

Sebastian, Basti, Wastl?!

Recognizing Named Entities in Bavarian Dialectal Data

Siyao Peng[▲][✉] Zihang Sun[▲] Huangyan Shan[▲] Marie Kolm[▲]
Verena Blaschke[▲][✉] Ekaterina Artemova[▲] Barbara Plank[▲][✉]

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

[✉] Munich Center for Machine Learning (MCML), Munich, Germany

{siyao.peng, b.plank}@lmu.de

Abstract

Named Entity Recognition (NER) is a fundamental task to extract key information from texts, but annotated resources are scarce for dialects. This paper introduces the first dialectal NER dataset for German, BARNER, with 161K tokens annotated on Bavarian Wikipedia articles (*bar-wiki*) and tweets (*bar-tweet*), using a schema adapted from German CoNLL 2006 and GermEval. The Bavarian dialect differs from standard German in lexical distribution, syntactic construction, and entity information. We conduct in-domain, cross-domain, sequential, and joint experiments on two Bavarian and three German corpora and present the first comprehensive NER results on Bavarian. Incorporating knowledge from the larger German NER (sub-)datasets notably improves on *bar-wiki* and moderately on *bar-tweet*. Inversely, training first on Bavarian contributes slightly to the seminal German CoNLL 2006 corpus. Moreover, with gold dialect labels on Bavarian tweets, we assess multi-task learning between five NER and two Bavarian-German dialect identification tasks and achieve NER SOTA on *bar-wiki*. We substantiate the necessity of our low-resource BARNER corpus and the importance of diversity in dialects, genres, and topics in enhancing model performance.

Keywords: Bavarian, German dialects, named entity recognition, low-resource languages, dataset

1. Introduction

Named Entity Recognition (NER) is a long-standing Natural Language Processing (NLP) task that extracts named entities (NEs) from texts and classifies them into a closed set of semantic types. A large number of NER datasets annotated for different genres and languages emerged after the seminal benchmark CoNLL 2003 shared task (Tjong Kim Sang and De Meulder, 2003), such as English (Baldwin et al., 2015; Strauss et al., 2016; Derczynski et al., 2017; Liu and Ritter, 2023), German (Benikova et al., 2014), code-switched and multilingual corpora (Aguilar et al., 2018; Piskorski et al., 2017; Singh, 2008; Liu et al., 2021; Fetahu et al., 2023), to name but a few. However, NER datasets for non-standard language varieties are scarce — there remains a demand for high-quality manual annotations on low-resource dialects.

Bavarian (German: *Bairisch*; Bavarian: *Boarisch*; ISO: 639-3; code: *bar*) is a West German dialect spoken in southern Germany, Austria, and northern Italy (South Tyrol). Bavarian has distinctive features in phonology, lexicons, and syntax compared to Standard German (Hinderling, 1984; Rowley, 2011). Given its large number of speakers (10M+, Rowley 2011), and regional variations in writings, we chose Bavarian to exemplify high-quality manual NER annotations on a non-standard language variety.

This paper presents BARNER, the first manually annotated NER dataset on a German dialect. BARNER contains 161K Bavarian tokens in two

genres: Wikipedia articles and tweets. We include coarse-grained person (PER), location (LOC), organization (ORG), and miscellaneous (MISC) entities strictly mirroring the CoNLL 2006 German guideline (Tjong Kim Sang and De Meulder, 2003), as well as fine-grained annotations on derived and partially-contained NEs and other entity types adapted from GermEval 2014 (Reimers et al., 2014). We conduct manual double annotations on half of the dataset and achieve high inter-annotator agreements. Primary and double BarNER annotations are publicly available on Github¹ as much as licenses allow. We also provide a training-centric NE annotation guideline on GitHub. We follow Bender and Friedman (2018) to include a data statement in Appendix B.

We highlight token- and entity-level distinctions between Bavarian and German NER. We compare our Bavarian wiki (*bar-wiki*) and tweet (*bar-tweet*) data to two German datasets within the same genres and the CoNLL 2006 German news dataset to illustrate how lexically distinct Bavarian is from German. We find cross-dialectal lexical dissimilarities to be larger than cross-genre in German. Moreover, the distribution of entity types and texts varies highly across dialects, genres, and topics — jointly referred to as “domains” in this paper.

We conduct in-domain, cross-domain, sequential, and joint NER experiments between Bavarian and German and across genres. We establish baseline NER scores on BARNER and demonstrate that directly applying German-trained

¹<https://github.com/mainlp/BarNER>

models achieves poor performance on Bavarian. Experiments show noticeable improvements in Bavarian by sequential and joint training and incorporating knowledge from larger German datasets. Inversely, training first on Bavarian can also help German NER, though to a smaller extent.

For multi-task learning (MTL), we train NER with Bavarian-German dialect identification. MTL scores SOTA on *bar-wiki* NER, 11.26 points higher on Span F1 than the in-domain baseline. Results demonstrate the efficacy of our multi-genre dialectal data in establishing low-resource evaluations and advancing high-resource performances.

2. Related Work

NER datasets for German The CoNLL 2003 shared task (Tjong Kim Sang and De Meulder, 2003) provides seminal NER datasets for German and English. Annotation guidelines and the German CoNLL dataset are subsequently updated to CoNLL 2006. On the other hand, German NoSta-D (Non-Standard German, Dipper et al. 2013) NER annotations were first implemented on five non-standard genre varieties – historical data, chat data, spoken data, learner data, and literary prose – and later extended to the prominent GermEval 2014 shared task (Benikova et al., 2014) on Wikipedia and online news. GermEval 2024 annotates nested NEs (though limited to two layers), and the -part and -deriv suffixes for compounds that partly contain NEs and words morphologically derived from NEs. Moreover, German NER datasets were expanded to various domains, such as parliament debates (Faruqui and Padó, 2010), historical newspapers (Neudecker, 2016; Hamdi et al., 2021), historical biodiversity literature (BIOfid, Ahmed et al. 2019), biographic interviews (Ruppenhofer et al., 2020), legal court decisions (Leitner et al., 2020), and traffic reports (Schiersch et al., 2018; Hennig et al., 2021). More fine-grained entity types are introduced in these corpora.

NER datasets for dialects and low-resource languages

The few dialectal NER datasets are predominately annotated for Arabic, where most surfaced in the last two years. These include Egyptian Arabic news and blogs (Darwish, 2013; Zirikly and Diab, 2014), Palestinian Arabic social media texts (Wojood, Jarrar et al. 2022), Narabizi (Latin-scripted vernacular Arabic) forums (NERDz, Touleb 2022), Algerian Facebook pages and YouTube channels (DzNER, Dahou and Cheragui 2023), and Darija (Moroccan vernacular Arabic) Wikipedia pages (DarNERcorp, Moussa and Mourhir 2023). Orasmaa et al. (2022) annotate Parish court records for north and south Estonian dialects (Finnic family). Besides, NER datasets

emerged for low-resource languages, such as Assamese (Pathak et al., 2022), Marathi (Litake et al., 2022), Kazakh (Yeshpanov et al., 2022), Vietnamese (Phan et al., 2022), and Sub-Saharan languages (Adelani et al., 2022).

Beyond gold annotations, Pan et al. (2017) create a silver WikiAnn dataset for 282 languages through Wikipedia knowledge mining and cross-lingual transfer, which includes Bavarian. However, many sentences in the small or medium-sized sub-corpora are unnatural. For example, the average sentence length is low (~5.26 tokens per sentence) in Bavarian, and half are hyperlinks, e.g., *Deutschland* [*Deutschland*]_{LOC} (the word for ‘Germany’ in Bavarian). Thus, there remains a gap for low-resource but high-quality NER datasets.

NER datasets for tweets Tweets form an unparalleled informal genre in NLP research given its (previously) short length limit, many users, and most timely posts (Finin et al., 2010; Scannell, 2022). Tweets contain NEs in texts, #hashtags, @mentions, and emojis. Finin et al. (2010) crowdsourced the earliest Twitter NER dataset for English. Ritter et al. (2011) ignored @mentions due to ambiguity and anonymity, and the corpus was used in Workshop on Noisy User-generated Text (WNUT) 2015-2017 shared tasks (Baldwin et al., 2015; Strauss et al., 2016; Derczynski et al., 2017). The Broad Twitter Corpus (Derczynski et al., 2016) samples from heterogeneous temporal, geographical, and social contexts. Tweebank-NER (Jiang et al., 2022) adds NEs to Universal Dependencies Tweebank V2 (Liu et al., 2018), and TweetNER7 (Ushio et al., 2022) is the largest tweet NER corpus to date. Beyond English, tweet NEs are annotated for German (Schiersch et al., 2018; Hennig et al., 2021), French (Lopez et al., 2017), Danish (Plank et al., 2020), Turkish (Küçük and Can, 2019; Çarık and Yeniterzi, 2022), Serbian and Croatian (Baksa et al., 2014, 2017; Ljubešić et al., 2019), etc.

3. BARNER

3.1. NER Taxonomy

We conduct NER annotations on Bavarian (*bar*) mirroring the CoNLL 2006 (Tjong Kim Sang and De Meulder, 2003) and GermEval 2014 (NoSta-D, Benikova et al. 2014) guidelines for German. We constrain our annotations to the narrowly defined but widely adapted flat NEs compatible with sequential BIO tagging. Namely, we exclude common nouns, pronouns, overlapping, or nested NEs.

We follow both guidelines to include the four major entity types: PER, ORG, LOC, and MISC. Since nominal derivation and compounding are similarly prevalent in Bavarian, we adapt NoSta-D’s strategy

to add *-deriv* and *-part* suffixes for tokens derived or partly containing NEs. For example, [*Italien-roas*]_{LOCpart} ‘tour of Italy’ partly contains the country Italy and [*eiropäischn*]_{LOCderiv} ‘European’ is an adjective derived from [*Eiropa*]_{LOC} ‘Europe.’ Moreover, annotators observe a relatively high frequency of NEs referring to languages (LANG), religions (RELIGION), events (EVENT), and works of art (WOA) during training, and we thus add them and their *-deriv* and *-part* suffixed variations to our NE tagset. These additional labels elaborate on entities in a given text and can be merged or discarded when comparing with other datasets (see §4.2). We refer to the PER/LOC/ORG/MISC tagset as coarse-grained and the extension with *-part*, *-deriv*, and other entity types as fined-grained.

3.2. Genre & Corpus Statistics

We conduct manual annotations on two mainstream genres: Wikipedia articles and Twitter (X) tweets. Wikipedia pages are carefully written and consistently updated by multiple contributors. We selected 40+ documents from Bavarian Wiki² with a wide topic coverage. Continuous sections of ~1.5K tokens from the beginning of documents were extracted to enable future document-level analyses.

Social media texts like tweets are noisier, less formal, and more dynamic (Ushio et al., 2022). Tweet collection for Bavarian is also more difficult. To sample enough data with author diversity, we snowballed from a list of 17 Bavarian ‘seed users’³ (Bhroin, 2015) to their friend circles on Twitter (i.e., the people they follow and follow them), under the assumption that dialect groups are closely connected on social media (Backstrom et al., 2006). All tweets⁴ from seed users and their friends are extracted to train a pilot dialect identifier to filter these users. We update the list of seed users iteratively until we reach 100K+ tokens on silver Bavarian tweets.

We further ensure the dialectal sanity of our tweet data by asking annotators to classify tweet sentences into one of the following categories: *bar* if the sentence is predominately Bavarian, *de* if predominately German, *other* if another language or dialect, and *na* if unintelligible⁵ – see §6 for our multi-task learning experiments combining NER with Dialect Identification (DID). As for our *bar-tweet* NER dataset, we only kept *bar*-labelled sen-

²<https://bar.wikipedia.org/wiki/Wikipedia:Hoamseitn>

³<http://indigenoustweets.com/bar/>

⁴4.4K+ tweets from 151 users were collected between Feb and May 2023, with no restriction on when the tweets were posted.

⁵Cohen’s kappa on intermediate and final dialect identifications are 82.47 and 85.66 between two annotators.


tences. During NER annotations, we include hash-tags (*#minga*]_{LOC} ‘Munich’) and emojis (_{LOC}) as our annotation targets. Mentions are anonymized to *@mention* and excluded from annotation.

Table 1 presents the number of tokens, sentences, and named entities in our two Bavarian NER sub-corpora: *bar-wiki* and *bar-tweet*. Both genres reach 75K+ tokens, a quarter the size of the German CoNLL 2006 corpus. However, we note that *bar-tweet* has proportionally much fewer entities than *bar-wiki* due to informality and tweet length limits (see §4 for comparisons with German).

Corpus	#Toks	#Sents	#Ents	%Ents/Toks
<i>bar-wiki</i>	75,687	3,574	4,192	5.54%
<i>bar-tweet</i>	86,090	7,459	2,486	2.89%

Table 1: *bar-wiki* and *bar-tweet* NER corpus statistics: numbers of tokens, sentences, entities, and percentages of entities over tokens.

3.3. Annotation Procedure & Agreement

The annotation project took five months and was conducted by three graduate students with computational linguistics backgrounds.⁶ These include one native Bavarian speaker and two majoring in German studies. For training, three annotators annotate the same documents independently during the first three months. Project coordinators hold discussion sessions biweekly and adjudicate the final version with annotators. After training, two annotators remained on the project and worked on different documents. We conduct two inter-annotator agreement (IAA) experiments (i.e., after training and final) between the two remaining annotators on the coarse- and fine-grained labels. Each IAA is evaluated on ~7K tokens, aligning with NER experiments’ development and test sets in §5. Table 2 presents the token- and entity-level IAA statistics.

Corpus	#Toks	Tagset	Token-level		Entity-level F1	
			Raw	Kappa	Untyped	Typed
Intermediate Agreement (Dev Set)						
<i>bar-wiki</i>	7.4K	fine	97.62	88.80	91.64	83.62
		coarse	97.93	89.76	91.82	86.13
<i>bar-tweet</i>	7.0K	fine	99.09	87.02	86.52	83.48
		coarse	99.29	88.19	87.21	85.64
Final Agreement (Test Set)						
<i>bar-wiki</i>	6.9K	fine	99.11	94.04	94.99	88.32
		coarse	99.24	94.64	94.84	90.52
<i>bar-tweet</i>	7.3K	fine	99.16	88.26	89.60	85.20
		coarse	99.27	88.22	88.10	87.62

Table 2: Inter-Annotator Agreement (IAA) on Bavarian wiki and tweet NER tagging, including token-level raw and Cohen’s kappa scores and entity-level untyped and typed F1 scores.

⁶The annotators were hired and compensated for their work following national salary rates.

Token- and entity-level agreements on Bavarian wiki and tweet are high; all Cohen’s kappa and typed span F1 are above 80, with an additional improvement on the final agreements. Since tweet entities are much more sparse than wiki (see Table 1), determining whether a noun phrase is an entity span is more difficult for tweets, resulting in a ~ 5 percentage point decrease in untyped F1. In contrast, entity typing is easier for tweets with smaller gaps between untyped and typed F1s. The IAA scores on coarse- and fine-grained entity types are also close. Our data release includes individual annotations on top of the adjudicated final version for future annotation disagreement studies.

4. Comparisons with German

To assess quantitatively and qualitatively how Bavarian NER differs from German, we compare our BARNER sub-corpora (*bar-wiki* and *bar-tweet*) with three German ones: the wiki portion of NoSta-D (henceforth *de-wiki*, Benikova et al. 2014), Mobile transportation tweets (*de-tweet*, Hennig et al. 2021), and the German CoNLL 2006 news (*de-news*, Tjong Kim Sang and De Meulder 2003). The first two share the same genres as BARNER; the last is the benchmark German NER dataset. §4.1 quantifies lexical similarities and §4.2 maps fine-grained tagsets into coarse-grained for entity type comparisons. §4.3 presents cross-genre and cross-dialectal distinctions in entity texts, and §4.4 supplements with annotators’ observations on differences in annotating Bavarian NER compared to German.

4.1. Lexical Similarities

This section demonstrates lexical distinctions from two aspects: between orthographically distinctive Bavarian and German; and among wiki, tweet, and news genres. We employ Jaccard similarity, which is frequently used in corpus analysis and measures the ratio of shared (i.e., intersection) tokens over concatenated (i.e., union) ones between datasets (Chen et al., 2022). Figure 1 presents the Jaccard similarities among the top 1K frequent tokens sampled from the Bavarian and German (sub-)corpora – see Appendix A for word clouds. We value variations in word choice and orthography across dialects and genres and thus compare surface token strings without normalization – analogous to how language models process texts.

Firstly, we observe that German *wiki*×*news* has the highest lexical similarity of 0.417 due to their formality and well-edited texts. Secondly, lexical distances between tweets and other same-dialect genres are similar for Bavarian and German, e.g., German *tweet*×*news* scores 0.229, *tweet*×*wiki* 0.195, and Bavarian *tweet*×*wiki* is 0.181. Lastly,

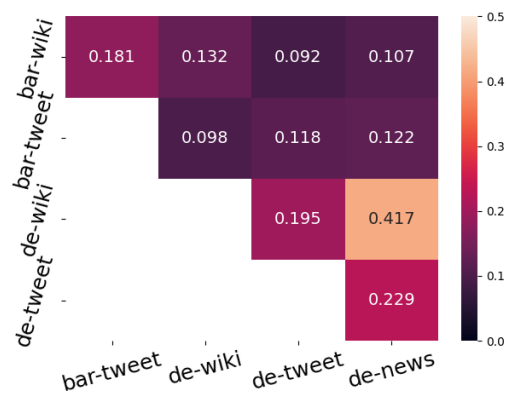


Figure 1: Heatmap of Jaccard similarities of top 1K frequent tokens across five (sub-)corpora: *bar-wiki*, *bar-tweet*, *de-wiki*, *de-tweet*, and *de-news*.

Jaccard similarities between dialects within the same genre are not distinguishable from cross-dialect and cross-genre scenarios. For example, *de-wiki*×*bar-wiki* 0.132 is close to *de-news*×*bar-tweet* 0.122. We can report from these observations that Bavarian is more lexically distinct from German than different text genres in German are from each other.

4.2. Entity Mappings & Distributions

The benchmark CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) dataset was updated to CoNLL 2006 with some guideline changes.⁷ These include: 1) adjectives derived from proper names (e.g., *deutsch* ‘adj. German’) are no longer marked as NEs; and 2) neither nouns derived from proper names (e.g., *Frankfurter* ‘people from Frankfurt’) 3) nor nominal compounds containing proper names (e.g., *SPD-Vorsitzender* ‘SPD-Chairman’) are marked. More recent NER datasets, such as BARNER, evolved from CoNLL 2006 and include additional entity types to capture language or domain-specific phenomena.

To conduct objective comparisons, we normalize all fine-grained tagsets to the CoNLL 2006 (*de-news*) coarse-grained one. Table 3 shows the tagset normalization rules for the four *bar/de-wiki/tweet* (sub-)datasets. These include removing nested, partly, and derived NEs and dropping or merging fine-grained entity types into MISC.

Figure 2 presents frequencies of the four coarse-grained entity types per 1K tokens in the five normalized (sub-)datasets. Firstly, we observe similar proportional frequencies of four entity types in *bar-wiki* and *de-wiki*, whereas *de-news* has a slightly lower ratio for MISC. Secondly, *de-tweet* exhibits an extreme outlier of 105.2 on LOC due

⁷<https://usermanual.wiki/Document/guide.820232904.pdf>

Normalization rules	<i>bar-wiki</i>	<i>bar-tweet</i>	<i>de-wiki</i>	<i>de-tweet</i>
Removing 2nd-level NEs			✓	
OTH -> MISC			✓	
EVENT/WOA -> MISC	✓	✓		✓
-part/deriv -> O	✓	✓	✓	
LANG/RELIGION -> O	✓	✓		
TIME/DISTANCE/NUMBER -> O				✓

Table 3: Tagset normalization on Bavarian and German NER (sub-)datasets.

to the deliberately sampled transportation tweets and the abundance of location entities, such as routes, streets, and cities. Lastly, *bar-tweet* has the least amount of entities proportionally, particularly low in ORG. The higher rank of PER entities than LOC and the frequent use of first-person pronouns in *bar-tweet* suggest that texts in the corpus are more personal than the other four (sub-)datasets, involving fewer proper nouns. Another contributing factor is the decision to anonymize and ignore all @mentions in annotations; nevertheless, PER is the most frequent NE type in *bar-tweet*. Even with normalized NER types, (sub-)datasets vary largely in the distribution of entity annotations, which can contribute positively or negatively when simultaneously or sequentially training on multiple NER (sub-)datasets as in §5.

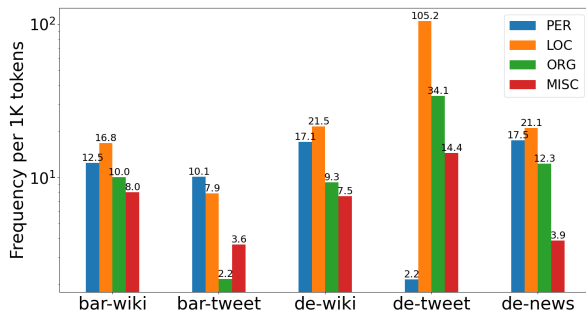


Figure 2: Frequencies of named entity types in German and Bavarian (sub-)datasets per 1K tokens (bars are displayed in log-scale).

4.3. Top Entity Texts

Besides the distribution of entity types, we also examine how entity texts differ across genres and dialects. Table 4 presents the top 10 entities from the five normalized (sub-)corpora. Some entities are dominant and shared across dialects, *Deutschland* ‘Germany’ in German or *Deitschland* in Bavarian. The most common cities differ between the two dialects: *Minga* ‘Munich’ is the most popular in Bavarian. In contrast, *Berlin* and *Frankfurt* frequently occur in the German corpora. Inevitably, different corpus designs result in the largest divergence in frequencies. We constructed our *bar-wiki*

sub-corpus to be around 1.5K tokens per document for future discourse analyses, which caused the document titles to be prominent entities. *bar-tweet* is sampled by a list of Bavarian-speaking users. Names of their frequently connected friends, e.g., *Michi*, *Gitte*, and *Gela*, are among the top entities in the dataset, particularly occurring in the greeting sentences. *de-wiki* contains the most city and country names. In the *de-news* corpus, mentions of the then-currency *Mark* and political parties like *SPD* or *CDU* are prominent. *de-tweet*, purposely sampled from the transportation domain, is dominated by local railway lines with or without hashtags, e.g., *#S3* and *S3*. These different entity distributions across (sub-)datasets contribute to the divergent performances when sequentially or jointly trained on different external NER (sub-)datasets in §5.

Corpus	Top 10 entities
<i>bar-wiki</i>	<i>Minga</i> _{LOC} , <i>Odysseus</i> _{PER} , <i>Nestroy</i> _{PER} , <i>Deutschland</i> _{LOC} , <i>Bundeswehr</i> _{ORG} , <i>Dochau</i> _{LOC} , <i>Los Angeles</i> _{LOC} , <i>Google</i> _{ORG} , <i>Doana</i> _{LOC} , <i>Bassau</i> _{LOC}
<i>bar-tweet</i>	<i>Gela</i> _{PER} , <i>Michi</i> _{PER} , <i>Minga</i> _{LOC} , <i>Gitte</i> _{PER} , <i>Stefan</i> _{PER} , <i>Jo Jo</i> _{PER} , <i>#wiesn</i> _{MISC} , <i>Bayern</i> _{LOC} , <i>Twitter</i> _{ORG} , <i>#Wiesn</i> _{MISC}
<i>de-wiki</i>	<i>Deutschland</i> _{LOC} , <i>Berlin</i> _{LOC} , <i>USA</i> _{LOC} , <i>München</i> _{LOC} , <i>Frankreich</i> _{LOC} , <i>Wien</i> _{LOC} , <i>Zweiten Weltkrieg</i> _{MISC} , <i>Europa</i> _{LOC} , <i>Österreich</i> _{LOC} , <i>Paris</i> _{LOC}
<i>de-tweet</i>	@ <i>SBahnBerlin</i> _{ORG} , <i>#S3</i> _{LOC} , <i>#S1</i> _{LOC} , <i>#S2</i> _{LOC} , <i>S1</i> _{LOC} , <i>S3</i> _{LOC} , <i>#S7</i> _{LOC} , <i>Unfall</i> _{MISC} , <i>#S4</i> _{LOC} , <i>#S6</i> _{LOC}
<i>de-news</i>	<i>Mark</i> _{MISC} , <i>Frankfurt</i> _{LOC} , <i>SPD</i> _{ORG} , <i>CDU</i> _{ORG} , <i>Grünen</i> _{ORG} , <i>Deutschland</i> _{LOC} , <i>Sachsenhausen</i> _{LOC} , <i>FR</i> _{ORG} , <i>Hessen</i> _{LOC} , <i>Weltbank</i> _{ORG}

Table 4: Top 10 frequent entities (cased) in the five normalized Bavarian and German (sub-)datasets.

4.4. Qualitative Observations

This section provides qualitative observations from our annotators when comparing Bavarian NER annotation to German. Person entities in Bavarian are the most distinct from German. Firstly, family names in Bavarian usually come before given names (Weiß, 1998, p. 71), e.g., *Dreßen* is the family name in *Dreßen Thomas*. Secondly, as diminutives are used more frequently in Bavarian, many given names have specific shortened versions as nicknames (Merkle, 1993, p. 108), e.g. *Sebastian* becomes *Basti* or *Wastl*. Furthermore, given names are typically preceded by definite articles (Weiß, 1998, p. 69–71), e.g., *d’Maria* and *da Michel*.

We also observe distinctions in the genitive case marking between the two dialects. Bavarian has three common grammatical cases: nominative, accusative, and dative. The genitive determiner in German is replaced by the combination of the preposition *vo* ‘from’ and a dative determiner in Bavarian (Merkle 1993, p. 96, Bülow et al. 2021). For example, ‘Association of National Olympic Committees’ is translated as *Vaeinigung vo de Nationoin Olympischn Komitees* in Bavarian with *vo*

de, rather than *Vereinigung der Nationalen Olympischen Komitees* in German with genitive *der*.

To summarize, §4 illustrates substantial variations in lexical distributions, entities, and grammatical constructions captured in annotations. These (sub-)datasets were organized to capture entities in designated topics, genres, and dialects.

5. NER Experiments & Results

This section examines whether neural architectures, language models, and external training resources help Bavarian NER tagging. §5.1 explains the basic experiment setups. §5.2 shows that in-domain NER tagging performances are lower in Bavarian than in German. §5.3 further illustrates degradations when testing on out-of-domain (OOD) (sub-)datasets. §5.4 and §5.5 show improvements through sequential and joint training with multiple Bavarian and German (sub-)datasets. However, NER on Bavarian or tweets is still more difficult than on German or the more prescriptive wiki and news genres.

5.1. Setups

We employ MaChAmp (van der Goot et al., 2021) for our NER experiments using the masked CRF decoder with BIO encoding and its default hyperparameters. MaChAmp allows joint and sequential training and achieves satisfying performances on the MultiCoNER shared task (Malmasi et al., 2022; Plank, 2022; van der Goot, 2022). We use two top-performing LMs, German *deepset/gbert-large*⁸ (GBERT, Chan et al. 2020) and multilingual *xlm-roberta-large*⁹ (XLM-R, Conneau et al. 2020). Experiments are conducted on five normalized German and Bavarian (sub-)datasets introduced in §4. Models are trained on an NVIDIA A100 GPU, and we report 3-run averages (avg) with standard deviations (std) on the Span F1 metric.

5.2. In-domain Training

We compare in-domain training of our Bavarian NER sub-corpora with three German ones, two from the same genres and one benchmark news corpus. Table 5 presents in-domain results with XLM-R and GBERT and the number of training tokens and entities in each dataset for comparison. NER is more difficult on *bar-wiki*, *bar-tweet*, and *de-tweet* since they are smaller in training size and represent non-mainstream variations in dialects or genres. Overall, German *news/wiki* text reaches F1 scores in the high 80s, while this drops to the 70s on the social media genre and for Bavarian. Moreover,

bar-tweet has sparser entity density (2.23%) than the other (sub-)datasets (see §4.2).

In-domain	<i>bar-wiki</i>	<i>bar-tweet</i>	<i>de-wiki</i>	<i>de-tweet</i>	<i>de-news</i>
#TrainToks	61.4K	71.8K	232.4K	47.0K	207.0K
#TrainEnts	2.7K	1.6K	12.9K	7.3K	10.0K
%Ents/Toks	4.40%	2.23%	5.55%	15.5%	4.83%
XLM-R	72.91 _{0.67}	77.55 _{0.64}	85.67 _{0.80}	77.14 _{0.69}	88.35 _{0.33}
GBERT	72.17 _{1.75}	73.30 _{6.98}	86.68 _{0.50}	79.75 _{0.62}	90.23 _{0.37}

Table 5: In-domain training results (3-run avg_{std} Span F1) and the number of training tokens and entities in five Bavarian and German (sub-)datasets.

In §5.3-§5.5, we experiment with cross-domain testing and sequential and joint training to optimize NER tagging on Bavarian. We choose XLM-R for further experiments given its higher Span F1s and lower standard deviations on Bavarian.

5.3. Cross-domain Evaluation

We conduct out-of-domain testing across five (sub-)datasets to assess whether and to what extent distinctions among training data surface in model performance. Table 6 presents results trained on the (sub-)datasets in the left columns and tested on the top rows. We include in-domain testing in the diagonal for comparison, and the best out-of-domain scenario on the target dataset is **bolded**.

Train\Test	<i>bar-wiki</i>	<i>bar-tweet</i>	<i>de-wiki</i>	<i>de-tweet</i>	<i>de-news</i>
<i>bar-wiki</i>	72.91 _{0.67}	35.33 _{2.37}	71.62 _{1.35}	55.39 _{2.07}	71.39 _{1.39}
<i>bar-tweet</i>	60.07 _{3.50}	77.55 _{0.64}	66.39 _{1.92}	48.63 _{2.70}	67.43 _{2.89}
<i>de-wiki</i>	57.45 _{5.29}	24.07 _{1.45}	85.67 _{0.80}	60.21 _{3.82}	81.91 _{2.42}
<i>de-tweet</i>	51.32 _{7.34}	37.32 _{2.23}	65.40 _{1.25}	77.14 _{0.69}	67.59 _{1.56}
<i>de-news</i>	48.73 _{4.98}	29.31 _{2.49}	78.66 _{0.24}	58.54 _{1.83}	88.35 _{0.33}

Table 6: Out-of-domain (OOD) evaluation results (3-run avg_{std} Span F1) across five (sub-)datasets.

Firstly, models trained on the larger *de-wiki* and *de-news* (sub-)corpora perform badly on Bavarian, 15.46 and 24.18 percentage points lower than in-domain performances on *bar-wiki*, 53.48 and 48.24 lower on *bar-tweet*. This shows that a model trained on the smaller but targeted Bavarian data (*de-wiki* and *de-news* are three times bigger) is more beneficial. *de-wiki* and *de-news* models also lose 10+ percentage points on *de-tweet*, but cross-genre degradations between *de-wiki* and *de-news* are relatively small. On the other hand, models trained on Bavarian also perform poorly on German data. These low cross-domain performances substantiate dissimilarities across Bavarian and German and the need for our NER annotations on Bavarian.

5.4. Sequential Training

Since *bar-wiki* and *bar-tweet* are smaller and do not target specific topics, we suffer from fewer training tokens and entities (see Table 5). We experiment

⁸<https://huggingface.co/deepset/gbert-large>

⁹<https://huggingface.co/xlm-roberta-large>

with sequential training to provide access to larger external entity resources. Table 7 demonstrates results where we first train on the (sub-)datasets in the left column and then continue training and evaluating on the top row target ones.

First\Target	<i>bar-wiki</i>	<i>bar-tweet</i>	<i>de-wiki</i>	<i>de-tweet</i>	<i>de-news</i>
<i>bar-wiki</i>	/	79.27 _{1.33} ↑	85.64 _{0.52}	77.26 _{0.83} ↑	88.50 _{0.54} ↑
<i>bar-tweet</i>	68.86 _{3.28}	/	85.57 _{0.16}	76.48 _{1.42}	88.84 _{0.45} ↑
<i>de-wiki</i>	73.67 _{1.16} ↑	78.07 _{1.00} ↑	/	76.76 _{0.81}	88.57 _{0.44} ↑
<i>de-tweet</i>	69.55 _{2.53}	76.32 _{0.82}	86.08 _{1.14} ↑	/	88.89 _{0.09} ↑
<i>de-news</i>	71.79 _{1.12}	78.65 _{1.87} ↑	85.11 _{1.28}	76.71 _{1.08}	/
In-domain	72.91 _{0.67}	77.55 _{0.64}	85.67 _{0.80}	77.14 _{0.69}	88.35 _{0.33}

Table 7: Sequential training results (3-run avg_{std} Span F1) on five (sub-)datasets. ↑ indicates an increase from in-domain baseline and the highest is bolded; e.g., **73.67**_{1.16}↑ is the avg_{std} first trained on *de-wiki* and then on *bar-wiki* which outperforms (↑) the in-domain baseline 72.91_{0.67} and achieves the best performance on *bar-wiki* (bolded).

Compared to in-domain training (the bottom row in Table 7), first training on one other dataset and then on the target improves performance on all five (sub-)datasets. Early training on same-genre *de-wiki* helps *bar-wiki* tagging by 0.76. *bar-tweet* achieves the largest gain of 1.67 by first training on *bar-wiki*, potentially since they belong to the same dialect and *bar-wiki* has higher entity density. First training on German wiki and news also helps *bar-tweet* but not on the same genre *de-tweet*, narrowly focusing on transportation texts. On German, *de-tweet*’s topic-heavy entities result in the highest performances on the much larger *de-wiki* (86.08) and *de-news* (88.89). All other Bavarian and German (sub-)datasets also slightly improve the performance on the seminal but 20-year-old *de-news*. To summarize, we show that sequential training can improve target data performance. We hypothesize that language diversity in genres, dialects, and topic-specific vocabularies contribute to these improvements.

5.5. Joint Training

In addition to sequential training, we analyze how training jointly with all five (sub-)datasets while developing and testing on the target influences model performances. We also pipeline joint training with another round of training on the target dataset (joint+seq). Table 8 demonstrates the results.

	<i>bar-wiki</i>	<i>bar-tweet</i>	<i>de-wiki</i>	<i>de-tweet</i>	<i>de-news</i>
joint	81.73 _{1.52} ↑	78.17 _{0.91} ↑	85.89 _{0.22} ↑	76.58 _{0.41}	87.62 _{0.46}
joint+seq	84.09 _{2.92} ↑	77.80 _{0.62} ↑	85.17 _{0.21}	75.78 _{1.08}	88.67 _{0.85} ↑
In-domain	72.91 _{0.67}	77.55 _{0.64}	85.67 _{0.80}	77.14 _{0.69}	88.35 _{0.33}

Table 8: Joint training results (3-run avg_{std} Span F1) on five (sub-)datasets with or without sequential training on the target dataset.

Joint training on five (sub-)datasets considerably

improves performance on *bar-wiki* (+8.82) and moderately on *bar-tweet* (+0.62), but not as much for the three German datasets. When adding pipeline training on the target, *bar-tweet* increases by another 2.36. Adding sequential to joint training also improves the large *de-news* benchmark.

To summarize, we here (§5) present four experiment scenarios, in-domain, cross-domain, sequential, and joint training, on five (sub-)datasets. We first establish in-domain NER baselines on BARNER. Secondly, we demonstrate the unsatisfying performances of directly applying German-trained models to Bavarian data, substantiating the necessity of our sizeable Bavarian annotations. Thirdly, we demonstrate first training on other (sub-)datasets and then training on the target one improves model performance. This could result from a larger training size, higher entity density, or more topic diversity. Lastly, jointly train on all five (sub-)datasets and then on the target achieves radical enhancement +11.18 on *bar-wiki* NER tagging.

6. Multi-Task Learning with Dialect Identification

Dialect identification (DID) is an NLP task discriminating between similar dialects. German is well-known for its intrinsic dialectal variation, but existing German DID datasets are geographically labeled and centralized on Swiss German, e.g., the VarDial 2017-2019 shared tasks (Zampieri et al., 2017, 2018, 2019). Given our gold selection of Bavarian wiki and tweet texts, this section conducts multi-task learning (MTL) between NER tagging and DID for Bavarian and German.

We conduct separate DID experiments for tweets and wikis. Our tweet DID dataset consists of all *bar-tweet* sentences (75K tokens) and an equal amount of German tweets from an in-house archive.¹⁰ For wiki DID, we extract 75K tokens from Bavarian and German wiki pages. We run the MaChAmp classification task with XLM-R and report averages on Micro F1 in the top-left corner of Table 9.

Models yield commendable performances on both DID datasets. Still, the average F1 score on tweet DID is 3.76 lower than on wiki. This results from the frequent code-switching between Bavarian and German in tweets and their shorter sentences, i.e., 10.64 tokens/sentence (incl. @mentions and #hashtags) in tweet DID data but 14.94 in wiki.

Furthermore, we evaluate multi-task learning with equal weights on the 5 NER and 2 DID tasks. As shown in Table 9, MTL is beneficial for our non-canonical data: It enhances both DID on the harder tweet genre and NER on two Bavarian sub-corpora

¹⁰We refrain from using Mobile for tweet DID due to its topic specificity.

	DID		NER				
	wiki	tweet	bar-wiki	bar-tweet	de-wiki	de-tweet	de-news
in-domain	99.93 _{0.06}	96.17 _{1.10}	72.91 _{0.67}	77.55 _{0.64}	85.67 _{0.80}	77.14 _{0.69}	88.35 _{0.33}
multi-task	99.85 _{0.06}	96.60 _{0.07}	84.17 _{2.07}	78.28 _{1.54}	85.45 _{0.34}	74.88 _{1.11}	88.01 _{0.42}
Best	in-domain	multi-task	multi-task	seq-bar-wiki	seq-de-tweet	seq-bar-wiki	seq-de-tweet
Model	99.93 _{0.06}	96.60 _{0.07}	84.17 _{2.07}	79.27 _{1.33}	86.08 _{1.14}	77.26 _{0.83}	88.89 _{0.09}
Improvement	/	+0.43	+11.26	+1.72	+0.41	+0.12	+0.54

Table 9: Performances on binary dialect identification (DID) between Bavarian and German, with multi-task learning (MTL) combining 5 NER and 2 DID (sub-)datasets, and our best setups (3-run avg_{std}).

but not on wiki DID or German NER.

Lastly, we summarize the best-performing models throughout §5-§6 on the seven tasks at the bottom of Table 9. Results show that MTL scores state-of-the-art on tweet DID and *bar-wiki* NER; the latter improvement is particularly rewarding. Still, most tasks, i.e., NER on the other four (sub-)datasets, achieve the best performances via first training on an external dataset, *bar-wiki* or *de-tweet*, and then on the target. We evince the advantages of data diversity in dialects, genres, and topics on dialect identification and named entity recognition.

7. Error Analysis

We perform error analyses on BARNER tagging to interpret how sequential, joint, and multi-task learning models improve over the in-domain baseline. Table 10 presents NER tagging errors from four aspects: orthography and word choices of Bavarian proper and common nouns, guidelines regarding emojis and hashtags, and faithfulness to the CoNLL 2006 rules.



Example	Trans/Expl	base	seq	joint	mtl
Bavarian: proper nouns					
[Traunviatl] _{LOC}	(district name)	×	×	✓	×
[Nordkar] _{LOC}	(a ski resort)	×	×	✓	✓
[Wastl] _{PER}	Sebastian	×	×	✓	✓
[Bechtolsheim Andy] _{PER}	Andy Bechtolsheim	✓	✓	✓	×
Bavarian: common nouns					
Haisl	house	×	×	✓	✓
Bazi	rascal	×	×	×	✓
Guideline: emoji & hashtags in tweets					
 _{LOC}	(Germany flag)	×	✓	✓	✓
[#minga] _{LOC}	#munich	×	×	✓	×
Guideline: faithful to CoNLL 2006					
Eiropa-Zentrale	Europe-center	×	×	✓	✓
[Prommenade] _{LOC} 23	(street name)	×	×	✓	✓
'Dr. [Karl Ritter von Görner] _{PER} '	(person name)	×	×	✓	✓
"[Eazhezogtum Owaöstareich ob da Enns] _{ORG} "	Archduchy of Austria above the Enns	×	✓	✓	✓

Table 10: NER error analysis: examples of gold entity annotations and whether our in-domain baseline (base), sequential (seq), joint, and multi-task (mtl) models tag them correctly (yes ✓ or no ×).

BARNER includes many new proper names such

as regional locations in Bavarian or Austria (*Nordkar* and *Traunviatl*), person names (*Verdi*), and the tradition of spelling the family name first (*Bechtolsheim* is the family name in *Bechtolsheim Andy*) opposite from German or English. Unseen spellings of common nouns (*Haisl* 'house' or *Bazi* 'rascal') also result in model errors wrongly predicting them as entities. Unlike English, all German nouns, including common nouns, are capitalized. This makes distinguishing common and proper nouns more difficult than in English. Additionally, tweet texts are more colloquial, and common and proper nouns are frequently written in lowercase. As a result, the base and sequential models perform similarly badly in these two scenarios. Models improve on Bavarian with additional knowledge from joint and multi-task training.

Incorporating more similarly annotated (sub-)corpora also helps adhere to the CoNLL 2006 guideline. In tweets, we found the joint and multi-task models more successful in tagging emojis () and hashtags (*#minga*). Across both genres, they are more faithful in excluding nominal compounds containing partly NEs (*Eiropa-Zentrale*), house numbers after street names (*Prommenade 23*), titles before person names (*Dr. Karl Ritter Görner*), but including post-nominal prepositional phrases that specify the entity (*Eazhezogtum Owaöstareich ob da Enns* 'en. Archduchy of Austria above the Enns'). Owing to the multilingual language model and vast German training data, Bavarian NER can achieve high 70s to low 80s performance, not much lower than on the German datasets.

However, lexical ambiguity is an unsolved issue for NER. The best-performing model is still confused with polysemes such as *Google*. For example, *Google* can refer to an office location (LOC), a company (ORG), or a web search product (MISC). Similarly, whether *#zoom* refers to the online conferencing platform (MISC) or the company (ORG) is hard to distinguish, particularly with limited context. The large portion of double-annotated documents in BARNER can empower future studies addressing the interplay between annotators' and models' disagreements.

Lastly, we perform error analyses on wiki and tweet dialect identification. All falsely classified

cases in wiki are short, i.e., titles or phrases, which provide minimal context and could be ambiguous even for human annotators. In addition to short sentences, tweet DID data exhibits more code-switching and challenges model performances. On one side, the sampled in-house German tweets contain few examples with individual Bavarian terms. Conversely, some *bar*-labeled sentences minimally contain Bavarian. For example, only ‘*Foisch*’ (‘wrong’, *Falsch* in German) is dialectal in ‘*Lektion 3 Foisch: Da kann ich nichts machen*’ (‘Lesson 3 wrong: There’s nothing I can do about it’). This reveals the limitations of binary sentence-level DID. Moreover, our tweet DID annotation guideline includes predominately Bavarian sentences as *bar* to upscale data coverage for NER but decreases DID performance on tweets. More granular token-level annotations on code-mixing and NEs (Solorio et al., 2014) could be a promising future direction.

8. Conclusion

This paper presents BARNER, a medium-sized, manually annotated named entity corpus for Bavarian Wikipedia and tweets. We show quantitatively and qualitatively lexical and entity-level distinctions between German and Bavarian and how they affect NER performance. Our observations reveal the necessity to call for more dialectal datasets. In addition to in-domain experiments, we establish state-of-the-art results on BARNER by incorporating German NER corpora through sequential and joint training and multi-task learning with Bavarian-German dialect identification. Moreover, we also observe German performance gains by integrating diversity in domains – genres, topics, and dialects.

Future research directions include aligning between mainstream languages and local dialects, studying fine-grained sub-dialectal (sub-regional) variations, and comparing translation-based and transfer-based approaches to dialectal NLP. Syntactic and discourse annotations on Bavarian and other dialects are also along the line (Blaschke et al., 2024). Our dialectal NER and DID annotations can further benefit spoken corpora, for example, identifying NEs and classifying dialect regions in transcriptions of spoken data, e.g., whether and which part(s) of an interview involves dialectal speakers. All in all, we sincerely hope our study inspires more work on providing low-resource dialectal corpora in the future (Blaschke et al., 2023) and quantifying their distinctiveness from higher-resourced standard language corpora.

Limitation

Firstly, we acknowledge that accessing previously existing and future tweet data from Twitter (X) is get-

ting increasingly challenging and unfavorable. However, X is a trendy social media platform with a large sample of colloquial and user-generated Bavarian texts complementing well-edited Wikipedia data. We will ensure we safely store raw and annotated data and distribute them properly to researchers.

Secondly, we did not collect full metadata informing the sub-variety of these Bavarian texts. Unfortunately, location-related metadata for tweets is no longer accessible to us, and such information is frequently missing from Wikipedia articles. We also did not collect demographic metadata such as age and gender of users. However, by inspecting a small sample, we hypothesize that the most represented dialect is Central Bavarian for both *bar-wiki* and *bar-tweet*.

Thirdly, manual dialect classification of German tweets is challenging. Tweets can be short in length and frequently code-mixed between English, standard German, Bavarian, and other dialects. Moreover, sub-varieties of German dialects can be vastly different, and one annotator may not be capable of identifying the tweet’s exact dialect label. For example, a Central Bavarian speaker can judge a tweet as non-Central-Bavarian and non-standard-German. However, they might be unable to tell if the tweet is in, e.g., a North Bavarian dialect or a neighboring East Franconian one. For these reasons, we decided to use a simplistic approach to classify each tweet sentence as either mostly German, mostly Bavarian, mostly other languages or dialects, or unable to tell – since dialect classification was mainly implemented to filter out purely standard German and other dialect tweets. Nevertheless, some BARNER tweets are dialect-classified by multiple annotators, and such dialect label variations could benefit future work.

Acknowledgements

We thank Mike Zhang, Rob van der Goot and Elisa Bassignana for giving feedback on earlier drafts of this paper. This work is supported by ERC Consolidator Grant DIALECT 101043235.

9. Bibliographical References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata

- Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajudeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. [MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. [Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task](#). In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.
- Sajawel Ahmed, Manuel Stoeckel, Christine Driller, Adrian Pachzelt, and Alexander Mehler. 2019. [BIOfid dataset: Publishing a German gold standard for named entity recognition in historical biodiversity literature](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 871–880, Hong Kong, China. Association for Computational Linguistics.
- Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. [Group formation in large social networks: Membership, growth, and evolution](#). In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, page 44–54, New York, NY, USA. Association for Computing Machinery.
- Krešimir Baksa, Dino Dolovic, Goran Glavaš, and Jan Šnajder. 2014. [Named entity recognition in croatian tweets](#). In *Ninth Language Technologies Conference, Information Society (IS-JT 2014)*, pages 85–89.
- Krešimir Baksa, Dino Golović, Goran Glavaš, and Jan Šnajder. 2017. [Tagging named entities in croatian tweets](#). *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 4(1):20–41.
- Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. [Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition](#). In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. [NoSta-D named entity annotation for German: Guidelines and dataset](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Niamh Ní Bhroin. 2015. [Social Media-Innovation: The Case of Indigenous Tweets](#). *The Journal of Media Innovations*, 2(1):89–106. Number: 1.
- Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. [Maibaam: A multi-dialectal bavarian universal dependency treebank](#).
- Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. [A survey of corpora for Germanic low-resource languages and dialects](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.
- Lars Bülow, Philip C. Vergeiner, and Stephan Elspaß. 2021. [Structures of adnominal possession in austria's traditional dialects: Variation and change](#). *Journal of Linguistic Geography*, 9(2):69–85.
- Buse Çarık and Reyhan Yeniterzi. 2022. [A Twitter corpus for named entity recognition in Turkish](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4546–4551, Marseille, France. European Language Resources Association.
- Branden Chan, Stefan Schweter, and Timo Möller. 2020. [German's next language model](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. [SemEval-2022 task 8: Multilingual news article](#)

- similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.
- 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.
- Abdelhalim Hafedh Dahou and Mohamed Amine Cheragui. 2023. [Dzner: A large algerian named entity recognition dataset](#). *Natural Language Processing Journal*, 3:100005.
- Kareem Darwish. 2013. [Named entity recognition using cross-lingual resources: Arabic as an example](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567, Sofia, Bulgaria. Association for Computational Linguistics.
- Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. [Broad Twitter corpus: A diverse named entity recognition resource](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.
- Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. [Results of the WNUT2017 shared task on novel and emerging entity recognition](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.
- Stefanie Dipper, Anke Lüdeling, and Marc Reznicek. 2013. Nosta-d: A corpus of german non-standard varieties. *Non-standard data sources in corpus-based research 5*, pages 69–76.
- Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of the 10th Conference on Natural Language Processing (KONVENS 2010)*.
- Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. [SemEval-2023 task 2: Fine-grained multilingual named entity recognition \(MultiCoNER 2\)](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2247–2265, Toronto, Canada. Association for Computational Linguistics.
- Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. [Annotating named entities in Twitter data with crowdsourcing](#). In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 80–88, Los Angeles. Association for Computational Linguistics.
- Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G. Moreno, and Antoine Doucet. 2021. [A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’21*, page 2328–2334, New York, NY, USA. Association for Computing Machinery.
- Leonhard Hennig, Phuc Tran Truong, and Aleksandra Gabryszak. 2021. [MOBIE: A German dataset for named entity recognition, entity linking and relation extraction in the mobility domain](#). In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 223–227, Düsseldorf, Germany. KONVENS 2021 Organizers.
- Robert Hinderling. 1984. Bairisch: Sprache oder Dialekt? [Bavarian: language or dialect?] In R. Harnisch. *Jahrbuch der Johann-Andreas-Schmeller-Gesellschaft 1983*, pages 47–64.
- Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. [Wojood: Nested Arabic named entity corpus and recognition using BERT](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3626–3636, Marseille, France. European Language Resources Association.
- Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. [Annotating the Tweepbank corpus on named entity recognition and building NLP models for social media analysis](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7199–7208, Marseille, France. European Language Resources Association.
- Dilek Küçük and Fazli Can. 2019. [A tweet dataset annotated for named entity recognition and stance detection](#).
- Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. [A dataset of German legal documents for named entity recognition](#). In *Proceedings of the Twelfth Language Resources and*

- Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.
- Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. [L3Cube-MahaNER: A Marathi named entity recognition dataset and BERT models](#). In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 29–34, Marseille, France. European Language Resources Association.
- Shuheng Liu and Alan Ritter. 2023. [Do CoNLL-2003 named entity taggers still work well in 2023?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8254–8271, Toronto, Canada. Association for Computational Linguistics.
- Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. [Parsing tweets into Universal Dependencies](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.
- Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. [Crossner: Evaluating cross-domain named entity recognition](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460.
- Nikola Ljubešić, Tomaž Erjavec, Vuk Batanović, Maja Miličević, and Tanja Samardžić. 2019. [Croatian twitter training corpus ReLDI-NormTagNER-hr 2.1](#). Slovenian language resource repository CLARIN.SI.
- Cédric Lopez, Ioannis Partalas, Georgios Balikas, Nadia Derbas, Amélie Martin, Coralie Reutenauer, Frédérique Segond, and Massih-Reza Amini. 2017. Cap 2017 challenge: Twitter named entity recognition. *arXiv preprint arXiv:1707.07568*.
- Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. [SemEval-2022 task 11: Multilingual complex named entity recognition \(MultiCoNER\)](#). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.
- Ludwig Merkle. 1993. *Bairische Grammatik*, 5th edition. Heinrich Hugendubel Verlag, Munich.
- Hanane Nour Moussa and Asmaa Mourhir. 2023. [Darnercorp: An annotated named entity recognition dataset in the moroccan dialect](#). *Data in Brief*, 48:109234.
- Clemens Neudecker. 2016. [An open corpus for named entity recognition in historic newspapers](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4348–4352, Portorož, Slovenia. European Language Resources Association (ELRA).
- Siim Orasmaa, Kadri Muischnek, Kristjan Poska, and Anna Edela. 2022. [Named entity recognition in Estonian 19th century parish court records](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5304–5313, Marseille, France. European Language Resources Association.
- 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.
- Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2022. [AsNER - annotated dataset and baseline for Assamese named entity recognition](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6571–6577, Marseille, France. European Language Resources Association.
- Uyen Phan, Phuong N.V Nguyen, and Nhung Nguyen. 2022. [A named entity recognition corpus for Vietnamese biomedical texts to support tuberculosis treatment](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3601–3609, Marseille, France. European Language Resources Association.
- Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. [The first cross-lingual challenge on recognition, normalization, and matching of named entities in Slavic languages](#). In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 76–85, Valencia, Spain. Association for Computational Linguistics.
- Barbara Plank. 2022. [Sliced at SemEval-2022 task 11: Bigger, better? massively multilingual LMs for multilingual complex NER on an](#)

- academic GPU budget. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1494–1500, Seattle, United States. Association for Computational Linguistics.
- Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. [DaN+: Danish nested named entities and lexical normalization](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Thomas Proisl and Peter Uhrig. 2016. [SoMaJo: State-of-the-art tokenization for German web and social media texts](#). In *Proceedings of the 10th Web as Corpus Workshop*, pages 57–62, Berlin. Association for Computational Linguistics.
- Nils Reimers, Judith ECKLE-KOHLER, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. Germeval-2014: Nested named entity recognition with neural networks. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 117–120.
- Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. [Named entity recognition in tweets: An experimental study](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Anthony R. Rowley. 2011. Bavarian: Successful dialect or failed language? In Joshua Fishman and Ofelia Garcia, editors, *Handbook of Language and Ethnic Identity*, volume 2 (The Success-Failure Continuum in Language and Ethnic Identity Efforts), pages 299–309. Oxford University Press.
- Josef Ruppenhofer, Ines Rehbein, and Carolina Flinz. 2020. [Fine-grained named entity annotations for German biographic interviews](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4605–4614, Marseille, France. European Language Resources Association.
- Kevin P. Scannell. 2022. [Managing Data from Social Media: The Indigenous Tweets Project](#). In *The Open Handbook of Linguistic Data Management*. The MIT Press.
- Martin Schiersch, Veselina Mironova, Maximilian Schmitt, Philippe Thomas, Aleksandra Gabryszak, and Leonhard Hennig. 2018. [A German corpus for fine-grained named entity recognition and relation extraction of traffic and industry events](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Anil Kumar Singh. 2008. [Named entity recognition for south and south East Asian languages: Taking stock](#). In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.
- Tamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. [Overview for the first shared task on language identification in code-switched data](#). In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.
- Axel Sorensen, Siyao Peng, Barbara Plank, and Rob van der Goot. 2024. [Eevee: An easy annotation tool for natural language processing](#).
- Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. [Results of the WNUT16 named entity recognition shared task](#). In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Samia Touileb. 2022. [NERDz: A preliminary dataset of named entities for Algerian](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 95–101, Online only. Association for Computational Linguistics.
- Asahi Ushio, Francesco Barbieri, Vitor Sousa, Leonardo Neves, and Jose Camacho-Collados. 2022. [Named entity recognition in Twitter: A dataset and analysis on short-term temporal shifts](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 309–319, Online only. Association for Computational Linguistics.

- Rob van der Goot. 2022. [MaChAmp at SemEval-2022 tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task multi-lingual learning for a pre-selected set of semantic datasets](#). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1695–1703, Seattle, United States. Association for Computational Linguistics.
- Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. [Massive choice, ample tasks \(MaChAmp\): A toolkit for multi-task learning in NLP](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.
- Helmut Weiß. 1998. *Syntax des Bairischen*. Max Niemeyer Verlag.
- Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. [KazNERD: Kazakh named entity recognition dataset](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 417–426, Marseille, France. European Language Resources Association.
- Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aeppli. 2017. [Findings of the VarDial evaluation campaign 2017](#). In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. [Language identification and morphosyntactic tagging: The second VarDial evaluation campaign](#). In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. [A report on the third VarDial evaluation campaign](#). In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16, Ann Arbor, Michigan. Association for Computational Linguistics.
- Ayah Zirikly and Mona Diab. 2014. [Named entity recognition system for dialectal Arabic](#). In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 78–86, Doha, Qatar. Association for Computational Linguistics.

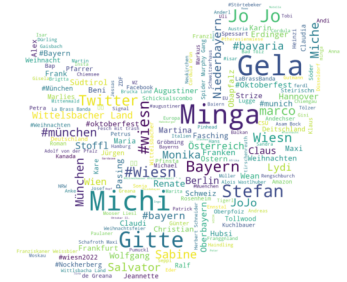
A. NE Word Clouds



(a) bar-wiki



(b) de-wiki



(c) bar-tweet



(d) de-tweet



(e) de-news

B. Data Statement for BARNER

We follow [Bender and Friedman \(2018\)](#) to include a data statement to alleviate potential ethical issues raised during BARNER data collection.

Header

- *Dataset Title:* BARNER
- *Dataset Curator(s):* Coordinators of this project, Siyao Peng, Verena Blaschke, Ekaterina Artemova and Barbara Plank, are academic employees at LMU Munich. Our three annotators, Zihang Sun, Huangyan Shan, and Marie Kolm, are master students at LMU Munich.
- *Dataset Version:* V1.0 as of March 19, 2024
- *Dataset Citation:* BARNER should be cited by citing this publication.
- *Data Statement Authors:* Same as the authors of this publication.
- *Data Statement Version:* V1.0 as of March 19, 2024
- *Data Statement Citation and DOI:* To cite this data statement, please cite this publication.
- *Links to versions of this data statement in other languages:* None

Executive Summary BARNER is the first named entity (NE) corpus for Bavarian German manually annotated by three annotators. The corpus includes 161K+ tokens and 6.6K+ entities balanced across two genres: Wikipedia articles and Twitter (X). We annotate both coarse-grained person (PER), location (LOC), organization (ORG), and miscellaneous (MISC) NEs following CoNLL 2006 German guidelines ([Tjong Kim Sang and De Meulder, 2003](#)), and fine-grained derived and partially contained NEs adapted from GermEval 2014 ([Reimers et al., 2014](#)). Half of BARNER is double-annotated with 83+ inter-annotator agreement on typed F1.

Curation Rationale The crucial research question behind BARNER and this paper is to examine whether and to what extent named entities surface differently between standard and dialectal language variations. We take Bavarian German as the dialectal counterpart and compare it with mainstream (high) German. We also contrast NEs in two representative genres: iteratively-refined Wikipedia texts versus spontaneously-iterated tweets. NE annotations follow the conventional BIO-encoding at the token level in both genres. Nevertheless, Wikipedia

Figure 3: NE word clouds of the five datasets.

articles are selections of continuous segments averaging over 1k+ tokens per document, whereas tweets are naturally much shorter instances in tokens.

Documentation for Source Datasets BARNER is directly sourced from raw Wikipedia and Twitter texts, not built on a pre-existing corpus.

Language Varieties Our dataset focuses on the Bavarian dialectal variant of German, i.e., ISO: 639-3; code: *bar*; referred to as *Bairisch* in standard German and *Boarisch* in Bavarian German. Even though we did not collect sub-dialectal metadata for BARNER, we can still claim that the most represented sub-dialect is Central Bavarian by inspecting a small sample.

Speaker Demographic We could not obtain full demographic information regarding Wikipedia editors and tweet users.

Annotator Demographic All three annotators are master students in their 20s, one male and two female. One annotator is a native Bavarian speaker, and the other two are native Mandarin Chinese speakers who majored in German during their bachelor studies and thus have full professional proficiency in German.

Speech Situation and Text Characteristics Wikipedia articles and tweets were collected between February and May 2023. Both are asynchronously written data directed at the general public. However, Wikipedia is more scripted, whereas Twitter simulates more spontaneous speech. Topic interests of the annotators contributed to the selection of Wikipedia articles. No topic restriction was imposed while scrapping Twitter posts.

Preprocessing and Data Formatting Twitter data were scrapped using Twitter API¹¹ conditioned on the list of Bavarian-speaking users, and Wikipedia texts are manually copy-pasted from webpages containing continuous text segments starting from the beginning of the page. All texts are automatically tokenized using SoMaJo's (Proisl and Uhrig, 2016) *de_CMC* model. Annotations are released in the CoNLL-styled tab-separated (*tsv*) format where a line is 1) a token and its tab-joined BIO-encoded NE tag, 2) hashtag-initial metadata information, or 3) a blank line separating sentences. Most NE annotations are done using local text editors except for the last two weeks, that was piloted on a newly released Eevee annotation

¹¹<https://developer.twitter.com/en/docs/twitter-api>

tool¹² (Sorensen et al., 2024). For Twitter data, we anonymize mentions to @Mention but allow NE annotations on #Hashtags and emojis.

Capture Quality We ensure the dialectal authenticity of *bar-tweet* by asking annotators to label each tweet sentence whether they are mostly Bavarian, German, another language or dialect, or unintelligible. We only keep mostly Bavarian tweets in our data release.

Limitations See the Limitation section of this paper.

Metadata

- *License*: CC-BY 4.0.¹³
- *Annotation Guidelines*: https://github.com/mainlp/BarNER/blob/main/MainLP_NER_Annotation_Guidelines.pdf
- *Annotation Process*: The annotators were hired and compensated for their work following national salary rates.
- *Dataset Quality Metrics*: Cohen's kappa.
- *Errata*: None so far. Please report errors by contacting the authors or opening an issue at <https://github.com/mainlp/BarNER/>.

Disclosures and Ethical Review This work is supported by ERC Consolidator Grant DIALECT 101043235.

Other None.

Glossary

- PER: Person
- LOC: Location
- ORG: Organization
- MISC: Miscellaneous
- LANG: Language
- RELIGION: Religion
- EVENT: Event
- WOA: work-of-art
- -part: partly containing a nominal NE
- -deriv: morphologically derived from a nominal NE

¹²<https://axelsorensen.github.io/EeveeTest/>

¹³<https://creativecommons.org/licenses/by/4.0/deed.en>