swis1247	-	-	-	✓	-	-	-	-	-	-	-
	central alemannic (a:ag)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:ai)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:ar)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:be)	swis1247	-	-	-	-	-	-	-	-	-	✓	✓
	central alemannic (a:bl)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:bs)	swis1247	-	-	-	-	-	-	-	-	-	✓	✓
	central alemannic (a:fr)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:gl)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:gr)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:lu)	swis1247	-	-	-	-	-	-	-	-	-	✓	✓
	central alemannic (a:nw)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:ow)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:sg)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:sh)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:so)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:sz)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:tg)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:ur)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:vs)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:zg)	swis1247	-	-	-	-	-	-	-	-	-	-	✓
	central alemannic (a:zh)	swis1247	-	✓	✓	-	-	-	-	-	-	✓	✓
	central bavarian	cent1967	-	-	-	-	-	-	-	-	-	-	✓
	german	stan1295	✓	✓	✓	✓	-	-	-	-	-	-	-
	kölsch	kols1241	-	-	-	✓	-	-	-	-	-	-	-
	limburgan	limb1263	-	-	-	✓	-	-	✓	✓	-	-	-
	luxemburgish	luxe1243	✓	-	-	✓	-	-	✓	✓	-	-	-
	pennsylvania german	penn1240	-	-	-	✓	-	-	-	-	-	-	-
	pfaelzisch-lothringisch	pala1330	-	-	-	✓	-	-	-	-	-	-	-
	upper saxon	uppe1465	-	-	-	-	-	-	-	-	-	-	✓
hindustani	fiji hindi	fiji1242	-	-	-	✓	-	-	-	-	-	-	-
	hindi	hind1269	✓	-	-	✓	-	-	-	-	-	-	-
inuit	alaskan inupiaq	inup1234	-	-	-	✓	-	-	-	-	-	-	-
	kalaallisut	kala1399	-	-	-	✓	-	-	-	-	-	-	-
italian romance	continental southern italian	neap1235	-	✓	✓	✓	-	-	-	-	-	-	-
	italian	ital1282	✓	✓	✓	✓	-	-	✓	✓	-	-	-
	italian (a:abruzzo)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:basilicata)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:calabria)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:campania)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:emilia-romagna)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:friuli-venezia giulia)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:lazio)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:liguria)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:lombardia)	ital1282	-	-	-	-	-	-	-	-	-	-	✓

Continued on next page

Table 8: Varieties represented in DialectBench.

Language clusters	Variety name	Glottocode	mBERT seen	UDP	POS	NER	SDQA	RCMC	NLI	TC	SC	DI	MT
	italian (a:marche)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:molise)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:piemonte)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:puglia)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:sardegna)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:sicilia)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:toscana)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:trentino-alto adige/südtirol)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:umbria)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:unknown)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (a:veneto)	ital1282	-	-	-	-	-	-	-	-	-	-	✓
	italian (r:casual, m:written, i:tweet)	ital1282	✓	✓	✓	-	-	-	-	-	-	-	-
	italian (r:formal, m:written, i:essay)	ital1282	✓	✓	✓	-	-	-	-	-	-	-	-
	italian romance (a:apulia, m:spoken, i:tarantino)	-	-	-	-	✓	-	-	-	-	-	-	-
	sicilian	sici1248	✓	-	-	✓	-	-	✓	✓	-	-	-
komi	komi	komi1267	-	-	-	✓	-	-	-	-	-	-	-
	komi-permyak	komi1269	-	✓	✓	✓	-	-	-	-	-	-	-
	komi-zyrian (m:spoken)	komi1268	-	✓	✓	-	-	-	-	-	-	-	-
	komi-zyrian (m:written)	komi1268	-	✓	✓	-	-	-	-	-	-	-	-
korean	korean (a:south-eastern, m:spoken)	kore1280	-	-	-	-	✓	-	-	-	-	-	-
	seoul (m:spoken)	seou1239	✓	-	-	-	✓	-	-	-	-	-	-
kurdish	central kurdish	cent1972	-	-	-	✓	-	-	✓	✓	-	-	-
	kurdish	kurd1259	-	-	-	✓	-	-	-	-	-	-	-
	northern kurdish	nort2641	-	-	-	-	-	-	✓	✓	-	-	-
	sine'i	sine1239	-	-	-	-	-	-	-	-	-	-	✓
	sorani	sora1257	-	-	-	-	-	-	-	-	-	-	✓
latvian	east latvian	east2282	-	-	-	✓	-	-	✓	✓	-	-	-
	latvian	latv1249	✓	-	-	✓	-	-	✓	✓	-	-	-
mari	eastern mari	east2328	-	-	-	✓	-	-	-	-	-	-	-
	western mari	west2392	-	-	-	✓	-	-	-	-	-	-	-
modern dutch	dutch	dutc1256	✓	-	-	✓	-	-	✓	✓	-	-	-
	western flemish	vlaa1240	-	-	-	✓	-	-	-	-	-	-	-
	zeeuws	zeeu1238	-	-	-	✓	-	-	-	-	-	-	-
neva	central and north pohjanmaa (a:ostrobothnian)	cent1985	-	-	✓	-	-	-	-	-	-	-	-
	estonian	esto1258	✓	-	✓	-	-	-	-	-	-	-	-
	finnish	finn1318	✓	-	✓	-	-	-	-	-	-	-	-
	häme (a:tavastian)	hame1240	-	-	✓	-	-	-	-	-	-	-	-
	neva (a:south-west trans)	-	-	-	✓	-	-	-	-	-	-	-	-
	savo (a:savonian)	savo1254	-	-	✓	-	-	-	-	-	-	-	-
	southeastern finnish (a:south-east)	sout1743	-	-	✓	-	-	-	-	-	-	-	-
	southwestern finnish (a:south-west)	sout2677	-	-	✓	-	-	-	-	-	-	-	-
norwegian	norwegian	norw1258	✓	-	-	-	-	-	-	-	-	-	-
	norwegian (a:eastern)	norw1258	-	-	-	-	-	-	-	-	-	-	✓
	norwegian (a:setesdal)	norw1258	-	-	-	-	-	-	-	-	-	-	✓
	norwegian (a:southwestern)	norw1258	-	-	-	-	-	-	-	-	-	-	✓
	norwegian (m:written, i:samnorsk)	-	-	-	-	✓	-	-	-	-	-	-	-
	norwegian bokmål (m:written)	norw1259	✓	✓	✓	✓	-	-	✓	✓	-	-	-
	norwegian nynorsk (m:written)	norw1262	✓	✓	✓	✓	-	-	✓	✓	-	-	-
	norwegian nynorsk (m:written, i:old)	norw1262	-	✓	✓	-	-	-	-	-	-	-	-
saami	north saami	nort2671	-	✓	✓	-	-	-	-	-	-	-	-
	skolt saami	skol1241	-	✓	✓	-	-	-	-	-	-	-	-
sabellic	umbrian	umbr1253	-	✓	✓	-	-	-	-	-	-	-	-
sardo-corsican	corsican	cors1241	-	-	-	✓	-	-	-	-	-	-	-
	sardinian	sard1257	-	-	-	✓	-	-	✓	✓	-	-	-
serbian-croatian-bosnian	bosnian standard	bosn1245	✓	-	-	✓	-	-	-	-	-	-	-
	croatian standard	croa1245	✓	-	-	✓	-	-	-	-	-	-	-
	serbian standard	serb1264	✓	-	-	✓	-	-	-	-	-	-	-
	serbian-croatian-bosnian	sout1528	✓	-	-	✓	-	-	-	-	-	-	-
sinitic	cantonese	cant1236	-	-	-	✓	-	-	✓	✓	-	-	-
	classical chinese	lite1248	-	✓	✓	✓	-	-	-	-	-	-	-
	classical-middle-modern sinitic (a:hongkong, o:traditional)	clas1255	-	✓	✓	-	-	-	-	-	-	-	-
	classical-middle-modern sinitic (o:simplified)	clas1255	✓	✓	✓	-	-	✓	✓	✓	-	-	-
	classical-middle-modern sinitic (o:traditional)	clas1255	✓	-	-	-	-	✓	✓	✓	-	-	-
	hakka chinese	hakk1236	-	-	-	✓	-	-	-	-	-	-	-
	mandarin chinese	mand1415	-	-	-	✓	-	-	-	-	-	-	-
	mandarin chinese (a:mainland, o:simplified)	mand1415	-	-	-	-	-	-	-	-	-	✓	-

Continued on next page

Table 8: Varieties represented in DialectBench.

Language clusters	Variety name	Glottocode	mBERT seen	UDP	POS	NER	SDQA	RCMC	NLI	TC	SC	DI	MT
	mandarin chinese (a:mainland, o:traditional, i:synthetic)	mand1415	-	-	-	-	-	-	-	-	-	✓	-
	mandarin chinese (a:taiwan, o:simplified)	mand1415	-	-	-	-	-	-	-	-	-	✓	-
	mandarin chinese (a:taiwan, o:traditional, i:synthetic)	mand1415	-	-	-	-	-	-	-	-	-	✓	-
	min nan chinese	minn1241	-	-	-	✓	-	-	-	-	-	-	-
	wu chinese	wuch1236	-	-	-	✓	-	-	-	-	-	-	-
sorbian	lower sorbian	lowe1385	-	-	-	✓	-	-	-	-	-	-	-
	upper sorbian	uppe1395	-	-	-	✓	-	-	-	-	-	-	-
sotho-tswana (s.30)	northern sotho	nort3233	-	-	-	✓	-	✓	✓	✓	-	-	-
	southern sotho	sout2807	-	-	-	✓	-	✓	✓	✓	-	-	-
southwestern shifted romance	brazilian portuguese	braz1246	-	✓	✓	-	-	-	-	-	-	✓	-
	extremaduran	extr1243	-	-	-	✓	-	-	-	-	-	-	-
	galician	gali1258	✓	-	-	✓	-	-	✓	✓	-	-	-
	latin american spanish	amer1254	-	-	-	-	-	-	-	-	-	✓	-
	mirandese	mira1251	-	-	-	✓	-	-	-	-	-	-	-
	occitan	occi1239	✓	-	✓	✓	-	-	✓	✓	-	-	✓
	portuguese (a:european)	port1283	-	✓	✓	-	-	-	✓	✓	-	✓	-
	portuguese (i:mix)	port1283	-	✓	✓	-	-	-	-	-	-	-	-
	portuguese (m:written)	port1283	✓	✓	✓	-	-	-	-	-	-	✓	-
	spanish	stan1288	✓	-	-	✓	-	-	✓	✓	-	✓	-
	spanish (a:europe)	stan1288	-	-	-	-	-	-	-	-	-	✓	-
swahili	swahili	swah1253	✓	-	-	-	-	-	-	-	-	-	-
	swahili (a:kenya)	swah1253	-	-	-	-	✓	-	-	-	-	-	-
	swahili (a:tanzania)	swah1253	-	-	-	-	✓	-	-	-	-	-	-
tigrinya	tigrinya (a:eritrea)	tigr1271	-	-	-	-	-	-	-	-	-	-	✓
	tigrinya (a:ethiopia)	tigr1271	-	-	-	-	-	-	-	-	-	-	✓
tupi-guarani subgroup i.a	mbyá guaraní (a:brazil)	mbya1239	-	✓	✓	-	-	-	-	-	-	-	-
	mbyá guaraní (a:paraguay)	mbya1239	-	✓	✓	-	-	-	-	-	-	-	-
	old guarani	oldp1258	-	✓	✓	-	-	-	-	-	-	-	-
west low german	west low german	west2357	✓	✓	✓	✓	-	-	-	-	-	-	-
yoruba	yoruba (a:central nigeria)	yoru1245	✓	-	-	-	-	-	-	-	-	-	✓

C.2 Language Clusters and Representative Varieties

Table 9: Language clusters and their standard varieties.

Task	Cluster Name	Cluster Representative
DEP. Parsing	albanian	albanian
	arabic	standard arabic
	eastern-western armenian	western armenian
	sinitic	classical-middle-modern sinitic (o:simplified)
	anglic	english
	gallo-rhaetian	french
	high german	german
	tupi-guarani subgroup i.a	old guarani
	italian romance	italian (r:formal, m:written, i:essay)
	komi	komi-zyrian (m:spoken)
	norwegian	norwegian bokmål (m:written)
	southwestern shifted romance	brazilian portuguese
	saami	north saami
POS Tagging	albanian	albanian
	anglic	english
	arabic	standard arabic
	eastern-western armenian	western armenian
	gallo-rhaetian	french
	high german	german
	italian romance	italian (r:formal, m:written, i:essay)
	komi	komi-zyrian (m:spoken)
	neva	finnish
	norwegian	norwegian bokmål (m:written)
	saami	north saami
	sinitic	classical-middle-modern sinitic (o:simplified)
	southwestern shifted romance	brazilian portuguese
	tupi-guarani subgroup i.a	mbyá guaraní (a:paraguay)
NER	anglic	english
	arabic	standard arabic
	circassian	adyghe
	common turkic	turkish
	eastern romance	romanian
	frisian	western frisian
	gallo-italian	piemontese
	gallo-rhaetian	french
	greater panjabic	western panjabi
	greek	modern greek
	high german	german
	hindustani	hindi
	inuit	alaskan inupiaq
	italian romance	italian
	komi	komi
	kurdish	kurdish

Continued on next page

Table 9: Language clusters and their clusters representatives.

Task	Cluster Name	Cluster Representative
	latvian	latvian
	mari	eastern mari
	modern dutch	dutch
	norwegian	norwegian bokmål (m:written)
	sardo-corsican	sardinian
	serbian-croatian-bosnian	croatian standard
	sinitic	mandarin chinese
	sorbian	lower sorbian
	sotho-tswana (s.30)	southern sotho
	southwestern shifted romance	spanish
EQA	anglic	southeast american english
	arabic	arabic (a:saudi-arabia)
	bengali	vanga (a:dhaka)
	korean	seoul (m:spoken)
	swahili	swahili (a:kenya)
MRC	anglic	english
	arabic	standard arabic
	sinitic	classical-middle-modern sinitic (o:simplified)
	sotho-tswana (s.30)	northern sotho
TC	anglic	english
	arabic	standard arabic
	common turkic	north azerbaijani
	gallo-italian	venetian
	high german	luxemburgish
	italian romance	italian
	kurdish	northern kurdish
	latvian	latvian
	norwegian	norwegian bokmål (m:written)
	sinitic	classical-middle-modern sinitic (o:simplified)
	sotho-tswana (s.30)	northern sotho
	southwestern shifted romance	spanish
SC	arabic	standard arabic
MT-Dialect	arabic	gulf arabic (a:riy)
	basque	basque (a:azkaine)
	bengali	vanga (a:dhaka)
	high german	central bavarian
	kurdish	sorani
	norwegian	norwegian (a:eastern)
	tigrinya	tigrinya (a:ethiopia)
	common turkic	turkish
	greek	cretan
MT-Region	italian romance	italian (a:umbria)
	high german	central alemannic (a:zh)
DId	anglic	english (a:uk)
	arabic	standard arabic
	greek	modern greek (r:casual, m:written, i:twitter)
	high german	central alemannic (a:zh)
	sinitic	mandarin chinese (a:mainland, o:simplified)
	southwestern shifted romance	brazilian portuguese
NLI	anglic	english
	arabic	standard arabic
	common turkic	north azerbaijani
	gallo-italian	venetian
	high german	luxemburgish
	italian romance	italian
	kurdish	northern kurdish
	latvian	latvian
	norwegian	norwegian bokmål (m:written)
	sinitic	classical-middle-modern sinitic (o:simplified)
	sotho-tswana (s.30)	northern sotho
	southwestern shifted romance	spanish

D Result Visualizations

D.1 Regional maps with aggregated Machine Translation scores

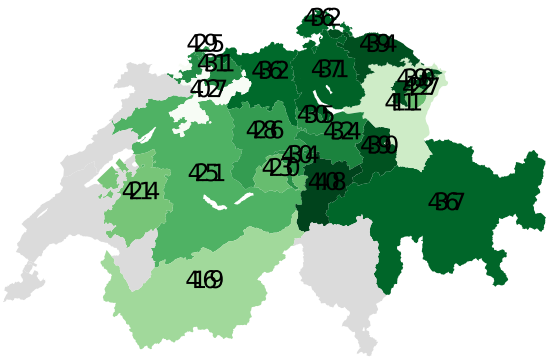


Figure 5: Map of Switzerland with aggregated BLEU scores of Swiss-German variety per region

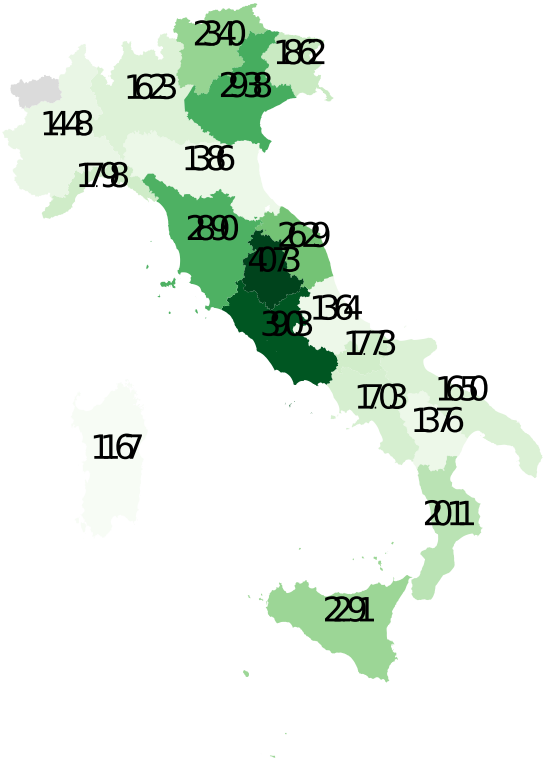


Figure 6: Map of Italy with aggregated BLEU scores of Italian variety per region

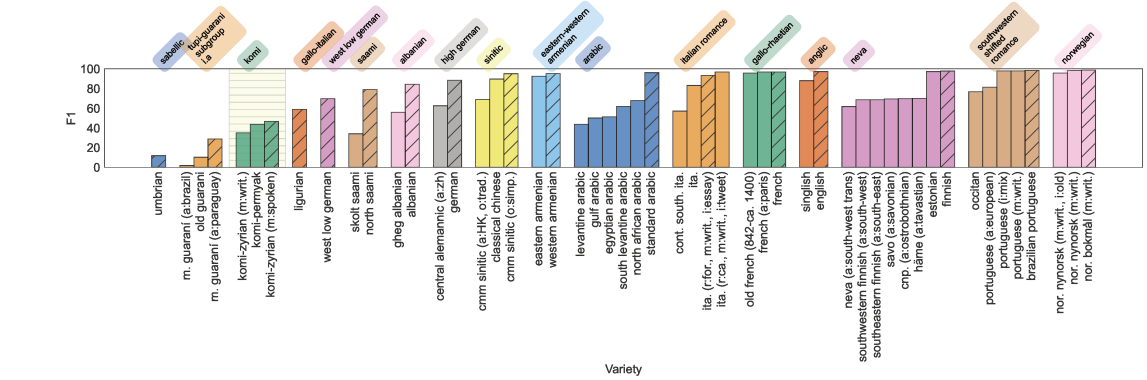
D.2 Task Specific Plot for Maximum Scores

D.3 Dialectal Gap visualizations (zeroshot)

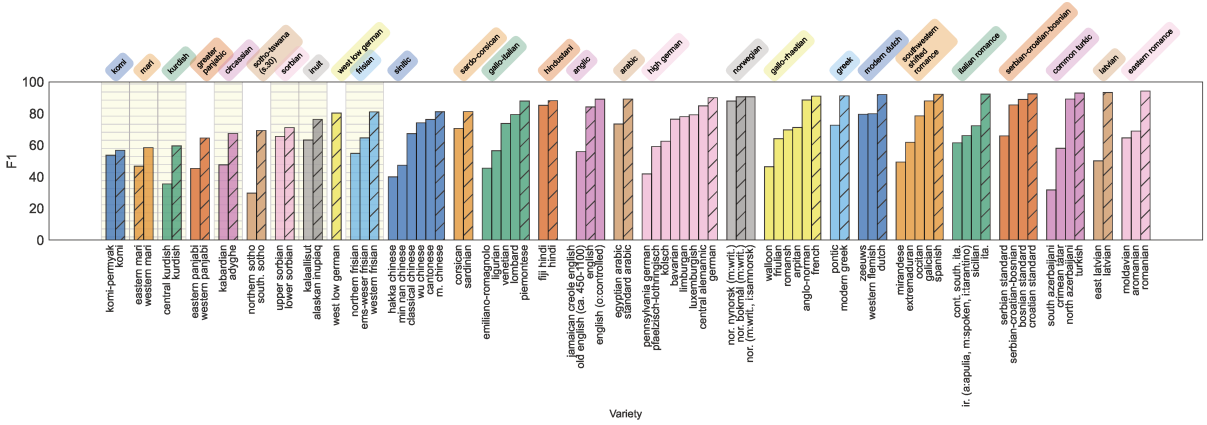
E Evaluation results

cluster	variety	UD-code	Zeroshot (mBERT)	Zeroshot (XLM-R)	Finetune (mBERT)	Finetune (XLM-R)
albanian	albanian	UD_Albanian-TSA	81.78	83.08	-	-
	gheg albanian	UD_Gheg-GPS	38.14	43.50	-	-
anglic	english	UD_English-EWT	91.55	91.10	91.55	91.1
	singlish	singlish	69.88	68.55	82.36	11.0
	african american vernacular english	TwitterAAE	50.53	52.53	-	-
arabic	standard arabic	UD_Arabic-PADT	50.87	55.91	88.47	2.48
	south levantine arabic	UD_South_Levantine_Arabic-MADAR	49.94	45.75	-	-
	north african arabic	UD_Maghrebi_Arabic_French-Arabizi	34.33	28.04	5.46	54.7
eastern-western armenian	western armenian	UD_Western_Armenian-ArmTDP	54.63	59.71	88.84	89.29
	eastern armenian	UD_Armenian-ArmTDP	53.27	62.63	85.55	87.01
gallo-italian	ligurian	UD_Ligurian-GLT	50.22	43.78	13.06	9.1
gallo-rhaetian	french	UD_French-ParTUT	80.87	82.17	93.93	92.09
	french (a:paris)	UD_French-ParisStories	57.64	61.36	77.54	77.54
	old french (842-ca. 1400)	UD_Old_French-SRCMF	56.12	46.52	91.51	89.78
high german	german	UD_German-LIT	72.65	76.50	-	-
	central alemannic (a:zh)	UD_Swiss_German-UZH	36.77	34.70	-	-
italian romance	italian	UD_Italian-PUD	78.99	78.58	-	-
	italian (r:formal, m:written, i:essay)	UD_Italian-MarkIT	76.83	79.38	86.34	83.81
	italian (r:casual, m:written, i:tweet)	UD_Italian-PoSTWITA	61.88	62.13	85.82	86.32
	continental southern italian	UD_Neapolitan-RB	30.00	50.00	-	-
komi	komi-zyrian (m:spoken)	UD_Komi_Zyrian-IKDP	26.89	32.14	-	-
	komi-permyak	UD_Komi_Permyak-UH	26.12	30.91	-	-
	komi-zyrian (m:written)	UD_Komi_Zyrian-Lattice	21.01	27.55	-	-
norwegian	norwegian bokmål (m:written)	UD_Norwegian-Bokmaal	79.55	82.95	93.57	93.49
	norwegian nynorsk (m:written)	UD_Norwegian-Nynorsk	76.45	76.76	93.15	93.14
	norwegian nynorsk (m:written, i:old)	UD_Norwegian-NynorskLIA	56.58	56.08	78.39	8.25
saami	north saami	UD_North_Sami-Giella	23.58	16.87	67.56	5.96
	skolt saami	UD_Skolt_Sami-Giellagas	19.41	28.28	-	-
sabellic	umbrian	UD_Umbrian-IKUVINA	33.21	28.75	-	-
sinitic	classical-middle-modern sinitic (a:hongkong, o:traditional)	UD_Chinese-HK	58.97	61.78	-	-
	classical-middle-modern sinitic (o:simplified)	UD_Chinese-GSDSimp	53.35	52.21	87.85	87.37
	classical chinese	UD_Classical_Chinese-Kyoto	46.72	35.14	19.65	18.11
southwestern shifted romance	portuguese (a:european)	UD_Portuguese-PUD	76.67	77.28	-	-
	portuguese (i:mix)	UD_Portuguese-Bosque	75.89	77.96	92.64	92.15
	brazilian portuguese	UD_Portuguese-GSD	73.30	74.21	94.37	93.75
	portuguese (m:written)	UD_Portuguese-CINTIL	69.34	72.94	83.35	84.31
tupi-guarani subgroup i.a	old guarani	UD_Guarani-OldTuDeT	22.58	29.03	-	-
	mbyá guaraní (a:paraguay)	UD_Mbya_Guarani-Thomas	13.51	11.15	-	-
	mbyá guaraní (a:brazil)	UD_Mbya_Guarani-Dooley	8.95	4.23	-	-
west low german	west low german	UD_Low_Saxon-LSDC	40.75	37.38	-	-

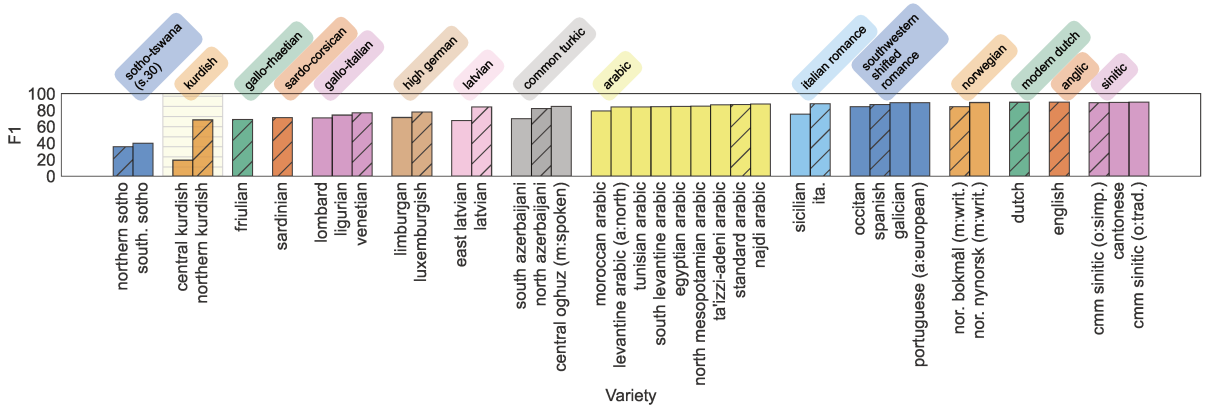
Table 10: Dependency parsing evaluation report comprising zeroshot score and in-language finetuning. We report UAS as evaluation score. Zeroshot scores are evaluated using model finetuned on Standerd English. If training data is not available, we skip those languages (mentioned as '-') for in-language finetuning.



(a) Parts of Speech Tagging

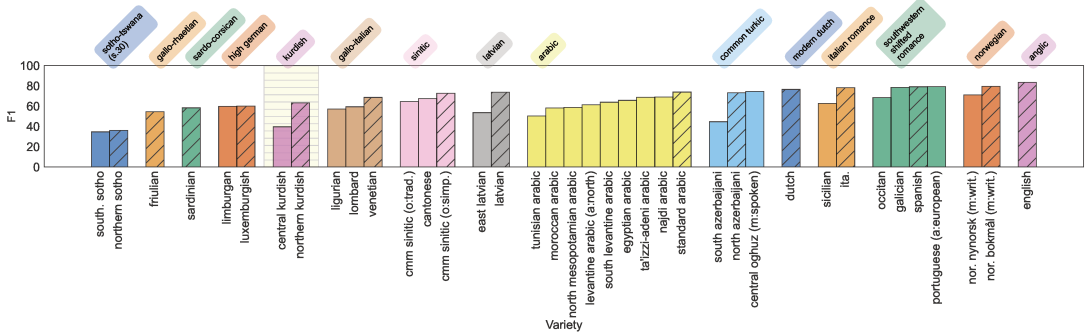


(b) Named Entity Recognition

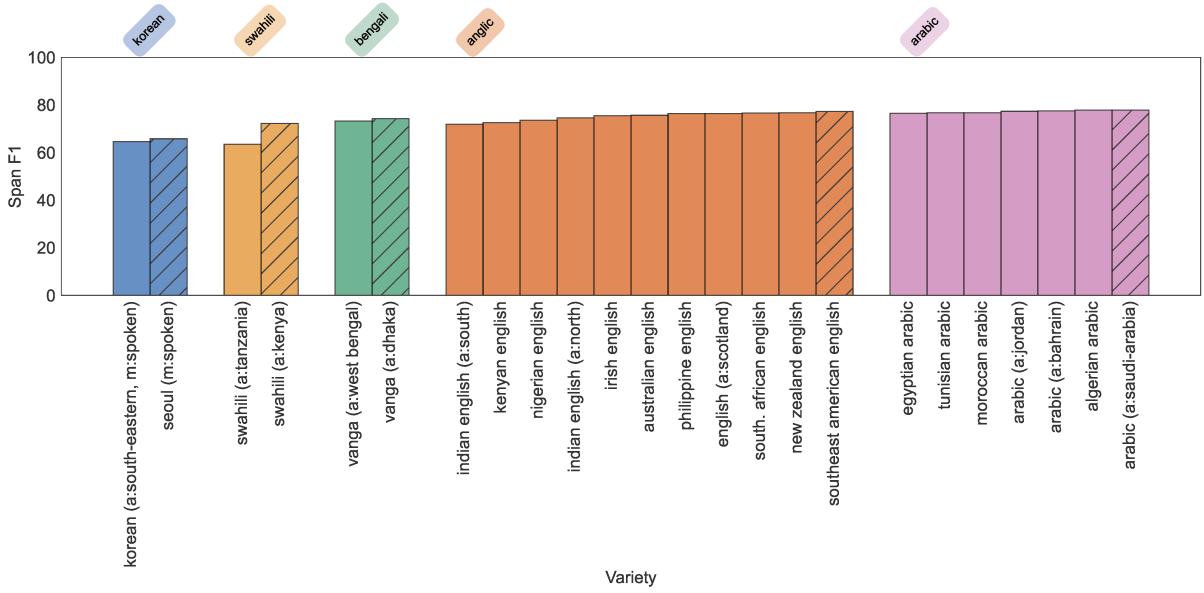


(c) Topic Classification

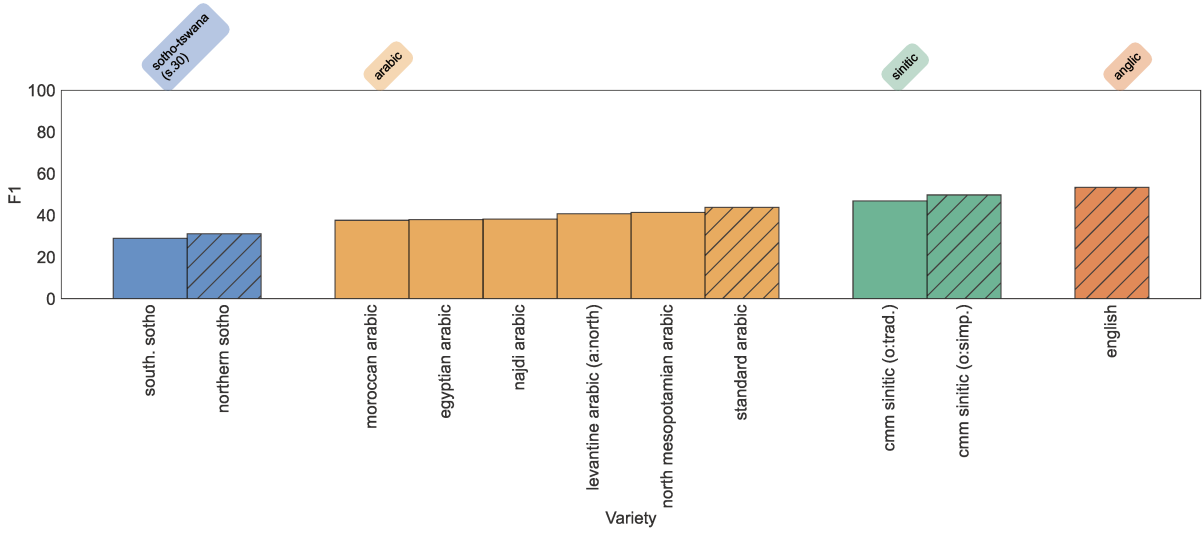
Figure 7: Task specific plot of maximum obtainable score for all varieties. The yellow-shaded region represents language clusters having no varieties seen during mBERT pertaining. The bars with colored stripes represent the standard variety of a cluster



(a) Natural Language Inference

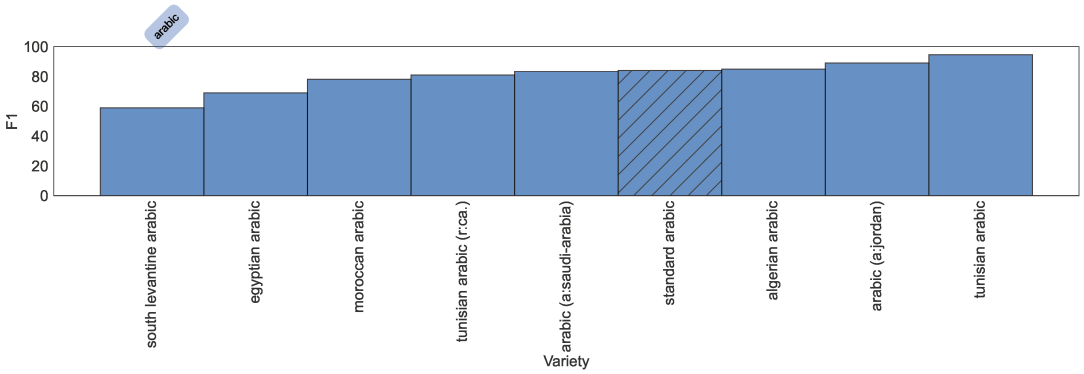


(b) Extractive Question Answering

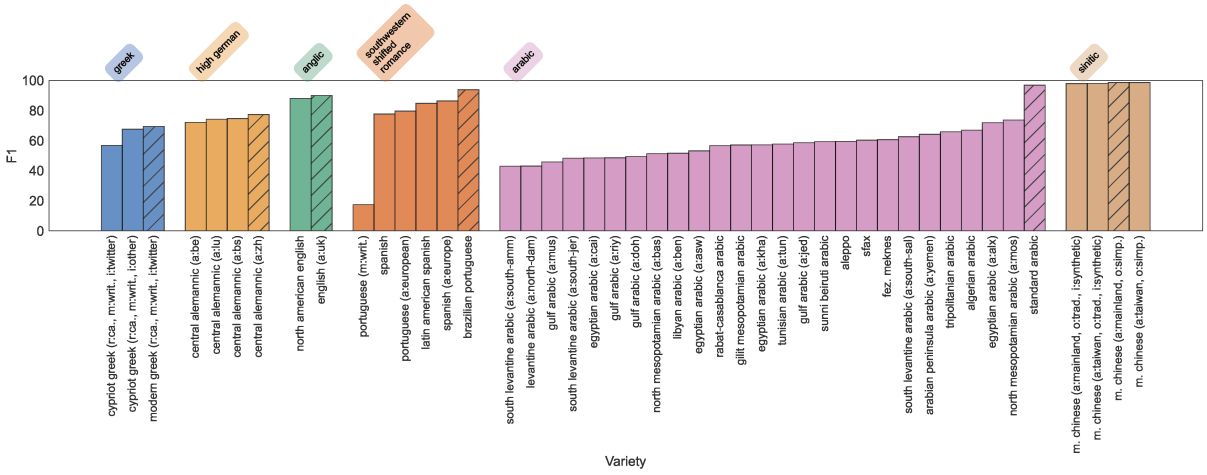


(c) Multiple Choice Machine Reading Comprehension

Figure 8: Task specific plot of maximum obtainable Linguistic Utility for all varieties. The yellow-shaded region represents language clusters having no varieties seen during mBERT pertaining. The bars with colored stripes represent the standard variety of a cluster. The dialect with the Rawlsian score in each cluster is that with the leftmost bar.

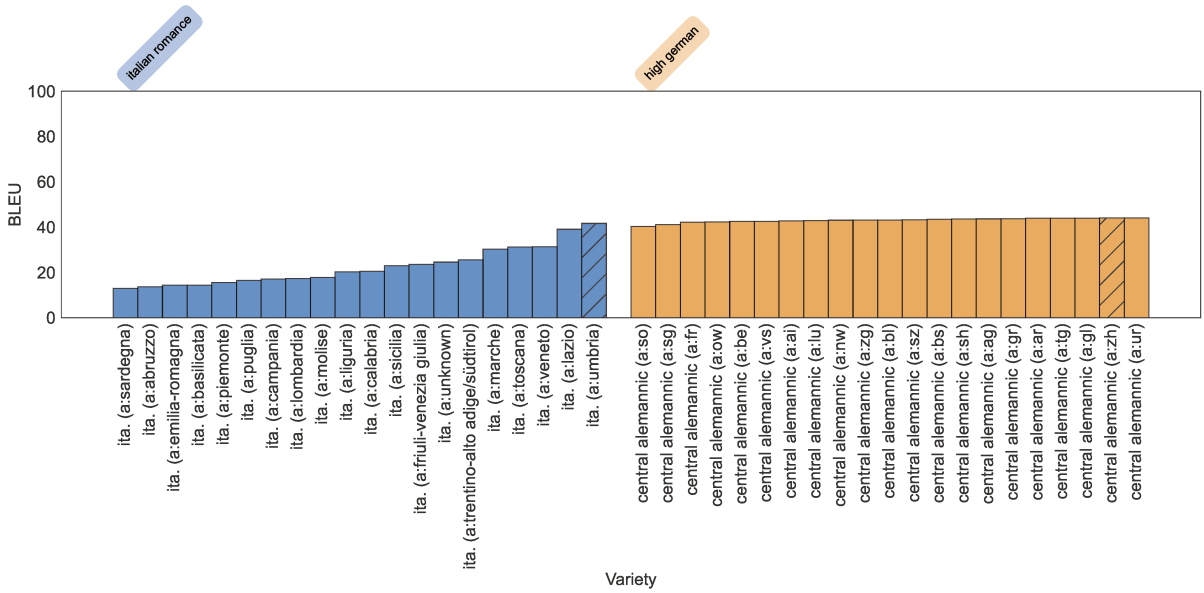


(a) Sentiment Analysis

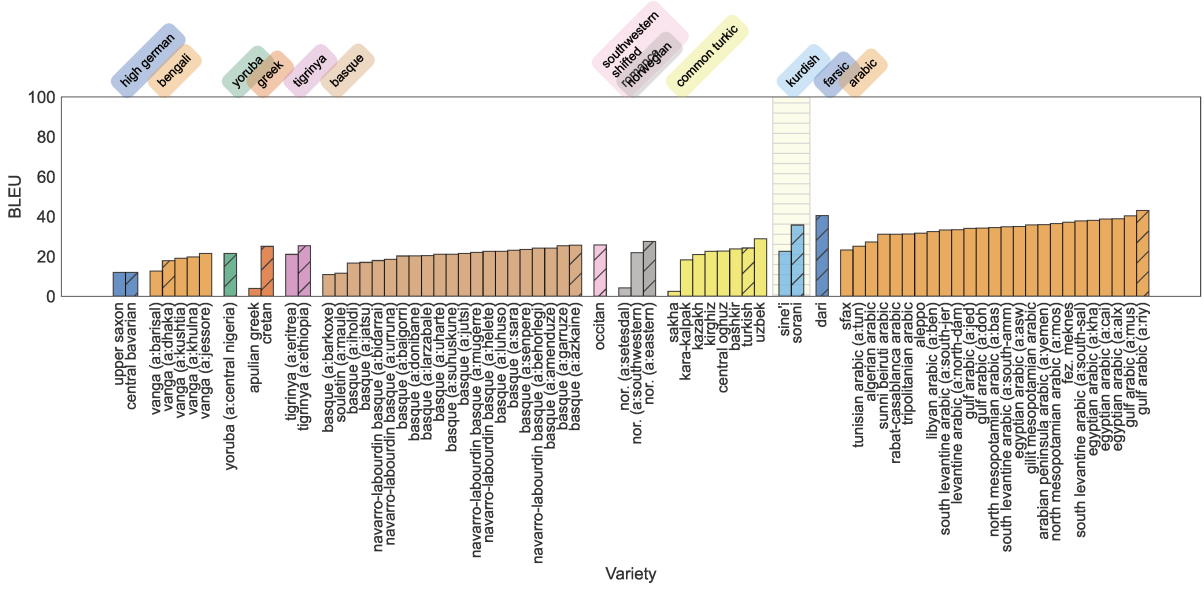


(b) Dialect Identification

Figure 9: Task specific plot of maximum obtainable Linguistic Utility for all varieties. The yellow-shaded region represents language clusters having no varieties seen during mBERT pertaining. The bars with colored stripes represent the standard variety of a cluster. The dialect with the Rawlsian score in each cluster is that with the leftmost bar.

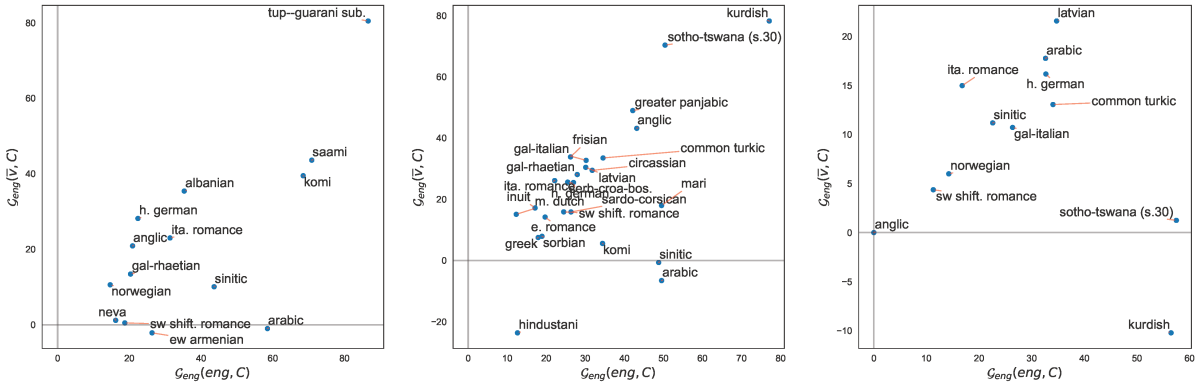


(a) Machine Translation (Region level aggregation)



(b) Machine Translation (Variety level)

Figure 10: Task specific plot of maximum obtainable Linguistic Utility for all varieties. The yellow-shaded region represents language clusters having no varieties seen during mBERT pertaining. The bars with colored stripes represent the standard variety of a cluster. The dialect with the Rawlsian score in each cluster is that with the leftmost bar.



(a) Parts-of-speech (POS) tagging (b) Named entity recognition (NER) (c) Natural language inference (NLI)

Figure 11: Dialectal Gap visualization for language clusters utilizing zero-shot cross-lingual transfer from Standard English. The x-axis is for cluster gap while comparing against Standard English variety. Values far from zero have a larger performance gap from English. The y-axis is for aggregated cluster gap while comparing against standard cluster variety and values far from zero have larger within cluster gap. Ideally, we want both of them to be close to zero.

cluster	variety	Dataset-code	dataset	Zeroshot (mBERT)	Zeroshot (XLM-R)	Finetune (mBERT)	Finetune (XLM-R)
albanian	albanian	UD_Albanian-TSA	ud	75.80	84.41	-	-
	gheg albanian	UD_Gheg-GPS	ud	48.96	55.84	-	-
anglic	english	UD_English-EWT	ud	96.41	97.16	96.41	97.16
	singlish	singlish	ud	76.27	77.55	87.38	87.96
arabic	south levantine arabic	UD_South_Levantine_Arabic-MADAR	ud	51.99	61.84	-	-
	standard arabic	UD_Arabic-PADT	ud	39.74	56.67	95.72	96.11
	gulf arabic	dar-glf	noisy	38.84	50.12	-	-
	egyptian arabic	dar-egy	noisy	36.14	51.39	-	-
	levantine arabic	dar-lev	noisy	32.66	43.76	-	-
	north african arabic	UD_Maghrebi_Arabic_French-Arabizi	ud	28.30	26.01	67.89	59.29
eastern-western armenian	eastern armenian	UD_Armenian-ArmTDP	ud	71.78	82.63	91.31	92.36
	western armenian	UD_Western_Armenian-ArmTDP	ud	70.27	75.31	94.86	95.14
gallo-italian	ligurian	UD_Ligurian-GLT	ud	58.90	52.78	13.16	5.09
gallo-rhaetian	french	UD_French-ParTUT	ud	84.36	85.47	96.88	96.42
	french (a:paris)	UD_French-ParisStories	ud	81.37	82.77	96.70	96.76
	old french (842-ca. 1400)	UD_Old_French-SRCMF	ud	64.70	59.41	95.64	95.64
high german	german	UD_German-LIT	ud	87.08	88.36	-	-
	central alemannic (a:zh)	UD_Swiss_German-UZH	ud	62.56	47.18	-	-
italian romance	italian	UD_Italian-PUD	ud	81.09	83.12	-	-
	italian (r:formal, m:written, i:essay)	UD_Italian-MarkIT	ud	80.00	81.87	93.35	92.70
	italian (r:casual, m:written, i:tweet)	UD_Italian-PoSTWITA	ud	73.71	76.45	95.80	96.83
	continental southern italian	UD_Neapolitan-RB	ud	30.00	57.14	-	-
komi	komi-zyrian (m:spoken)	UD_Komi_Zyrian-IKDP	ud	41.25	46.66	-	-
	komi-permyak	UD_Komi_Permyak-UH	ud	29.52	43.67	-	-
	komi-zyrian (m:written)	UD_Komi_Zyrian-Lattice	ud	20.40	35.12	-	-
neva	finnish	UD_Finnish-TDT	ud	81.29	86.21	96.05	97.76
	estonian	UD_Estonian-EDT	ud	80.34	85.17	96.49	97.20
	hame (a:tavastian)	murre-HAAM	noisy	55.63	70.08	-	-
	central and north pohjanmaa (a:ostrobothnian)	murre-POH	noisy	55.09	69.68	-	-
	savo (a:savonian)	murre-SAV	noisy	54.70	69.42	-	-
	southeastern finnish (a:south-east)	murre-KAA	noisy	51.68	68.71	-	-
	southwestern finnish (a:south-west)	murre-LVA	noisy	49.80	68.54	-	-
	neva (a:south-west trans)	murre-LOU	noisy	42.57	61.72	-	-
norwegian	norwegian bokmål (m:written)	UD_Norwegian-Bokmaal	ud	88.53	89.55	98.19	98.67
	norwegian nynorsk (m:written)	UD_Norwegian-Nynorsk	ud	85.06	85.81	97.83	98.32
	norwegian nynorsk (m:written, i:old)	UD_Norwegian-NynorskLIA	ud	73.25	79.29	95.47	95.72
saami	north saami	UD_North_Sami-Giella	ud	35.92	32.13	78.89	71.50
	skolt saami	UD_Skolt_Sami-Giellagas	ud	20.26	34.15	-	-
sabellic	umbrian	UD_Umbrian-IKUVINA	ud	11.90	5.44	-	-
sinitic	classical-middle-modern sinitic (a:hongkong, o:traditional)	UD_Chinese-HK	ud	68.99	35.49	-	-
	classical-middle-modern sinitic (o:simplified)	UD_Chinese-GSDSimp	ud	58.26	30.92	94.72	95.02
	classical chinese	UD_Classical_Chinese-Kyoto	ud	35.80	20.85	89.62	89.49
southwestern shifted romance	portuguese (a:european)	UD_Portuguese-PUD	ud	80.08	81.38	-	-
	brazilian portuguese	UD_Portuguese-GSD	ud	78.63	80.12	98.19	98.50
	portuguese (i:mix)	UD_Portuguese-Bosque	ud	78.48	79.85	97.71	97.81
	occitan	ROci	noisy	76.84	65.80	-	-
	portuguese (m:written)	UD_Portuguese-CINTIL	ud	76.19	78.76	97.67	97.87
tupi-guarani subgroup i.a	mbyá guaraní (a:paraguay)	UD_Mbya_Guarani-Thomas	ud	27.89	28.77	-	-
	old guarani	UD_Guarani-OldTuDeT	ud	8.96	10.30	-	-
	mbyá guaraní (a:brazil)	UD_Mbya_Guarani-Dooley	ud	1.94	0.59	-	-
west low german	west low german	UD_Low_Saxon-LSDC	ud	69.65	54.93	-	-

Table 11: Parts-of-speech evaluation report comprising zeroshot score and in-language finetuning. We report F1 as evaluation score. Zeroshot scores are evaluated using model finetuned on Standerd English. If training data is not available, we skip those languages (mentioned as '-') for in-language finetuning.

Table 12: Named entity recognition (NER) evaluation report comprising zeroshot score and in-group finetuning. We report F1 as evaluation score. Zeroshot scores are evaluated using model finetuned on Standerd English. If training data is not available, we skip those languages (mentioned as '-') for in-language finetuning.

cluster	variety	target-code	source	dataset	support	Zeroshot (mBERT)	Zeroshot (XLM-R)	Finetune (mBERT)	Finetune (XLM-R)
anglic	english (o:controlled)	simple	en	wikiann	1000	89.07	86.03	89.07	86.03
	english	en	en	wikiann	10000	84.15	82.11	84.15	82.11
	old english (ca. 450-1100)	ang	en	wikiann	100	54.41	55.94	54.41	55.94
	jamaican creole english	jam	en	wikiann	0	0.00	0.00	0.00	0.00
arabic	egyptian arabic	arz	ar	wikiann	100	43.82	50.58	67.77	73.33
	standard arabic	ar	ar	wikiann	10000	41.12	41.76	89.10	87.85
circassian	adyghe	ady	en	wikiann	693	67.33	54.03	-	-
	kabardian	kbd	en	wikiann	1482	47.51	34.79	-	-
common turkic	turkish	tr	tr	wikiann	10000	73.56	75.71	92.90	91.80
	north azerbaijani	az	az	wikiann	1000	67.31	61.01	89.17	88.26
	crimean tatar	crh	tr	wikiann	100	47.81	40.57	57.99	52.67
	south azerbaijani	azb	az	wikiann	2567	31.67	11.16	30.22	22.14
eastern romance	romanian	ro	ro	wikiann	10000	74.62	70.82	94.17	93.47
	aromanian	roa-rup	el	wikiann	732	64.78	62.66	68.92	63.33
	moldavian	mo	ro	wikiann	345	63.31	56.53	60.60	64.56
frisian	western frisian	fy	nl	wikiann	1000	80.17	71.96	81.06	78.21
	ems-weser frisian	stq	nl	wikiann	1085	59.45	57.47	64.55	55.15
	northern frisian	frr	nl	wikiann	100	46.67	46.22	54.78	52.63
gallo-italian	piemontese	pms	it	wikiann	100	79.53	71.10	87.88	78.72
	lombard	lmo	it	wikiann	100	72.49	68.17	79.37	78.66
	venetian	vec	it	wikiann	100	62.71	55.20	73.73	67.51
	ligurian	lij	it	wikiann	100	45.36	34.75	56.51	47.79
	emiliano-romagnolo	eml	it	wikiann	100	33.55	33.23	42.80	45.45
gallo-rhaetian	french	fr	fr	wikiann	10000	79.16	76.52	90.96	89.00
	anglo-norman	nrm	fr	wikiann	1281	66.78	88.56	71.98	71.92
	arpitan	frp	it	wikiann	2358	63.30	63.67	68.38	71.13
	romansh	rm	it	wikiann	100	56.88	55.19	69.69	67.58
	friulian	fur	it	wikiann	100	51.41	50.75	64.12	56.30
	walloon	wa	fr	wikiann	100	46.33	41.27	45.19	42.50
greater panjabic	western panjabi	pnb	pa	wikiann	100	64.46	53.78	17.82	0.00
	eastern panjabi	pa	pa	wikiann	100	32.90	45.25	30.63	0.00
greek	modern greek	el	el	wikiann	10000	71.76	72.96	91.18	90.68
	pontic	pnt	el	wikiann	291	66.37	69.79	68.45	72.58
high german	german	de	de	wikiann	10000	79.08	75.67	89.97	87.80
	central alemannic	als	de	wikiann	100	75.36	65.15	84.84	79.02
	luxemburgish	lb	nl	wikiann	1000	71.61	49.22	79.22	58.71
	limburgan	li	nl	wikiann	100	63.03	63.72	78.03	73.15
	bavarian	bar	de	wikiann	100	56.62	55.96	76.36	68.84
	kölsch	ksh	nl	wikiann	100	54.80	39.42	62.50	48.51
	pfaelzisch-lothringisch	pfl	de	wikiann	1092	49.40	47.19	59.12	50.14
	pennsylvania german	pdc	de	wikiann	100	41.76	41.79	39.39	38.76
hindustani	fiji hindi	hif	hi	wikiann	715	81.29	79.67	74.28	85.15
	hindi	hi	hi	wikiann	1000	65.74	65.77	88.11	85.83
inuit	alaskan inupiaq	ik	en	wikiann	431	76.27	70.56	-	-
	kalaallisut	kl	en	wikiann	1403	63.20	60.20	-	-
italian romance	italian	it	it	wikiann	10000	81.44	78.35	92.21	90.22
	italian romance (a:apulia, m:spoken, i:tarantino)	roa-tara	it	wikiann	3811	62.97	66.02	60.34	64.45
	sicilian	scn	it	wikiann	100	60.26	54.46	72.25	60.87
	continental southern italian	nap	it	wikiann	100	57.35	54.81	61.48	58.33
komi	komi	kv	en	wikiann	2464	56.78	41.78	-	-
	komi-permyak	koi	en	wikiann	2798	53.62	47.80	-	-
kurdish	kurdish	ku	ku	wikiann	100	31.67	59.48	13.70	0.00
	central kurdish	ckb	ku	wikiann	1000	6.90	35.49	1.17	0.00
latvian	latvian	lv	lv	wikiann	10000	69.36	71.07	93.24	92.30
	east latvian	ltg	lv	wikiann	1036	48.27	48.86	50.10	49.95

Continued on next page

Table 12: Named entity recognition (NER) evaluation report comprising zeroshot score and in-group finetuning. We report F1 as evaluation score. Zeroshot scores are evaluated using model finetuned on Standerd English. If training data is not available, we skip those languages (mentioned as '-') for in-language finetuning.

cluster	variety	target-code	source	dataset	support	Zeroshot (mBERT)	Zeroshot (XLM-R)	Finetune (mBERT)	Finetune (XLM-R)
mari	eastern mari	mhr	mhr	wikiann	100	46.67	40.94	32.70	0.00
	western mari	mrj	mhr	wikiann	6036	38.29	58.46	5.51	0.00
modern dutch	dutch	nl	nl	wikiann	10000	82.03	80.42	91.96	90.51
	western flemish	vls	nl	wikiann	100	73.36	72.08	77.66	79.85
	zeeuws	zea	nl	wikiann	100	65.98	66.20	79.55	76.09
norwegian	norwegian nynorsk (m:written)	nynorsk	nynorsk	norwegian_ner	1511	87.58	87.86	87.58	87.86
	norwegian (m:written, i:samnorsk)	samnorsk	samnorsk	norwegian_ner	3450	86.96	90.55	86.96	90.55
	norwegian bokmål (m:written)	bokmaal	bokmaal	norwegian_ner	1939	85.82	90.54	85.82	90.54
sardo-corsican	sardinian	sc	it	wikiann	917	67.32	65.26	80.87	81.17
	corsican	co	it	wikiann	100	56.65	56.41	70.59	66.06
serbian-croatian-bosnian	croatian standard	hr	hr	wikiann	10000	77.59	78.12	92.40	91.02
	bosnian standard	bs	hr	wikiann	1000	69.93	74.88	87.50	88.86
	serbian standard	sr	hr	wikiann	10000	64.38	60.71	63.68	65.86
	serbian-croatian-bosnian	sh	hr	wikiann	10000	38.92	69.14	85.05	85.43
sinitic	wu chinese	wuu	zh	wikiann	100	71.89	35.80	74.15	63.73
	min nan chinese	zh-min-nan	zh	wikiann	100	44.68	47.30	21.40	15.08
	cantonese	zh-yue	zh	wikiann	10000	43.73	26.55	76.19	71.97
	mandarin chinese	zh	zh	wikiann	10000	42.86	24.71	81.12	77.22
	classical chinese	zh-classical	zh	wikiann	100	28.03	16.78	67.28	62.39
	hakka chinese	hak	zh	wikiann	100	27.43	31.84	36.55	40.00
sorbian	lower sorbian	dsb	hsb	wikiann	862	71.04	68.80	3.03	0.00
	upper sorbian	hsb	hsb	wikiann	100	65.44	65.48	38.20	0.00
sotho-tswana (s.30)	southern sotho	st	en	wikiann	339	64.36	69.26	-	-
	northern sotho	nso	en	wikiann	720	19.08	29.66	-	-
southwestern shifted romance	galician	gl	es	wikiann	10000	81.98	80.27	87.99	86.14
	spanish	es	es	wikiann	10000	72.80	70.73	92.17	90.32
	occitan	oc	it	wikiann	100	72.00	67.58	78.50	75.35
	mirandese	mwl	es	wikiann	100	46.20	44.07	49.29	42.81
	extremaduran	ext	es	wikiann	100	44.83	38.33	61.82	40.00
west low german	west low german	nds	de	wikiann	100	80.29	66.44	79.40	70.99

Table 20: Zero-shot results for Machine Translation. We evaluate NLLB_600m and NLLB_1_3bn by translating each dialectal variety to English. For all languages without reference translations in English, we evaluate dialectal translations using the standard language's translation as the reference. For varieties with parallel data like Yoruba, Turkish, Farsi, Tigrinya and Bengali we use the English reference provided in the dataset.

language_group	variety	NLLB_600m-bleu	NLLB_1_3bn-bleu
arabic	gulf arabic (a:riy)	43.07	43.07
	gulf arabic (a:mus)	40.32	40.32
	egyptian arabic (a:alx)	38.90	38.90
	egyptian arabic (a:cai)	38.71	38.71
	egyptian arabic (a:kha)	38.10	38.10
	south levantine arabic (a:south-sal)	37.79	37.79
	fez. meknes	37.19	37.19
	north mesopotamian arabic (a:mos)	36.47	36.47
	arabian peninsula arabic (a:yemen)	35.93	35.93
	gilit mesopotamian arabic	35.78	35.78
	egyptian arabic (a:asw)	35.02	35.02
	south levantine arabic (a:south-amm)	34.88	34.88
	north mesopotamian arabic (a:bas)	34.41	34.41
	gulf arabic (a:doh)	34.20	34.20
	gulf arabic (a:jed)	34.10	34.10
	levantine arabic (a:north-dam)	33.36	33.36

Continued on next page

Table 20: Machine Translation

language_group	variety	NLLB_600m-bleu	NLLB_1_3bn-bleu
	south levantine arabic (a:south-jer)	33.27	33.27
	libyan arabic (a:ben)	32.45	32.45
	aleppo	31.64	31.64
	tripolitanian arabic	31.21	31.21
	rabat-casablanca arabic	31.17	31.17
	sunni beiruti arabic	31.07	31.07
	algerian arabic	27.25	27.25
	tunisian arabic (a:tun)	25.10	25.10
	sfax	23.24	23.24
basque	basque (a:azkaine)	25.61	22.58
	navarro-labourdin basque (a:helete)	22.53	22.03
	basque (a:garruze)	25.40	22.00
	navarro-labourdin basque (a:behorlegi)	24.19	21.35
	basque (a:luhuso)	22.58	20.95
	basque (a:amenduze)	24.21	20.86
	basque (a:senpere)	23.47	20.35
	basque (a:sara)	23.04	20.25
	basque (a:jutsi)	21.46	19.99
	navarro-labourdin basque (a:mugerre)	22.00	19.71
	basque (a:baigorri)	20.21	18.80
	basque (a:uharte)	21.10	18.57
	basque (a:suhuskune)	21.10	18.57
	basque (a:donibane)	20.28	18.03
	basque (a:larzabale)	20.37	17.88
	navarro-labourdin basque (a:bidarrai)	17.94	16.90
	navarro-labourdin basque (a:urruna)	18.50	16.25
	basque (a:iholdi)	16.68	15.01
	basque (a:jatsu)	17.01	14.56
	basque (a:barkoxe)	10.93	10.82
	souletin (a:maule)	11.55	10.36
bengali	vanga (a:jessore)	20.71	21.44
	vanga (a:khulna)	18.96	19.73
	vanga (a:kushtia)	17.36	19.12
	vanga (a:dhaka)	17.18	17.85
	vanga (a:barisal)	11.33	12.68
common turkic	uzbek	24.68	28.82
	turkish	22.24	24.22
	bashkir	20.88	23.78
	central oghuz	18.82	22.65
	kirghiz	18.68	22.53
	kazakh	19.56	20.96
	kara-kalpak	14.31	18.23
	sakha	2.53	2.43
farsic	dari	37.02	40.49
greek	cretan	25.03	21.34
	apulian greek	3.92	3.97
high german	central bavarian	9.49	12.01
	upper saxon	11.99	8.63
kurdish	sorani	35.74	34.88
	sine'i	22.56	22.60
norwegian	norwegian (a:eastern)	24.00	27.49
	norwegian (a:southwestern)	21.88	15.81
	norwegian (a:setesdal)	2.24	4.23
southwestern shifted romance	occitan	20.73	25.74
tigrinya	tigrinya (a:ethiopia)	22.40	25.36
	tigrinya (a:eritrea)	18.64	21.02
yoruba	yoruba (a:central nigeria)	21.46	18.10
	central alemannic (a:zh)	43.71	44.06
	central alemannic (a:gl)	43.90	43.96

Continued on next page

high german

Table 20: Machine Translation

language_group	variety	NLLB_600m-bleu	NLLB_1_3bn-bleu
	central alemannic (a:tg)	43.94	43.62
	central alemannic (a:bs)	42.94	43.48
	central alemannic (a:gr)	43.67	43.31
	central alemannic (a:ag)	43.62	43.26
	central alemannic (a:sh)	43.62	43.01
	central alemannic (a:zg)	43.05	42.95
	central alemannic (a:ai)	42.27	42.72
	central alemannic (a:ar)	43.90	42.67
	central alemannic (a:sz)	43.24	42.57
	central alemannic (a:vs)	41.69	42.52
	central alemannic (a:ur)	44.08	42.41
	central alemannic (a:be)	42.51	42.36
	central alemannic (a:lu)	42.86	42.22
	central alemannic (a:nw)	43.04	41.89
	central alemannic (a:bl)	43.11	41.68
	central alemannic (a:fr)	42.13	41.05
	central alemannic (a:ow)	42.29	40.89
	central alemannic (a:sg)	41.11	40.55
	central alemannic (a:so)	40.27	39.39
italian romance	italian (a:umbria)	40.73	41.62
	italian (a:lazio)	39.03	32.23
	italian (a:veneto)	29.38	31.27
	italian (a:toscana)	28.90	31.24
	italian (a:marche)	26.29	30.22
	italian (a:trentino-alto adige/südtirol)	23.40	25.56
	italian (a:unknown)	22.25	24.54
	italian (a:friuli-venezia giulia)	18.62	23.51
	italian (a:sicilia)	22.91	23.00
	italian (a:calabria)	20.11	20.48
	italian (a:liguria)	17.98	20.22
	italian (a:lombardia)	16.23	17.32
	italian (a:campania)	17.03	17.03
	italian (a:molise)	17.73	15.98
	italian (a:puglia)	16.50	15.64
	italian (a:piemonte)	14.48	15.54
	italian (a:basilicata)	13.76	14.36
	italian (a:emilia-romagna)	13.86	14.35
	italian (a:sardegna)	11.67	12.98
	italian (a:abruzzo)	13.64	11.02

F Highest performing and lowest performing varieties

G Low-resource variety performing better in zero-shot NER

H Cluster level Result Summaries with *Demographic Utility* and Standard Deviation report

I In-Context Learning Details

I.1 Prompts

We adapt the prompts from Super-NaturalInstructions (Wang et al., 2022) for our in-context learning experiments.

Sentiment Analysis. For sentiment analysis we provide 4 few-shot examples in the prompt. The prompt template is given below:

```
In this task, you are given a piece of text. Your task is to classify the sentiment of the text based
on its content.

Sentence: <Sentence Example 1>
Label: <Label for Example 1, Positive, negative, neutral>

...

Sentence: <Sentence Example k>
```

cluster	variety	target-code	src	mBERT (acc)	XLM-R (acc)	mBERT (F1)	XLM-R (F1)
anglic	english	eng_Latn	eng_Latn	81.90	83.35	81.95	83.43
arabic	standard arabic	arb_Arab	eng_Latn	65.57	73.83	65.57	73.85
	najdi arabic	ars_Arab	eng_Latn	59.42	69.02	59.14	68.94
	ta'izzi-adeni arabic	acq_Arab	eng_Latn	58.84	68.72	58.64	68.62
	moroccan arabic	ary_Arab	eng_Latn	54.65	58.30	54.61	58.14
	egyptian arabic	arz_Arab	eng_Latn	54.53	65.87	53.86	65.70
	south levantine arabic	ajp_Arab	eng_Latn	54.09	64.13	53.42	63.81
	north mesopotamian arabic	acm_Arab	eng_Latn	52.95	58.94	52.84	58.75
	levantine arabic (a:north)	apc_Arab	eng_Latn	52.14	61.74	51.40	61.31
	tunisian arabic	aeb_Arab	eng_Latn	47.54	50.18	47.42	50.20
common turkic	north azerbaijani	azj_Latn	eng_Latn	59.76	73.17	59.20	73.17
	central oghuz (m:spoken)	tur_Latn	eng_Latn	59.14	74.47	58.37	74.52
	south azerbaijani	azb_Arab	eng_Latn	46.05	41.82	44.58	39.24
gallo-italian	venetian	vec_Latn	eng_Latn	65.15	68.52	64.99	68.55
	lombard	lmo_Latn	eng_Latn	59.38	56.39	59.34	56.16
	ligurian	lij_Latn	eng_Latn	57.82	57.70	56.70	57.16
high german	luxemburgish	ltz_Latn	eng_Latn	60.34	47.33	60.01	46.21
	limburgan	lim_Latn	eng_Latn	50.80	59.88	50.31	59.75
italian romance	italian	ita_Latn	eng_Latn	73.71	78.08	73.71	78.19
	sicilian	scn_Latn	eng_Latn	62.69	56.17	62.66	55.82
kurdish	central kurdish	ckb_Arab	eng_Latn	40.98	44.93	37.40	39.59
	northern kurdish	kmr_Latn	eng_Latn	39.10	63.45	33.93	63.26
latvian	latvian	lvs_Latn	eng_Latn	60.14	73.57	59.95	73.63
	east latvian	ltg_Latn	eng_Latn	48.62	54.19	47.02	53.54
norwegian	norwegian bokmål (m:written)	nob_Latn	eng_Latn	72.44	79.44	72.45	79.51
	norwegian nynorsk (m:written)	nno_Latn	eng_Latn	68.08	71.16	68.10	71.06
sinitic	classical-middle-modern sinitic (o:simplified)	zho_Hans	eng_Latn	68.52	72.61	68.54	72.57
	classical-middle-modern sinitic (o:traditional)	zho_Hant	eng_Latn	61.72	64.89	61.48	64.49
	cantonese	yue_Hant	eng_Latn	60.44	68.02	60.27	67.41
sotho-tswana (s.30)	northern sotho	nso_Latn	eng_Latn	39.38	40.18	35.06	35.98
	southern sotho	sot_Latn	eng_Latn	39.36	39.30	34.62	34.16
southwestern shifted romance	spanish	spa_Latn	eng_Latn	75.13	79.00	75.15	79.09
	portuguese (a:european)	por_Latn	eng_Latn	73.73	79.18	73.73	79.22
	galician	glg_Latn	eng_Latn	73.39	78.48	73.39	78.55
	occitan	oci_Latn	eng_Latn	68.48	63.09	68.47	62.96

Table 13: Natural language inference (NLI) evaluation report using zeroshot cross-lingual transfer from Standard English. We report F1 as evaluation score. NLI uses F1 as evaluation matric. We prepare a translate-train dataset to perform this evaluation.

cluster	variety	target-code	count	Finetune (mBERT)	Finetune (XLM-R)	Zeroshot (mBERT)	Zeroshot (XLM-R)
anglic	irish english	english–irl	494	73.00	67.98	68.62	62.45
	southeast american english	english–usa	494	73.51	67.95	68.56	62.97
	new zealand english	english–nzl	494	73.62	68.50	68.21	63.05
	southern african english	english–zaf	494	73.50	68.06	68.13	63.19
	english (a:scotland)	english–gbr	494	73.30	67.57	67.90	62.49
	australian english	english–aus	494	73.23	68.04	67.71	62.33
	nigerian english	english–nga	494	72.53	66.56	67.39	62.37
	philippine english	english–phl	494	73.75	67.29	67.35	61.54
	indian english (a:south)	english–ind_s	494	70.96	66.04	65.43	61.72
	kenyan english	english–kenya	494	70.63	65.64	65.30	60.28
	indian english (a:north)	english–ind_n	494	70.28	65.32	64.46	60.84
arabic	moroccan arabic	arabic–mar	324	70.94	65.14	50.94	49.56
	tunisian arabic	arabic–tun	324	71.36	65.31	50.86	50.01
	arabic (a:jordan)	arabic–jor	324	70.60	65.77	50.81	49.56
	arabic (a:bahrain)	arabic–bhr	324	70.96	65.86	49.88	50.36
	arabic (a:saudi-arabia)	arabic–sau	324	70.96	65.58	49.29	49.63
	egyptian arabic	arabic–egy	324	70.06	65.34	48.72	48.71
	algerian arabic	arabic–dza	324	68.63	64.71	48.71	48.65
bengali	bengali (a:west bengal)	bengali–ind	107	67.64	70.89	28.06	38.24
	bengali (a:dhaka)	bengali–dhaka	107	68.98	66.84	27.17	37.32
korean	seoul (m:spoken)	korean–korn	60	9.60	28.08	6.89	20.24
	korean (a:south-eastern, m:spoken)	korean–kors	60	9.27	26.96	6.89	19.66
swahili	swahili (a:kenya)	swahili–kenya	1000	72.06	71.90	46.20	46.22
	swahili (a:tanzania)	swahili–tanzania	1000	71.08	70.08	45.11	46.00

Table 14: Extractive dialectal question answering evaluation on SD-QA development set. We report span F1 as evaluation score. Zeroshot scores are evaluated using model finetuned on Standerd English whereas, we use combined training set for supervised finetuning

cluster	variety	target-code	count	Finetune (mBERT)	Finetune (XLM-R)	Zeroshot (mBERT)	Zeroshot (XLM-R)	ICL Mistral
anglic	english (a:scotland)	english–gbr	440	76.38	70.34	71.82	63.15	70.18
	southern african english	english–zaf	440	76.66	71.18	71.49	63.87	71.14
	new zealand english	english–nzl	440	76.71	71.39	71.22	63.69	70.95
	australian english	english–aus	440	75.66	70.89	71.20	62.28	69.23
	southeast american english	english–usa	440	77.26	71.50	71.17	63.71	71.76
	irish english	english–irl	440	75.52	70.73	70.92	62.15	70.64
	philippine english	english–phl	440	76.37	70.64	70.47	62.22	70.11
	nigerian english	english–nga	440	73.61	68.33	69.10	61.27	68.10
	indian english (a:north)	english–ind_n	440	74.62	68.03	68.84	61.25	68.99
	kenyan english	english–kenya	440	72.59	66.68	68.72	58.64	64.91
	indian english (a:south)	english–ind_s	440	71.93	66.88	66.49	60.36	65.13
arabic	arabic (a:bahrain)	arabic–bhr	921	77.52	72.11	53.25	53.28	55.33
	arabic (a:jordan)	arabic–jor	921	77.35	71.29	52.72	53.72	54.93
	arabic (a:saudi-arabia)	arabic–sau	921	77.88	72.11	52.72	53.24	55.66
	algerian arabic	arabic–dza	921	77.85	72.34	52.56	53.52	55.18
	tunisian arabic	arabic–tun	921	76.72	71.64	52.28	52.94	55.03
	moroccan arabic	arabic–mar	921	76.73	71.57	51.86	52.17	53.92
	egyptian arabic	arabic–egy	921	76.53	70.75	51.80	51.99	54.66
bengali	bengali (a:west bengal)	bengali–ind	113	68.62	73.27	32.30	36.39	50.06
	bengali (a:dhaka)	bengali–dhaka	113	67.37	74.24	31.79	35.52	51.73
korean	seoul (m:spoken)	korean–korn	276	10.15	31.91	7.26	19.62	65.79
	korean (a:south-eastern, m:spoken)	korean–kors	276	9.92	31.01	7.22	20.08	64.63
swahili	swahili (a:tanzania)	swahili–tanzania	472	63.54	62.30	38.24	39.38	55.70
	swahili (a:kenya)	swahili–kenya	472	72.25	70.53	37.97	41.59	58.71

Table 15: Extractive dialectal question answering evaluation on SD-QA test set. We report span F1 as evaluation score. Zeroshot scores are evaluated using model finetuned on Standerd English whereas, we use combined training set for supervised finetuning

cluster	variety	dialect-code	support	precision (mBERT)	precision (XLM-R)	recall (mBERT)	recall (XLM-R)	F1 (mBERT)	F1 (XLM-R)
anglic	english (a:uk)	english:en-gb	249	98.10	89.57	83.13	71.59	90.00	79.58
	north american english	english:en-us	349	93.27	88.14	83.38	82.09	88.05	85.01
arabic	aleppo	arabic:ale	200	59.50	58.50	59.50	48.35	59.50	52.94
	algerian arabic	arabic:alg	272	79.00	65.50	58.09	62.68	66.95	64.06
	arabian peninsula arabic (a:yemen)	arabic:san	177	60.50	55.50	68.36	56.63	64.19	56.06
	egyptian arabic (a:alx)	arabic:alx	192	70.50	74.50	73.44	66.82	71.94	70.45
	egyptian arabic (a:asw)	arabic:asw	221	56.00	52.00	50.68	45.02	53.21	48.26
	egyptian arabic (a:cai)	arabic:cai	130	35.50	36.50	54.62	72.28	43.03	48.50
	egyptian arabic (a:kha)	arabic:kha	244	63.50	55.50	52.05	44.05	57.21	49.12
	fez. meknes	arabic:fes	196	60.00	59.50	61.22	56.40	60.61	57.91
	gilit mesopotamian arabic	arabic:bag	203	57.50	47.50	56.65	49.48	57.07	48.47
	gulf arabic (a:doh)	arabic:doh	205	50.00	43.50	48.78	45.55	49.38	44.50
	gulf arabic (a:jed)	arabic:jed	196	58.00	46.00	59.18	40.89	58.59	43.29
	gulf arabic (a:mus)	arabic:mus	178	38.50	49.50	43.26	42.67	40.74	45.83
	gulf arabic (a:riy)	arabic:riy	311	62.00	56.50	39.87	37.92	48.53	45.38
	levantine arabic (a:north-dam)	arabic:dam	148	37.50	24.50	50.68	42.98	43.10	31.21
	libyan arabic (a:ben)	arabic:ben	238	56.50	47.50	47.48	52.78	51.60	50.00
	north mesopotamian arabic (a:bas)	arabic:bas	186	49.50	39.00	53.23	49.68	51.30	43.70
	north mesopotamian arabic (a:mos)	arabic:mos	188	71.50	70.00	76.06	69.31	73.71	69.65
	rabat-casablanca arabic	arabic:rab	153	50.00	40.00	65.36	60.61	56.66	48.19
	sfax	arabic:sfx	215	62.50	64.50	58.14	48.13	60.24	55.13
	south levantine arabic (a:south-amm)	arabic:amm	177	40.50	35.00	45.76	35.53	42.97	35.26
	south levantine arabic (a:south-jer)	arabic:jer	202	48.50	47.00	48.02	40.34	48.26	43.42
	south levantine arabic (a:south-sal)	arabic:sal	167	46.00	66.50	55.09	59.11	50.14	62.59
	standard arabic	arabic:msa	244	75.00	98.00	61.48	95.61	67.57	96.79
	sunni beiruti arabic	arabic:bei	192	58.00	52.50	60.42	68.18	59.18	59.32
	tripolitanian arabic	arabic:tri	201	66.00	60.00	65.67	60.30	65.84	60.15
	tunisian arabic (a:tun)	arabic:tun	164	52.50	37.00	64.02	56.49	57.69	44.71
greek	cypriot greek (r:casual, m:written, i:other)	greek:cg_other	81	74.14	84.48	53.09	56.32	61.87	67.59
	cypriot greek (r:casual, m:written, i:twitter)	greek:cg_twitter	36	51.11	44.44	63.89	68.97	56.79	54.05
	modern greek (r:casual, m:written, i:twitter)	greek:smg_twitter	94	89.83	100.00	56.38	53.15	69.28	69.41
high german	central alemannic (a:be)	swiss-dialects:be	389	72.89	51.58	71.21	62.42	72.04	56.48
	central alemannic (a:bs)	swiss-dialects:bs	340	73.50	57.83	75.88	61.14	74.67	59.44
	central alemannic (a:lu)	swiss-dialects:lu	335	72.91	65.13	75.52	59.47	74.19	62.17
	central alemannic (a:zh)	swiss-dialects:zh	359	78.84	73.33	75.77	63.73	77.27	68.19
sinitic	mandarin chinese (a:mainland, o:simplified)	mandarin_simplified:m	986	97.90	96.10	99.29	90.66	98.59	93.30
	mandarin chinese (a:mainland, o:traditional, i:synthetic)	mandarin_traditional:m	977	96.80	92.10	99.08	95.74	97.93	93.88
	mandarin chinese (a:taiwan, o:simplified)	mandarin_simplified:t	1014	99.30	90.10	97.93	95.85	98.61	92.89
	mandarin chinese (a:taiwan, o:traditional, i:synthetic)	mandarin_traditional:t	1023	99.10	95.90	96.87	92.39	97.97	94.11
southwestern shifted romance	brazilian portuguese	portuguese:pt-br	627	96.94	92.35	90.91	84.98	93.83	88.51
	latin american spanish	spanish:es-ar	207	81.06	9.25	88.89	91.30	84.79	16.80
	portuguese (a:european)	portuguese:pt-pt	349	91.45	83.64	70.49	63.92	79.61	72.46
	portuguese (m:written)	portuguese:pt	15	9.70	0.00	86.67	0.00	17.45	0.00
	spanish	spanish:es	290	74.21	74.53	81.38	47.69	77.63	58.16
	spanish (a:europe)	spanish:es-es	492	90.99	83.33	82.11	78.89	86.32	81.05

Table 16: Dialect Identification evaluation using language cluster specific datasets. We finetune a classificaiton model using either mBERT or XLM-R and then evaluate on the test data.

cluster	variety	target-code	src	mBERT (acc)	XLM-R (acc)	mBERT (F1)	XLM-R (F1)
arabic	standard arabic	arb_Arab	arb_Arab	85.25	83.96	86.71	82.27
	ta'izzi-adeni arabic	acq_Arab	arb_Arab	84.96	82.05	86.44	81.98
	najdi arabic	ars_Arab	arb_Arab	84.80	84.39	87.41	83.33
	north mesopotamian arabic	acm_Arab	arb_Arab	82.97	80.95	84.77	80.36
	south levantine arabic	ajp_Arab	arb_Arab	81.82	80.16	84.16	79.05
	levantine arabic (a:north)	apc_Arab	arb_Arab	81.59	80.15	83.76	79.88
	egyptian arabic	arz_Arab	arb_Arab	81.02	76.38	84.43	81.03
	tunisian arabic	aeb_Arab	arb_Arab	79.45	72.88	83.97	77.33
	moroccan arabic	ary_Arab	arb_Arab	73.87	79.14	78.76	78.55
common turkic	north azerbaijani	azj_Latn	azj_Latn	80.46	79.87	82.00	79.55
	central oghuz (m:spoken)	tur_Latn	azj_Latn	79.10	84.41	80.61	79.51
	south azerbaijani	azb_Arab	azj_Latn	65.90	67.08	69.71	68.37
gallo-italian	venetian	vec_Latn	ita_Latn	76.72	70.68	75.07	74.28
	lombard	lmo_Latn	ita_Latn	69.92	59.90	70.65	64.56
	ligurian	lij_Latn	lij_Latn	66.81	63.42	74.03	57.78
high german	luxemburgish	ltz_Latn	nld_Latn	74.74	58.50	77.86	64.83
	limburgan	lim_Latn	nld_Latn	71.09	65.83	71.12	65.73
italian romance	italian	ita_Latn	ita_Latn	87.67	84.92	86.68	85.83
	sicilian	scn_Latn	ita_Latn	75.22	59.71	72.70	59.47
kurdish	northern kurdish	kmr_Latn	ckb_Arab	33.23	68.21	10.45	5.71
	central kurdish	ckb_Arab	ckb_Arab	13.10	19.37	16.86	12.38
latvian	latvian	lvs_Latn	lvs_Latn	76.35	83.75	80.63	82.80
	east latvian	ltg_Latn	lvs_Latn	55.67	65.02	63.69	67.42
norwegian	norwegian nynorsk (m:written)	nno_Latn	nob_Latn	85.66	79.94	89.20	79.06
	norwegian bokmål (m:written)	nob_Latn	nob_Latn	83.81	82.90	83.82	84.14
sinitic	classical-middle-modern sinitic (o:traditional)	zho_Hant	zho_Hans	89.82	86.80	89.02	86.39
	cantonese	yue_Hant	zho_Hans	89.45	86.46	88.71	87.64
	classical-middle-modern sinitic (o:simplified)	zho_Hans	zho_Hans	88.74	86.38	88.86	89.15
sotho-tswana (s.30)	northern sotho	nso_Latn	nso_Latn	35.62	28.16	34.86	13.55
	southern sotho	sot_Latn	nso_Latn	32.55	32.31	39.93	19.08
southwestern shifted romance	portuguese (a:european)	por_Latn	spa_Latn	88.13	89.10	88.10	87.74
	galician	glg_Latn	spa_Latn	86.99	89.00	86.93	87.83
	spanish	spa_Latn	spa_Latn	86.74	85.93	84.87	86.55
	occitan	oci_Latn	lij_Latn	84.12	74.80	78.53	62.56

Table 17: Topic classification evaluation using SIB-200 language data with dialectal presence. We report span F1 as evaluation score. Zeroshot scores are evaluated using model finetuned on Standerd English whereas, we use in-group training for supervised finetuning

lang-group	variety	dialect	MBERT_Acc	MBERT_F1	XLMR_Acc	XLMR_F1	mistral7b
arabic	tunisian arabic	aeb_arab	94.55	94.56	94.61	94.62	73.3
	algerian arabic	arq_arab	84.98	85.00	84.70	84.69	76.0
	arabic (a:jordan)	jor_arab	82.96	82.90	89.07	89.00	82.2
	arabic (a:saudi-arabia)	sau_arab	81.38	65.97	83.40	67.66	79.8
	tunisian arabic (r:casual)	aeb_latn	80.95	65.65	79.80	63.70	62.3
	standard arabic	arb_arab	80.63	70.01	83.96	72.91	65.7
	moroccan arabic	ary_arab	78.08	61.50	77.41	55.55	58.4
	egyptian arabic	arz_arab	67.03	40.00	69.03	47.89	50.0
	south levantine arabic	ar_lb	58.38	34.63	58.90	32.72	0.0

Table 18: Sentiment Analysis results. In addition to, using mBERT and XLM-R as the base models, we also perform in-context learning to evaluate the performance of large language models (Mistral-7B).

lang-group	language	dialect	acc (mBERT)	acc (XLM-R)	mBERT (F1)	XLM-R (F1)
anglic	english	eng_Latn	52.22	53.56	51.97	53.44
arabic	standard arabic	arb_Arab	39.00	43.78	39.01	43.78
	levantine arabic (a:north)	apc_Arab	38.89	40.78	38.64	40.71
	north mesopotamian arabic	acm_Arab	38.11	41.33	37.99	41.35
	moroccan arabic	ary_Arab	37.00	37.67	36.94	37.61
	egyptian arabic	arz_Arab	36.22	38.00	36.21	37.98
	najdi arabic	ars_Arab	36.00	38.11	36.05	38.16
sinitic	classical-middle-modern sinitic (o:simplified)	zho_Hans	50.11	47.22	49.79	47.10
	classical-middle-modern sinitic (o:traditional)	zho_Hant	47.00	45.11	46.88	44.76
sotho-tswana (s.30)	northern sotho	nso_Latn	31.11	29.78	31.18	29.72
	southern sotho	sot_Latn	28.56	29.11	28.52	29.00

Table 19: Multiple-choice machine reading comprehension evaluation using Belebele dataset languages with dialectal presence. We report span F1 as evaluation score. We use combined finetuning using the aggregated training data provided with Belebele evaluationd data.

```
Label: <Label for Example k, Positive, negative, neutral>

Sentence: <Test Example Input>
Label:
```

Extractive Question Answering. We provide 2 few-shot examples i.e. $k = 2$ due to the long-form nature of text for this task.

```
This task is about writing a correct answer for the reading comprehension task. Based on the information provided in a given passage, you should identify the shortest continuous text span from the passage that serves as an answer to the given question. Avoid answers that are incorrect or provides incomplete justification for the question.

Passage: <Passage for Example 1>
Question: <Question for Example 1>
Answer: <Answer for Example 1>

...

Passage: <Passage for Example k>
Question: <Question for Example k>
Answer: <Answer for Example k>

Passage: <Passage for Test Example>
Question: <Question for Test Example>
Answer:
```

Task (Dataset)	Varieties with Highest Performance		Varieties with Lowest Performance	
DEP parsing (UD)	anglic/english*	italian romance/italian (r:formal, m:written, i:essay)*	tupi-guarani subgroup i.a/old guarani*	arabic/north african arabic
	albanian/albanian*	southwestern shifted romance/-portuguese (a:european)	komi/komi-zyrian (m:written)*	italian romance/continental southern italian
	gallo-rhaetian/french*	norwegian/norwegian nynorsk (m:written)*	saami/skolt saami*	komi/komi-zyrian (m:spoken)†
	norwegian/norwegian bokmål (m:written)*	southwestern shifted romance/-portuguese (i:mix)*	tupi-guarani subgroup i.a/mbyá guaraní (a:paraguay)	komi/komi-permyak†
	italian romance/italian*	southwestern shifted romance/brazilian portuguese*	tupi-guarani subgroup i.a/mbyá guaraní (a:brazil)*	saami/north saami*†
EQA (SD-QA-test)	anglic/english (a:scotland)	anglic/irish english	swahili/swahili (a:kenya)*	arabic/algerian arabic†
	anglic/southern african english	anglic/philippine english	bengali/vanga (a:west bengal)*†	arabic/tunisian arabic†
	anglic/new zealand english	anglic/nigerian english	bengali/vanga (a:dhaka)*†	arabic/moroccan arabic†
	anglic/australian english	anglic/indian english (a:north)	korean/seoul (m:spoken)*†	arabic/egyptian arabic*†
	anglic/southeast american english*	anglic/kenyan english	korean/korean (a:south-eastern, m:spoken)*†	swahili/swahili (a:tanzania)*
TC (SIB-200)	sinitic/classical-middle-modern sinitic (o:traditional)*†	italian romance/italian*	latvian/east latvian*	arabic/moroccan arabic†
	anglic/english*	southwestern shifted romance/-galician*	sotho-tswana (s.30)/northern sotho*	high german/limburgan
	sinitic/cantonese*†	southwestern shifted romance/spanish*	kurdish/northern kurdish†	gallo-italian/lombard
	sinitic/classical-middle-modern sinitic (o:simplified)*†	norwegian/norwegian nynorsk (m:written)*	sotho-tswana (s.30)/southern sotho*	gallo-italian/ligurian
	southwestern shifted romance/-portuguese (a:european)	arabic/standard arabic*†	kurdish/central kurdish†	common turkic/south azerbaijani†
MRC (Belebele)	anglic/english*	arabic/north mesopotamian arabic†	arabic/moroccan arabic†	sinitic/classical-middle-modern sinitic (o:simplified)*†
	sinitic/classical-middle-modern sinitic (o:simplified)*†	arabic/moroccan arabic†	arabic/egyptian arabic*†	sinitic/classical-middle-modern sinitic (o:traditional)*†
	sinitic/classical-middle-modern sinitic (o:traditional)*†	arabic/egyptian arabic*†	arabic/najdi arabic†	arabic/standard arabic*†
	arabic/standard arabic*†	arabic/najdi arabic†	sotho-tswana (s.30)/northern sotho*	arabic/levantine arabic (a:north)†
	arabic/levantine arabic (a:north)†	sotho-tswana (s.30)/northern sotho*	sotho-tswana (s.30)/southern sotho*	arabic/north mesopotamian arabic†
NER (Wikiann)	anglic/english (o:controlled)*	modern dutch/dutch*	sinitic/classical chinese†	mari/western mari†
	norwegian/norwegian nynorsk (m:written)*	southwestern shifted romance/-galician*	sinitic/hakka chinese*†	gallo-italian/emiliano-romagnolo
	norwegian/norwegian (m:written, i:samnorsk)	italian romance/italian*	sotho-tswana (s.30)/northern sotho*	greater panjabic/eastern panjabi*†
	norwegian/norwegian bokmål (m:written)*	hindustani/fiji hindi†	kurdish/central kurdish†	common turkic/south azerbaijani†
	anglic/english*	frisian/western frisian*†	anglic/jamaican creole english	kurdish/kurdish*†
NLI (XNLI-translate-test)	anglic/english*	norwegian/norwegian bokmål (m:written)*	common turkic/south azerbaijani†	arabic/north mesopotamian arabic†
	southwestern shifted romance/spanish*	sinitic/classical-middle-modern sinitic (o:simplified)*†	kurdish/central kurdish†	arabic/levantine arabic (a:north)†
	southwestern shifted romance/-portuguese (a:european)	southwestern shifted romance/occitan	sotho-tswana (s.30)/northern sotho*	high german/limburgan
	italian romance/italian*	norwegian/norwegian nynorsk (m:written)*	sotho-tswana (s.30)/southern sotho*	arabic/tunisian arabic†
	southwestern shifted romance/-galician*	arabic/standard arabic*†	kurdish/northern kurdish†	latvian/east latvian*
POS tagging (UD)	anglic/english*	gallo-rhaetian/french (a:paris)	tupi-guarani subgroup i.a/mbyá guaraní (a:paraguay)	sinitic/classical chinese†
	norwegian/norwegian bokmål (m:written)*	neva/finnish*	komi/komi-zyrian (m:written)*	arabic/levantine arabic†
	high german/german*	italian romance/italian*	saami/skolt saami*	italian romance/continental southern italian
	norwegian/norwegian nynorsk (m:written)*	neva/estonian*	tupi-guarani subgroup i.a/old guarani*	komi/komi-permyak†
	gallo-rhaetian/french*	southwestern shifted romance/-portuguese (a:european)	tupi-guarani subgroup i.a/mbyá guaraní (a:brazil)*	arabic/north african arabic
writing system (non-MT)	Latin (77.2%)		Latin (44.2%)	
dialect (non-MT)	Standard (7%)		Standard (4.5%)	

Table 21: Varieties with the highest and lowest performance (in terms of raw evaluation score) on various tasks in the zero-shot setting. On the left are the top 10. We find that most of these varieties are high-resource standard varieties, and only a few high-resource non-standard varieties. At right are the bottom 10, which are mainly low-resource, non-standard varieties. We use the following notations for variety type and writing system quantification: * marks standard varieties (for some clusters, here we consider multiple varieties as standard because of the substantial resource presence of both of the varieties; e.g. portuguese/european and portuguese/mix) and † notes mix and non-Latin writing system.

Cluster	High-resource variety	Script	Low-resource variety	Script
Hindustani	Hindi	Devanagari	Fiji hindi	Latin
Greater panjabic	Eastern panjabi	Devanagari	Western panjabi	Arabic
Sotho-tswana (s.30)	Northern sotho	Latin	Southern sotho	Latin

Table 22: NER cases where low-resource varieties perform better in zero-shot. The script plays a significant role here as most of the cases, high-resource scripts such as Latin, and Arabic could uplift the performance of a low-resource variety over a high-resource non-Latin script.

Table 23: Language clusters with their *linguistic utility*, demographic utility, variety with minimum *linguistic utility* and standard deviation. Task: DEP Parsing

clusters	*linguistic utility*	*demographic utility*	variety (minimum ling. u.)	standard deviation
albanian	63.3	69.1	(gheg albanian, 43.5)	28.0
anglic	75.5	90.7	(african american vernacular english, 52.5)	20.4
arabic	64.4	83.3	(south levantine arabic, 49.9)	21.0
eastern-western armenian	88.2	87.5	(eastern armenian, 87.0)	1.6
gallo-italian	50.2	50.2	(ligurian, 50.2)	-
gallo-rhaetian	87.7	93.4	(french (a:paris), 77.5)	8.8
high german	56.6	76.5	(central alemannic (a:zh), 36.8)	28.1
italian romance	75.4	76.5	(continental southern italian, 50.0)	17.3
komi	30.2	30.1	(komi-zyrian (m:written), 27.6)	2.4
norwegian	88.4	-	(norwegian nynorsk (m:written, i:old), 78.4)	8.6
saami	47.9	67.0	(skolt saami, 28.3)	27.8
sabellic	33.2	-	(umbrian, 33.2)	-
sinitic	65.4	-	(classical chinese, 46.7)	20.8
southwestern shifted romance	87.2	91.6	(portuguese (a:european), 77.3)	7.9
tupi-guarani subgroup i.a	17.2	22.0	(mbyá guaraní (a:brazil), 9.0)	10.5
west low german	40.8	40.8	(west low german, 40.8)	-

Table 24: Language clusters with their *linguistic utility*, demographic utility, variety with minimum *linguistic utility* and standard deviation. Task: POS Tagging

clusters	*linguistic utility*	*demographic utility*	variety (minimum ling. u.)	standard deviation
albanian	70.1	74.3	(gheg albanian, 55.8)	20.2
anglic	92.6	97.1	(singlish, 88.0)	6.5
arabic	61.9	81.4	(levantine arabic, 43.8)	18.9
eastern-western armenian	93.8	93.0	(eastern armenian, 92.4)	2.0
gallo-italian	58.9	58.9	(ligurian, 58.9)	-
gallo-rhaetian	96.4	96.9	(old french (842-ca. 1400), 95.6)	0.7
high german	75.5	88.4	(central alemannic (a:zh), 62.6)	18.2
italian romance	82.6	80.9	(continental southern italian, 57.1)	17.9
komi	41.8	41.6	(komi-zyrian (m:written), 35.1)	6.0
neva	75.4	97.7	(neva (a:south-west trans), 61.7)	13.9
norwegian	97.6	-	(norwegian nynorsk (m:written, i:old), 95.7)	1.6
saami	56.5	78.3	(skolt saami, 34.1)	31.6
sabellic	11.9	-	(umbrian, 11.9)	-
sinitic	84.5	-	(classical-middle-modern sinitic (a:hongkong, o:traditional), 69.0)	13.7
southwestern shifted romance	90.5	95.7	(occitan, 76.8)	10.5
tupi-guarani subgroup i.a	13.7	18.5	(mbyá guaraní (a:brazil), 1.9)	13.7
west low german	69.7	69.7	(west low german, 69.7)	-

Table 25: Language clusters with their *linguistic utility*, demographic utility, variety with minimum *linguistic utility* and standard deviation. Task: NER

clusters	*linguistic utility*	*demographic utility*	variety (minimum ling. u.)	standard deviation
anglic	57.3	83.9	(jamaican creole english, 0.0)	40.9
arabic	81.2	86.0	(egyptian arabic, 73.3)	11.1
circassian	57.4	54.1	(kabardian, 47.5)	14.0
common turkic	67.9	85.1	(south azerbaijani, 31.7)	28.8
eastern romance	75.9	79.3	(moldavian, 64.6)	16.0
frisian	66.8	80.4	(northern frisian, 54.8)	13.3
gallo-italian	68.6	68.0	(emiliano-romagnolo, 45.5)	17.3
gallo-rhaetian	71.8	90.3	(walloon, 46.3)	16.5
greater panjabic	54.9	53.2	(eastern panjabi, 45.2)	13.6
greek	81.9	90.1	(pontic, 72.6)	13.2
high german	71.5	88.0	(pennsylvania german, 41.8)	15.9
hindustani	86.6	88.1	(fiji hindi, 85.2)	2.1
inuit	69.7	63.5	(kalaallisut, 63.2)	9.2
italian romance	73.0	88.4	(continental southern italian, 61.5)	13.6
komi	55.2	56.6	(komi-permyak, 53.6)	2.2
kurdish	47.5	54.8	(central kurdish, 35.5)	17.0
latvian	71.7	88.7	(east latvian, 50.1)	30.5
mari	52.6	47.4	(eastern mari, 46.7)	8.3
modern dutch	83.8	91.2	(zeeuws, 79.5)	7.1
norwegian	89.7	-	(norwegian nynorsk (m:written), 87.9)	1.5
sardo-corsican	75.9	79.8	(corsican, 70.6)	7.5
serbian-croatian-bosnian	83.1	81.0	(serbian standard, 65.9)	11.9
sinitic	64.3	78.3	(hakka chinese, 40.0)	16.8
sorbian	68.3	67.4	(upper sorbian, 65.5)	3.9
sotho-tswana (s.30)	49.5	41.1	(northern sotho, 29.7)	28.0
southwestern shifted romance	74.0	92.1	(mirandese, 49.3)	18.1
west low german	80.3	80.3	(west low german, 80.3)	-

Table 26: Language clusters with their *linguistic utility*, demographic utility, variety with minimum *linguistic utility* and standard deviation. Task: NLI

clusters	*linguistic utility*	*demographic utility*	variety (minimum ling. u.)	standard deviation
anglic	83.4	83.4	(english, 83.4)	-
arabic	63.3	70.0	(tunisian arabic, 50.2)	7.1
common turkic	64.1	62.3	(south azerbaijani, 44.6)	16.9
gallo-italian	61.7	63.6	(ligurian, 57.2)	6.0
gallo-rhaetian	54.6	54.6	(friulian, 54.6)	-
high german	59.9	59.8	(limburgan, 59.7)	0.2
italian romance	70.4	77.1	(sicilian, 62.7)	11.0
kurdish	51.4	55.5	(central kurdish, 39.6)	16.7
latvian	63.6	71.5	(east latvian, 53.5)	14.2
modern dutch	76.5	76.5	(dutch, 76.5)	-
norwegian	75.3	-	(norwegian nynorsk (m:written), 71.1)	6.0
sardo-corsican	58.3	58.3	(sardinian, 58.3)	-
sinitic	68.2	67.4	(classical-middle-modern sinitic (o:traditional), 64.5)	4.1
sotho-tswana (s.30)	35.3	35.6	(southern sotho, 34.6)	1.0
southwestern shifted romance	76.3	79.1	(occitan, 68.5)	5.2

Table 27: Language clusters with their *linguistic utility*, demographic utility, variety with minimum *linguistic utility* and standard deviation. Task: TC

clusters	*linguistic utility*	*demographic utility*	variety (minimum ling. u.)	standard deviation
anglic	89.7	89.7	(english, 89.7)	-
arabic	84.5	85.7	(moroccan arabic, 79.1)	2.4
common turkic	78.7	77.3	(south azerbaijani, 69.7)	7.9
gallo-italian	73.8	73.7	(lombard, 70.6)	3.0
gallo-rhaetian	68.8	68.8	(friulian, 68.8)	-
high german	74.5	72.7	(limburgan, 71.1)	4.8
italian romance	81.4	86.8	(sicilian, 75.2)	8.8
kurdish	43.8	52.3	(central kurdish, 19.4)	34.5
latvian	75.6	82.0	(east latvian, 67.4)	11.5
modern dutch	89.6	89.6	(dutch, 89.6)	-
norwegian	86.7	-	(norwegian bokmål (m:written), 84.1)	3.6
sardo-corsican	71.0	71.0	(sardinian, 71.0)	-
sinitic	89.5	89.5	(classical-middle-modern sinitic (o:simplified), 89.2)	0.3
sotho-tswana (s.30)	37.8	36.9	(northern sotho, 35.6)	3.1
southwestern shifted romance	87.2	86.9	(occitan, 84.1)	2.3

Table 28: Language clusters with their *linguistic utility*, demographic utility, variety with minimum *linguistic utility* and standard deviation. Task: DId

clusters	*linguistic utility*	*demographic utility*	variety (minimum ling. u.)	standard deviation
anglic	89.0	88.4	(north american english, 88.0)	1.4
arabic	58.1	89.0	(south levantine arabic (a:south-amm), 43.0)	11.3
greek	64.6	-	(cypriot greek (r:casual, m:written, i:twitter), 56.8)	6.8
high german	74.5	-	(central alemannic (a:be), 72.0)	2.1
sinitic	98.3	98.3	(mandarin chinese (a:mainland, o:traditional, i:synthetic), 97.9)	0.4
southwestern shifted romance	73.3	82.7	(portuguese (m:written), 17.4)	27.9

Table 29: Language clusters with their *linguistic utility*, demographic utility, variety with minimum *linguistic utility* and standard deviation. Task: SA

clusters	*linguistic utility*	*demographic utility*	variety (minimum ling. u.)	standard deviation
arabic	80.3	81.4	(levantine/south, 58.9)	10.7

Table 30: Language clusters with their *linguistic utility*, demographic utility, variety with minimum *linguistic utility* and standard deviation. Task: MRC

clusters	*linguistic utility*	*demographic utility*	variety (minimum ling. u.)	standard deviation
anglic	53.4	53.4	(english, 53.4)	-
arabic	39.9	42.1	(moroccan arabic, 37.6)	2.4
sinitic	48.3	-	(classical-middle-modern sinitic (o:traditional), 46.9)	2.1
sotho-tswana (s.30)	30.1	30.5	(southern sotho, 29.0)	1.5

Table 31: Language clusters with their *linguistic utility*, demographic utility, variety with minimum *linguistic utility* and standard deviation. Task: EQA

clusters	*linguistic utility*	*demographic utility*	variety (minimum ling. u.)	standard deviation
anglic	75.2	75.4	(indian english (a:south), 71.9)	1.8
arabic	77.2	77.0	(egyptian arabic, 76.5)	0.6
bengali	73.8	-	(vanga (a:west bengal), 73.3)	0.7
korean	65.2	64.6	(korean (a:south-eastern, m:spoken), 64.6)	0.8
swahili	67.9	64.1	(swahili (a:tanzania), 63.5)	6.2