Evaluating Pixel Language Models on Non-Standardized Languages

Alberto Muñoz-Ortiz **Verena Blaschke** **Barbara Plank**
Universidade da Coruña, CITIC, Spain
LMU Munich, Center for Information and Language Processing (CIS), Germany
Munich Center for Machine Learning (MCML), Germany
alberto.munoz.ortiz@udc.es, {verena.blaschke,b.plank}@lmu.de

Abstract

We explore the potential of pixel-based models for transfer learning from standard languages to dialects. These models convert text into images that are divided into patches, enabling a continuous vocabulary representation that proves especially useful for out-of-vocabulary words common in dialectal data. Using German as a case study, we compare the performance of pixel-based models to token-based models across various syntactic and semantic tasks. Our results show that pixel-based models outperform token-based models in part-of-speech tagging, dependency parsing and intent detection for zero-shot dialect evaluation by up to 26 percentage points in some scenarios, though not in Standard German. However, pixel-based models fall short in topic classification. These findings emphasize the potential of pixel-based models for handling dialectal data, though further research should be conducted to assess their effectiveness in various linguistic contexts.

1 Introduction

Despite being spoken by millions of people worldwide, dialects and other non-standard language forms are largely underrepresented in Natural Language Processing (NLP) systems. Although pretrained language models (PLMs) achieve strong results for languages seen during training, where more data is available, their performance declines with out-of-domain dialects.

One of the primary factors contributing to the poor performance of PLMs on non-standard language varieties is tokenization, as tokenizers frequently break dialects into sub-tokens that lack meaning. Modifying tokenization has been shown to improve performance on non-standard data (Aepli and Sennrich, 2022; Blaschke et al., 2023; Srivastava and Chiang, 2023a,b).

In this context, dialectal variations can be viewed as a form of perturbation: tokenizing dialect data often produces tokens that are not meaningful. However, despite these variations, native speakers of the standard language can still comprehend dialects up to certain point due to linguistic and visual similarities. This suggests that visual cues may help models address the tokenization challenges posed by dialectal variations more effectively.

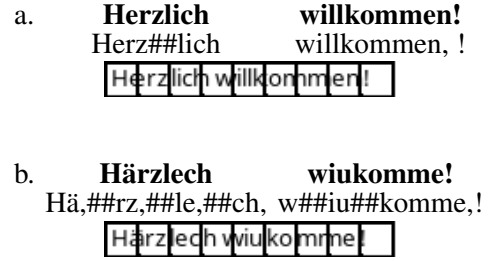

Figure 1: "Welcome!" in Standard German (a) and the Swiss German Bern dialect (b) tokenized using DBMDZ German BERT and rendered and split in patches by PIXEL. Standard German is tokenized in a more meaningful way, whereas the Bernese dialect form results in multiple non-meaningful sub-tokens due to variations in spelling.

To address the limitations of traditional tokenization, visual text representations convert text into images divided into patches, offering an alternative approach. Prior studies, such as Salesky et al. (2021) and Rust et al. (2023), have demonstrated that this approach effectively manages diverse scripts and languages without expanding the vocabulary, outperforming token-based approaches in syntactic tasks and machine translation, but not in semantic tasks. Strategies like structured rendering address this issue (Lotz et al., 2023).

Following this approach, we explore the use of pixel-based models to enhance NLP performance on dialects. Using German as a case study, we pretrain a pixel-based model from scratch, which we release it publicly.[1] We compare its performance

[1] https://huggingface.co/amunozo/pixel-base-german

Proceedings of the 31st International Conference on Computational Linguistics, pages 6412–6419
January 19–24, 2025.

to token-based models pretrained on the same data. Our results show that pixel-based models outperform token-based ones in syntactic tasks for zero-shot dialect evaluation but not in Standard German. In sentence-level semantic tasks, it excels in intent detection but is outperformed by token-based models in topic classification.

2 Pixel-based models

The **Pix**el-based **E**ncoder of **L**anguage (Rust et al., 2023), or PIXEL, is a model that casts language modeling as a visual recognition task. It is composed of: (1) **Text Renderer**: The text rendered transforms a string of text into a RGB image, divided into equal-sized patches of 16x16 pixels. (2) **Encoder**: Once the text is converted into patches, the image fed to a Vision Transformer (He et al., 2022) architecture that processes it. PIXEL uses a 12-layer transformer with a total of 86M parameters as encoder. (3) **Decoder or Task-Specific Head**: Instead of masking tokens, PIXEL masks spans of patches, using a decoder to reconstruct the masked patches as pretraining task. This decoder is discarded after pretraining, being replaced by a task-specific classification head for fine-tuning.

Using a continuous vocabulary allows PIXEL to handle multiple languages and scripts without the need of expanding its vocabulary. Also, this makes it robust to orthographic noise, such as typos or non-standard spellings, as it can generalize over orthographic variations which would break the tokenizations of token-based models. Finally, it avoids the high computational cost of a large vocabulary.

However, despite comparable parameter counts, PIXEL requires more fine-tuning steps to converge than a token-based model for the same data (Rust et al., 2023). Additionally, rendering text as images significantly increases disk space usage compared to plain text. While dynamic rendering during training or inference could alleviate storage concerns, it increases computational overhead.

3 Experiments

3.1 Setup

We investigate how pixel-based models pretrained on monolingual data compare to BERT (Devlin et al., 2019) when evaluated on dialectal data. To this end, we pretrain a pixel-based model on monolingual German data from the DBMDZ corpus,[2] following Rust et al. (2023).[3]

We choose a monolingual model as they perform competitively with multilingual models on dialect data (Bernier-Colborne et al., 2022; Castillo-lópez et al., 2023), and also due to computational constraints. We select German as our study language because of its wide range of dialectal variations, which show different degrees of standardization and are supported by available annotated data.

As a baseline, we use the cased[4] and uncased[5] versions of DBMDZ German BERT, pretrained on the same data. For simplicity, we will refer to the models as `bert-cased`, `bert-uncased`, and `pixel`. We fine-tune the three PLMs on part-of-speech (POS) tagging, dependency parsing, topic classification, and intent detection. For POS tagging, dependency parsing and topic classification, the models are trained on Standard German and evaluated on dialects. For intent detection, we train on both Standard German and dialects. The results were averaged over five runs. We followed]] the hyperparameters and setup for pretraining and fine-tuning from Rust et al. (2023). Detailed information about the datasets is available in the Appendix (Table 5).

3.2 German non-standard varieties

We evaluate our model on four non-standard language varieties related to German. **Bavarian** and **Alemannic** are dialect groups spoken in the South of the German-speaking area. They are pronounced differently than Standard German (which is expressed when the dialects are written), and their vocabulary and grammar also show differences to Standard German (Merkle, 1993; Christen, 2019). Neither dialect group has any widely adopted orthography. For Alemannic, we focus on Swiss German and Alsatian German. **Low Saxon** is a regional language spoken in Northern Germany and parts of the Netherlands. It is not standardized and encompasses multiple dialects (Wiesinger, 1983). Finally, we include **code-switched Turkish–German** data. The code-switching occurs on the level of morphemes, words and phrases.

[2] https://huggingface.co/datasets/stefan-it/german-dbmdz-bert-corpus

[3] We use the code from https://github.com/xplip/pixel

[4] https://huggingface.co/dbmdz/bert-base-german-cased

[5] https://huggingface.co/dbmdz/bert-base-german-uncased

		GSD			HDT		
Language	**Model**	**Acc.**	**UAS**	**LAS**	**Acc.**	**UAS**	**LAS**
German GSD	bert-cased	**96.2**	89.6	85.6	90.5	83.8	77.9
	bert-uncased	96.2	**89.8**	**85.8**	90.9	**84.3**	**78.5**
	pixel	95.2	86.1	81.3	**91.5**	82.5	76.3
German HDT	bert-cased	**89.9**	**89.5**	**84.1**	**98.6**	97.8	**96.9**
	bert-uncased	89.8	89.3	83.9	98.5	**97.8**	96.9
	pixel	89.6	88.5	82.6	98.5	96.9	95.8
Bavarian MaiBaam	bert-cased	**54.6**	53.0	35.6	43.1	32.6	23.2
	bert-uncased	46.1	44.7	28.7	33.1	26.2	18.4
	pixel	54.5	**54.0**	**38.3**	**48.4**	**39.5**	**29.4**
Low Saxon LSDC	bert-cased	33.3	**34.1**	17.8	17.3	9.5	5.9
	bert-uncased	33.6	32.7	16.8	17.4	8.6	5.3
	pixel	**37.2**	32.9	**18.1**	**23.9**	**14.1**	**8.2**
Turkish–German SAGT	bert-cased	**56.1**	**42.5**	**32.4**	54.7	**38.6**	**31.8**
	bert-uncased	54.4	40.8	32.1	53.8	38.3	31.5
	pixel	55.6	40.2	29.8	**55.5**	37.3	29.9
Swiss German UZH	bert-cased	58.2	50.1	33.3	45.1	31.3	22.6
	bert-uncased	50.6	41.9	26.6	35.5	26.7	18.4
	pixel	**59.2**	**51.5**	**35.8**	**54.9**	**39.6**	**29.7**
Swiss German NOAH's	bert-cased	63.1	—	—	54.2	—	—
	bert-uncased	55.3	—	—	45.2	—	—
	pixel	**63.4**	—	—	**62.1**	—	—
Alsatian BISAME	bert-cased	45.8	—	—	34.1	—	—
	bert-uncased	49.6	—	—	30.4	—	—
	pixel	**53.3**	—	—	**48.2**	—	—

Table 1: **POS tagging and dependency parsing performance** (in %) of models trained on German GSD and HDT and tested on different dialects.

3.3 Syntactic tasks

We cover POS tagging and dependency parsing together in this subsection, as both tasks are evaluated using the same datasets and show similar results.

Data For both POS tagging and dependency parsing, we use treebanks from Universal Dependencies (UD) (Nivre et al., 2020; de Marneffe et al., 2021) along with two non-UD datasets for Alemannic: NOAH's Corpus (Hollenstein and Aepli, 2014) and Alsatian Bisame GSW (STIH, 2020). The models were trained on two Standard German treebanks: GSD and HDT.

Results Table 1 shows the POS-tagging accuracy and (un)labelled attachment scores (UAS, LAS) of the models trained on Standard German and evaluated on dialects.

When trained on the GSD dataset, `bert-cased` performs best on Standard German treebanks and Turkish German code-switching, while `pixel` outperforms BERT on the Alemannic treebanks, Low Saxon, and Alsatian. It also performs comparably to BERT on Bavarian for POS tagging and outperforms it for dependency parsing.

When trained on HDT, `pixel` widens the performance gap, outperforming BERT on most dialect treebanks except HDT itself. Although all models experience a decline in accuracy, UAS and LAS, `pixel` demonstrates greater robustness.

In both tagging and parsing, `pixel` outperforms BERT during zero-shot evaluation on German dialects. BERT shows contrasting results: `bert-uncased` achieves the best performance on Standard German but performs significantly worse on dialects. Since nouns in German are capitalized, `bert-cased` likely leverages this feature to compensate for poor tokenization when processing dialects.

Accuracy per POS tag To gain deeper insights, we calculate the average accuracy per POS tag for each model trained on the Standard German treebanks and evaluated on dialects (Table 2).

For GSD, `pixel` outperforms BERT for all tags except `DET`, `NOUN`, `PROPN`, `ADV`, and `X`, which is the only tag where `bert-cased` outperforms `pixel` when trained on HDT. While these results are difficult to fully explain, there are plausible explanations for certain tags. For example, memorization plays a role for `PROPN`, which favors token-based models, and proper nouns might vary less between languages. Furthermore, words tagged as `NUM` and `PUNCT` exhibit visual similarities within each group, which benefits `pixel`.

LAS per dependency length To help explain why relative performance in POS tagging is better than in dependency parsing, we measured LAS performance based on dependency lengths, as in Rust et al. (2023). Figure 2 plots LAS per dependency length for models trained on GSD. Results on Standard German and Turkish German diverge as dependency length grows. Moreover, `bert-cased` and `bert-uncased` show similar results.

For dialects, however, `pixel` achieves higher LAS across all lengths for Bavarian and Alemannic, and the performance gap neither consistently widens nor narrows. For Low Saxon, where overall results are lower, `pixel`'s performance relative to BERT improves with increasing distance, surpassing BERT at distances of 3 and beyond, but not at shorter distances. Interestingly, the `pixel`'s poorer handling of long dependencies observed for Standard German and in Rust et al. (2023) is not observed when evaluation on dialects.

Lastly, we observe that `bert-uncased` performs considerably worse than `bert-cased`, unlike for Standard German.

Src	Model	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
GSD	bert-cased	42.9	55.5	**47.1**	29.8	69.6	**31.2**	0.0	**51.5**	58.7	16.7	40.4	**89.4**	99.3	41.0	5.6	43.8	**11.2**
	bert-uncased	32.5	55.1	44.1	22.9	71.2	27.6	0.0	40.9	58.2	18.4	41.1	87.2	99.3	**43.5**	0.3	39.6	8.0
	pixel	**49.3**	**60.8**	45.2	**31.7**	**75.9**	26.5	0.0	50.7	**63.9**	**21.8**	**46.5**	87.1	**99.8**	40.4	0.0	**52.8**	5.1
HDT	bert-cased	39.5	31.8	21.8	19.6	51.7	17.9	14.9	49.9	59.8	13.3	33.3	52.3	97.1	29.6	0.0	29.4	**66.6**
	bert-uncased	28.9	30.5	21.1	18.8	50.5	15.6	17.6	31.6	59.5	13.0	26.2	53.5	96.3	27.2	0.0	20.7	58.8
	pixel	**50.3**	**47.7**	**32.3**	**28.8**	**56.6**	**21.9**	**21.9**	**53.3**	**64.8**	**17.5**	**36.9**	**63.5**	**99.4**	**35.1**	0.0	**45.3**	64.2

Table 2: **Average accuracy per POS tag** (in %) for models trained on GSD and HDT when evaluating on dialect treebanks.

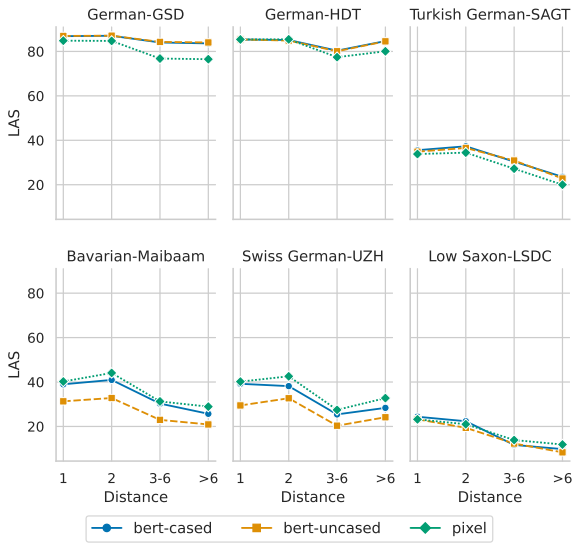

Figure 2: **Labelled attachment scores (in %) for different dependency distances** for models trained on German GSD.

Src	Model	de	ag	be	bs	gr	lu	sg	vs	zh
all	bert-cased	50.5	59.4	58.1	61.0	46.1	60.4	60.6	58.8	55.1
	bert-uncased	**50.7**	**63.5**	**58.1**	60.5	**46.7**	**61.9**	**62.2**	**60.9**	**58.1**
	pixel	44.5	57.6	57.1	**60.3**	43.8	57.3	58.6	58.1	55.1
ch	bert-cased	42.2	55.6	58.6	**63.9**	47.8	58.6	61.4	59.6	56.9
	bert-uncased	**45.7**	**59.6**	58.4	61.8	**48.2**	**62.7**	**62.2**	**62.6**	**60.5**
	pixel	36.3	57.1	**58.9**	58.2	41.1	56.8	57.8	58.6	54.9
de	bert-cased	50.0	**34.0**	**34.2**	**40.0**	30.4	**30.3**	**41.7**	**32.8**	**32.0**
	bert-uncased	**52.4**	22.3	19.5	25.6	34.4	21.1	29.2	26.5	25.2
	pixel	45.4	30.9	30.1	36.2	**35.3**	28.0	33.8	32.1	30.1
gr	bert-cased	44.0	46.7	49.1	49.7	48.2	**48.6**	51.2	49.5	45.6
	bert-uncased	**47.3**	**50.0**	**51.7**	**50.8**	**49.1**	44.2	**52.2**	**50.8**	**45.7**
	pixel	41.0	42.1	48.8	44.4	43.8	44.2	48.1	41.9	37.5

Table 3: **Topic classification accuracy** (in %) for models trained in the four training setups and evaluated on various targets in the SwissDial dataset. Key: de: Standard German, ag: Aargau, be: Bern, bs: Basel-Stadt, gr: Grisons, lu: Lucerne, sg: St. Gallen, vs: Valais, zh: Zurich, ch: all Swiss dialects.

3.4 Topic Classification

Data We use SwissDial (Dogan-Schönberger et al., 2021), a parallel corpus of Standard German and 8 Swiss German dialects, annotated with topic labels. We trained the models on four different datasets: `all` (a combination of Standard German and the 8 Swiss German dialects), `ch` (only the 8 Swiss German dialects), `de` (only Standard German), and `gr` (the Grisons dialect, which is the dialect that contains the most examples in the dataset and has the largest number of them in common with the Standard German data). We evaluate all the models on each variety.

Results The results are shown in Table 3. BERT models outperform `pixel` in most setups: `bert-cased` performs best when trained only on Standard German, and `bert-uncased` in the rest.

`Pixel` only outperforms BERT in two cases: on the Basel-Stadt dialect when trained on `all`, and on the Grisons dialect when trained on `de`. Although two improvements are not enough to draw conclusions, we observe that in both cases, the models have been trained on at least some Standard German data and evaluated on Swiss German dialects.

Transfer learning from Standard German to Swiss dialects is competitive with BERT, but the opposite is not true: BERT models trained on dialect data and tested on Standard German outperform `pixel`, likely due to more efficient tokenization.

3.5 Intent Detection

Data We use xSID 0.5 (van der Goot et al., 2021; Aepli et al., 2023; Winkler et al., 2024), a cross-lingual slot and intent detection dataset. We use the machine-translated German training set, the (human-translated) German test set (de) in addition to one Swiss German (gsw) and two Bavarian (de_ba, de_st) test sets. We additionally use the translated and naturalistic Bavarian intent classification test sets introduced by Winkler et al. (2024) (MAS:de-ba, nat:de-ba).

Results Table 4 shows the accuracies on the test sets. `Pixel` outperforms BERT for every dialect except MAS:de-ba. The differences are substantial

Model	de	M:ba	nat:ba	de-ba	de-st	gsw
bert-cased	98.2	20.0	65.4	78.6	79.8	62.4
bert-uncased	**99.2**	**26.2**	65.4	74.8	76.2	57.6
pixel	97.0	23.3	**76.2**	**79.2**	**88.6**	**83.8**

Table 4: **Intent classification accuracy** (in %) for models trained on Standard German and evaluated on German dialects.

for Swiss German, and notable but more modest for the Bavarian dialects. These results show promising signs of pixel for certain semantic tasks when evaluating on dialects.

3.6 Discussion

Pixel-based models has been show to underperform in sequence classification tasks (Rust et al., 2023). This can be attributed to the uniform rendering of text into 16×16 pixel patches. Unlike token classification, where words consistently map to similar patches, sequence classification introduces variability due to sentence progression, leading to slight variations in representation for the same word. Lotz et al. (2023) explored patch multiplicity reduction strategies, like pairing two characters per patch or using monospace fonts.

In our experiments, results in topic classification match this trend. However, pixel-based models outperform token-based ones in intent detection. This disparity may arise due to the dataset complexity, as topic classification on Standard German (∼50% accuracy) is inherently more challenging than intent detection (∼100% accuracy).

4 Conclusion

We presented a study on the use of pixel-based pretrained language models for zero-shot dialect evaluation, using German as a case study. Pixel-based models achieved higher scores than both cased and uncased token-based models when trained on Standard German and evaluated on German dialects for POS tagging, dependency parsing, and intent detection. However, they lagged behind token-based models in topic classification and for all tasks when evaluated on Standard German.

Pixel-based models showed promising results, particularly in intent detection, highlighting their potential in handling linguistic diversity. While their performance in topic classification indicates areas for further refinement and study, these models offer a novel approach to addressing dialectal NLP tasks. The current limitations in the availability of dialectal datasets present challenges for conducting a comprehensive evaluation, but we argue that pixel models have the potential to expand their utility across a broader range of NLP tasks, providing robust and adaptable language processing solutions, especially in low-resource contexts.

This work highlights the potential of pixel-based models in tackling the challenges posed by non-standard language varieties. With sufficient computational resources and data, multilingual pixel-based models could prove valuable for low-resource languages by bypassing tokenization and vocabulary limitations. However, the resource constraints commonly associated with low-resource language research (Ahia et al., 2021) may hinder their practical adoption. While the model's success with German dialects is encouraging, its generalizability to other languages and dialects remains uncertain.

Limitations

This study is constrained by several factors. Due to the high computational cost of pretraining language models, we focused on a single language, German, which limited our ability to explore other languages with different morphological or syntactic structures, or multilingual approaches. The scarcity of annotated data for dialectal varieties further restricted the scope of our experiments, excluding potential tasks and languages. While we included multiple dialects, the data imbalance and annotations quality may have introduced biases.

While our results show promise for German dialects, the generalization of these findings to other languages and language families remains uncertain and requires further investigation.

Acknowledgments

We thank Huangyan Shan for her assistance in identifying and cleaning appropriate datasets for this work, as well as for contributing to some preliminary visualizations.

This work was funded by the European Research Council (ERC) Consolidator Grant DIALECT 101043235; SCANNER-UDC (PID2020-113230RB-C21) funded by MICIU/AEI/10.13039/501100011033; Xunta de Galicia (ED431C 2024/02); GAP (PID2022-139308OA-I00) funded by MICIU/AEI/10.13039/501100011033/ and by ERDF, EU; Grant PRE2021-097001 funded

by MICIU/AEI/10.13039/501100011033 and by ESF+ (predoctoral training grant associated to project PID2020-113230RB-C21); LATCHING (PID2023-147129OB-C21) funded by MICIU/AEI/10.13039/501100011033 and ERDF; and Centro de Investigación de Galicia "CITIC", funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

References

Noëmi Aepli and Simon Clematide. 2018. Parsing Approaches for Swiss German. In *Proceedings of the 3rd Swiss Text Analytics Conference (SwissText), Winterthur, Switzerland*.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gabriel Bernier-Colborne, Serge Leger, and Cyril Goutte. 2022. Transfer learning improves French cross-domain dialect identification: NRC @ VarDial 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 109–118, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

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 *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France. Association for Computational Linguistics.

Galo Castillo-lópez, Arij Riabi, and Djamé Seddah. 2023. Analyzing zero-shot transfer scenarios across Spanish variants for hate speech detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 1–13, Dubrovnik, Croatia. Association for Computational Linguistics.

Helen Christen. 2019. Alemannisch in der Schweiz. In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum: Ein internationales Handbuch der Sprachvariation*, volume 4, pages 246–279. De Gruyter Mouton, Berlin, Boston.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken swiss german. *arXiv preprint arXiv:2103.11401*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

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

Jonas Lotz, Elizabeth Salesky, Phillip Rust, and Desmond Elliott. 2023. Text rendering strategies for pixel language models. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 10155–10172, Singapore. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

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

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Phillip Rust, Jonas F Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels. In *The Eleventh International Conference on Learning Representations*.

Elizabeth Salesky, David Etter, and Matt Post. 2021. Robust open-vocabulary translation from visual text representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7235–7252, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Janine Siewert and Jack Rueter. 2024. The Low Saxon LSDC dataset at Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15976–15981, Torino, Italia. ELRA and ICCL.

Aarohi Srivastava and David Chiang. 2023a. BERTwich: Extending BERT's capabilities to model dialectal and noisy text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15510–15521, Singapore. Association for Computational Linguistics.

Aarohi Srivastava and David Chiang. 2023b. Fine-tuning BERT with character-level noise for zero-shot transfer to dialects and closely-related languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 152–162, Dubrovnik, Croatia. Association for Computational Linguistics.

STIH. 2020. Bisame_gsw (alsacien) : corpus brut et annoté.

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

Peter Wiesinger. 1983. Die Einteilung der deutschen Dialekte. In Werner Besch, Ulrich Knoop, Wolfgang Putschke, and Herbert E. Wiegand, editors, *Ergebnisse dialektologischer Beschreibungen: Areale Bereiche deutscher Dialekte im Überblick*, pages 807–960. De Gruyter Mouton, Berlin, Boston.

Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, and Barbara Plank. 2024. Slot and intent detection resources for Bavarian and Lithuanian: Assessing translations vs natural queries to digital assistants. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14898–14915, Torino, Italia. ELRA and ICCL.

Özlem Çetinoğlu and Çağrı Çöltekin. 2022. Two languages, one treebank: Building a Turkish–German code-switching treebank and its challenges. *Language Resources and Evaluation*, pages 1–35.

A Datasets

Table 5 lists the downstream task datasets used for training and evaluation.

Dataset	Varieties	Train		Dev		Test		Task	Licence
		Sent	Word	Sent	Word	Sent	Word		
Token-level tasks									
UD German HDT 2.10 (Borges Völker et al., 2019)	German	153.0k	2799.1k	18.4k	324.8k	18.5k	331.7k	P, D	BY-SA 4.0 (annotations)
UD German GSD 2.10 (McDonald et al., 2013)	German	13.8k	263.8k	799	12.5k	977	16.5k	P, D	BY-SA 4.0
UD Swiss German UZH 2.10 (Aepli and Clematide, 2018)	Swiss German	—	—	—	—	100	1.4k	P, D	BY-SA 4.0
UD Low Saxon LSDC 2.10 (Siewert and Rueter, 2024)	Low Saxon	—	—	—	—	83	2.5k	P, D	BY-SA 4.0
UD Turkish German SAGT 2.10 (Çetinoğlu and Çöltekin, 2022)	Code-switched Turkish–German	—*	—*	—*	—*	805	13.9k	P, D	BY-SA 4.0
UD Bavarian MaiBaam 2.14 (Blaschke et al., 2024)	Bavarian	—	—	—	—	1.1k	15.0k	P, D	BY-SA 4.0
NOAH's corpus (Hollenstein and Aepli, 2014)	Swiss German	—	—	—	—	7.3k	113.6k	P	BY 4.0 (annotations)
Alsatian Bisame GSW (STIH, 2020)	Alsatian	—	—	—	—	382	8.2k	P	BY-NC-SA 3.0
Sentence-level tasks									
SwissDial (Dogan-Schönberger et al., 2021)	German, 8×Swiss German	*2.5k–4.1k sents per variety, split 80:10:10*						T	BY-NC 4.0
xSID 0.5 (van der Goot et al., 2021; Aepli et al., 2023; Winkler et al., 2024)	German, 2×Bavarian, Swiss German	—		—*		4×500		I	BY 4.0
NaLiBaSID MAS:de-ba (Winkler et al., 2024)	Bavarian	—		—		2.0k		I	not specified
NaLiBaSID nat:de-ba (Winkler et al., 2024)	Bavarian	—		—		315		I	not specified

Table 5: **Training and evaluation datasets used in our experiments.** P = part-of-speech tagging, D = dependency parsing, T = topic classification, I = intent classification. *The original dataset comes with training and/or development splits, but we do not use them.