BERnaT: Basque Encoders for Representing Natural Textual Diversity

BERnaT: Modelos Discriminativos en Euskera para Representar la Diversidad del Lenguaje Natural

Ekhi Azurmendi, Joseba Fernandez de Landa, Jaione Bengoetxea, Maite Heredia, Julen Etxaniz, Mikel Zubillaga, Ander Soraluze, Aitor Soroa
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{ekhi.azurmendi,joseba.fernandezdelanda,jaione.bengoetxea,
maite.heredia,julen.etxaniz,mikel.zubillaga,ander.soraluze,asoroa}@ehu.eus

Abstract: Language models depend on massive text corpora that are often filtered for quality, a process that can unintentionally exclude non-standard linguistic varieties and reinforce representational biases. In this paper, we argue that language models should aim to capture the full spectrum of language variation (dialectal, historical, informal, etc.) rather than relying solely on standardized text. Focusing on the Basque language, we construct new corpora combining standard, social media, and historical sources, and pretrain the BERnaT family of encoder-only models in three configurations: standard, diverse, and combined. We further propose an evaluation framework that separates Natural Language Understanding (NLU) tasks into standard and diverse subsets to assess linguistic generalization. Results show that models trained on both standard and diverse data often outperform those trained on standard corpora, improving performance across task types without compromising standard benchmark accuracy. These findings highlight the importance of linguistic diversity in building inclusive, generalizable language models.
Keywords: Training, Low-Resource Languages, Corpus, Linguistic Diversity, Evaluation

Resumen: Los modelos de lenguaje dependen de enormes corpus textuales que se suelen someter a un filtrado de calidad que puede conllevar la exclusión de variedades lingüísticas no estándar y el refuerzo de sesgos de representatividad. Los modelos deberían aspirar a la representación de todo el espectro de variación lingüística, en lugar de limitarse al texto estandarizado. Utilizamos el euskera como caso de estudio y creamos la familia de modelos discriminativos BERnaT, entrenada con fuentes de mayor diversidad: redes sociales y textos históricos. El preentrenamiento de los modelos se realiza con tres combinaciones de textos: estándar, diversos y combinados. Nuestra propuesta de evaluación consiste en separar las tareas entre subconjuntos de lenguaje estándar y no estándar, para poder medir la generalización lingüística de los modelos. Los resultados demuestran que los modelos entrenados con una mayor diversidad lingüística tienden a superar a aquellos basados solo en texto estándar. Así, es posible mejorar la generalización lingüística en muchas tareas sin perder precisión en las evaluaciones más tradicionales. Esto evidencia la importancia de la diversidad lingüística para lograr modelos más inclusivos con una mayor capacidad de generalización.
Palabras clave: Entrenamiento, Idiomas con pocos recursos, Corpus, Diversidad Lingüistica, Evaluación

1 *Introduction*

The development of large-scale language models requires large quantities of textual data that is often crawled from the web and subsequently filtered to ensure its quality. The amount of text that is left out as well as its characteristics will vary according to various factors, such as the target domain or language, as well as to differences in the beliefs of authors as to what constitutes noise (Laurençon et al., 2022; Etxaniz et al., 2024). Unwanted text may include code, text in languages other than the target language, or pure gibberish.

In the process of corpus creation and filtering, some authors may opt to exclude language varieties that, although representative of natural language use, do not meet their standards of quality - either intentionally, or by default through the selection of a reliably high-quality subset of the data (Wenzek et al., 2020; Artetxe et al., 2022; Soldaini et al., 2024; Palomar-Giner et al., 2024; Etxaniz et al., 2024). In contrast, in this paper we claim that any corpus gathered with the final objective of training a foundational model should try to capture the full diversity of language, including variation across dialects (diatopic), registers (diaphasic), social groups (diastratic), and even time (diachronic). Modeling this diversity can ensure that the resulting model will be capable of handling a broader range of domains and language varieties.

In practice, it has been proven that models trained with less diverse data suffer from representation biases and problems in performance (Blodgett, Green, and O'Connor, 2016). Even if this is not intended, a selection of texts to train a model or a quality filter performed without proper care is likely to be skewed towards powerful groups (Gururangan et al., 2022). These studies are mostly centered around English and other high-resource languages. Conversely, in this work we focus on Basque, a low-resource and morphologically rich language, and conduct our experiments using encoder-based architectures. Our research's aim is therefore to contribute to the body of resources for the Basque language and develop new monolingual models designed for specific Natural Language Understanding (NLU) tasks.

As shown in Figure 1, we use three different corpora combinations to pretrain a collection of models. The BERnaT$_{\text{standard}}$ models are pretrained using a widely used standard corpus, whereas BERnaT$_{\text{diverse}}$ models use a newly introduced diverse corpus that covers different varieties of Basque. Finally, the BERnaT models are pretrained with both the standard and the diverse corpora.

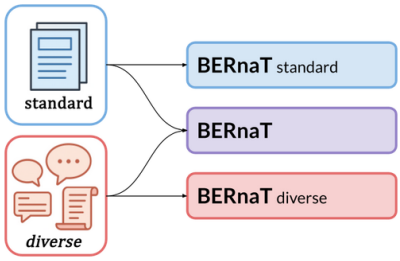

Figure 1: Summary of corpora combinations used to train BERnaT models. *Standard Corpus*: high-quality standard Basque from sources like News or Wikipedia. *Diverse Corpora*: social media and historical texts capturing informal, dialectal, and pre-standard Basque.

We propose a new configuration of NLU tasks to evaluate BERnaT models, splitting existing benchmark tasks into standard and diverse subsets depending on the register and source of the annotated data, thus emulating the standard/diverse separation. This evaluation framework is designed to facilitate the assessment of model performance under language diversity conditions, enabling a more comprehensive understanding of how well the models generalize across different varieties and linguistic contexts.

The following are the principal findings from our experiments, which may help dataset creation and model development for other languages:

- The inclusion of non-standard text in model pretraining is helpful for NLU tasks that deal with non-standard language, such as text from social media, dialectal variations and historical texts.
- The performance gain of combined BERnaT models is more pronounced when the data used for fine-tuning is scarce.
- Including diverse text doesn't degrade the results of traditional NLU tasks written in standard varieties (BasqueGLUE, NLI, POS).
- Even when including diverse corpora in training, models perform worse in diverse

evaluations, showing that these tasks remain more challenging.

- Model size has an impact on performance, with smaller models achieving better results if specialized either on standard or diverse texts, and larger models obtaining better when they combine both.

Beyond the insights previously discussed, this work contributes several resources[1] that support future research in Basque language technology: (i) the new family of BERnaT encoder models, available in three sizes and three training configurations, designed to process linguistically diverse text; (ii) the largest diverse language corpus for Basque, containing both contemporary social media and historical texts; and (iii) a comprehensive suite of benchmark datasets, comprising both established and new tasks, categorized into standard and diverse linguistic domains.

2 Related work

Approaches to language diversity. The exclusion of varieties outside the standard can be motivated by underlying language ideologies (Walsh, 2021), which shape researchers' beliefs about which varieties are suitable for corpus curation; or the convenience of a more standardized and homogeneous language in language modeling and evaluation (Jørgensen, Hovy, and Søgaard, 2015). Nonetheless, it often leads to misrepresentation of social groups outside of the mainstream (Gururangan et al., 2022; Blodgett et al., 2020) and worse performance and ability to generalize to unseen language varieties (Markl and McNulty, 2022; Bengoetxea, Gonzalez-Dios, and Agerri, 2025; Srirag, Sahoo, and Joshi, 2025). When faced with this challenge, previous approaches may include the *normalization* of text before its processing, or domain adaptation of models (Han and Baldwin, 2011; Eisenstein, 2013; Kuparinen, Miletić, and Scherrer, 2023; Alhafni et al., 2024).

Basque language diversity and standardization. In the process of gathering a diverse corpus, we have identified sources which include texts characterized by diatopic and diaphasic variability, that is, different dialects and registers, as well as a wide range of topics and domains. Our main sources of non-standard language include a) texts written before the development of *Euskara Batua*, the standardized form of the Basque language (Hualde and Zuazo, 2007; Oñederra, 2016), that use non-standard dialectal orthography (Soraluze Irureta et al., 2019); and b) more recent texts gathered from social media, which include innovative spellings as well as a more conversational and relaxed register (Elordui and Aiestaran, 2022; Fernandez de Landa et al., 2024).

Encoder models and corpora. While decoder-only LLMs are popular due to their text generation capabilities, encoder-only models are still frequently used for classification, clustering and retrieval, even beating decoders that are much larger (Weller et al., 2025; Gisserot-Boukhlef et al., 2025). The majority of encoder models have been trained on standard corpora, such as Wikipedia or journalistic texts, both in English (Liu et al., 2019) and multilingual settings (Conneau et al., 2020). However, several efforts have also focused on training encoder models on diverse corpora from social media in English (Nguyen, Vu, and Tuan Nguyen, 2020) and other languages (Barbieri, Espinosa Anke, and Camacho-Collados, 2022), and on evaluating them on that specific source (Barbieri et al., 2020). Recently, works have centered on training encoder models with modern techniques and corpora scales closer to LLMs for English (Warner et al., 2024), and the same approach has been extended to cover around 1.8K languages (Marone et al., 2025). However, these models have been mostly trained on standard filtered texts, with no special focus on training or evaluating on diverse corpora.

Basque encoder models and corpora. Current SoTA Basque monolingual encoders such as RoBERTa EusCrawl (Artetxe et al., 2022) and ElhBERTeu (Urbizu et al., 2022) were mostly trained on standard text that mostly comes from news websites and Wikipedia. (Artetxe et al., 2022) found that using other lower-quality corpora, such as mC4 (Xue et al., 2021) and CC-100 (Conneau et al., 2020), provided similar results compared to Euscrawl. Recently, bigger corpora such as Latxa corpus (Etxaniz et al., 2024) and ZelaiHandi (San Vicente et al., 2024) have been introduced, but no encoder models have

[1] Code: https://github.com/hitz-zentroa/BERnaT Resources: https://hf.co/collections/HiTZ/bernat

Source	Docs	Words	Diversity
Wikipedia	409k	51M	$0.1e^{-2}{\pm}0.2e^{-1}$
Egunkaria	176k	39M	$0.3e^{-2}{\pm}0.2e^{-1}$
EusCrawl-v1.1	1.79M	359M	$0.4e^{-2}{\pm}0.4e^{-1}$
Colossal-oscar	234k	105M	$1.8e^{-2}{\pm}0.8e^{-1}$
CulturaX	1.31M	541M	$1.9e^{-2}{\pm}0.8e^{-1}$
HPLT-v1	375k	120M	$3.0e^{-2}{\pm}1.1e^{-1}$
Booktegi	166	3M	$7.3e^{-2}{\pm}1.3e^{-1}$
BSM [new]	13k	188M	$1.4e^{-1}{\pm}1.7e^{-1}$
EKC [new]	338	21M	$7.3e^{-1}{\pm}1.7e^{-1}$

Table 1: Corpora sizes and diversity analysis for *latxa-corpus-v1.1* (upper part) and our proposed *Diverse Corpora* (bottom part). Diversity is calculated per document as described in Section 3.3, and we also include the standard deviation.

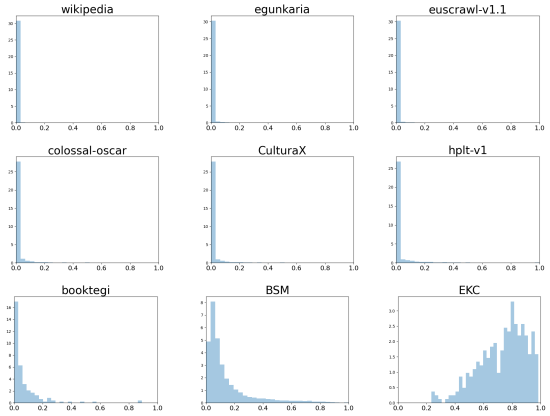

Figure 2: Visualization of diversity distribution for latxa standard corpora (*wikipedia, egunkaria, euscrawl-v1.1, colossal-oscar, CulturaX, hplt-v1, booktegi*) and newly added *BSM* and *EKC* non-standard corpora.

been trained on these. These corpora, despite having more diverse sources, still lack language diversity, which is very minimal as shown in Section 3. In contrast, our new corpora include a larger number of diverse texts.

3 Training Corpora

To represent diverse language usage in Basque, we combine an existing large-scale standard corpus (*latxa-corpus-v1.1*) with a newly curated corpus called *Diverse Corpora*, which is composed of varied and non-standard collections of texts (c.f. Section 3.2). This approach allows us to capture both formal, well-edited language and informal, dialectal, or historical varieties. We perform an empirical analysis of the selected corpora, quantifying their linguistic diversity and confirming their alignment with the intended categories, therefore ensuring that our pretraining data reflects a broad spectrum of Basque language use.

3.1 Standard Language Corpus

We utilize the largest standard Basque corpus currently available, *latxa-corpus-v1.1* (Etxaniz et al., 2024). The corpus comprises 4.3M documents in standard Basque, summing up to 1.22B words. Its sources include EusCrawl-v1.1, extracted using ad-hoc scrapers (Artetxe et al., 2022), as well as Egunkaria, containing content from the daily Basque newspaper Euskaldunon Egunkaria (2001–2006), and Booktegi, which provides a collection of literature books in Basque. Additional sources include Wikipedia, as well as datasets derived from Common Crawl, namely CulturaX (Nguyen et al., 2024), Colossal OSCAR (Abadji et al., 2022), and HPLT v1 (de Gibert et al., 2024). To ensure linguistic consistency and quality, *latxa-corpus-v1.1* was systematically normalized, deduplicated, and exhaustively filtered (Etxaniz et al., 2024).

3.2 Diverse Corpora

We create *Diverse Corpora* by combining text from two linguistically rich and varied sources: social media and historical texts. Social media captures informal, spontaneous, and up-to-date language use, including dialectal variation, slang, and code-switching. Historical texts, in turn, provide access to pre-standard Basque, which is highly rich in dialectal diversity. *Diverse Corpora* represents the largest and most diverse non-standard corpus constructed to date, with its overall size being half that of EusCrawl-v1.1 (refer to Table 1).

BSM: The Basque Social Media corpus (BSM) provides a valuable source of personal, spontaneous, and varied written content. It includes standard language as well as a wide range of dialectal, slang, informal, and code-switched data (Fernandez de Landa, Agerri, and Alegria, 2019). Building on the same techniques and extending previously collected resources (Fernandez de Landa et al., 2024), BSM contains approximately 11 million posts produced by more than 13,000 Basque-speaking users, amounting to around 188 million words. We perform a data augmentation step that effectively doubles the dataset size by reordering the existing material in two distinct ways. Specifically, we divide BSM into

Dataset	Task	Train/Dev	Test	Domain
BHTC [1]	Topic classification	10,442	1,854	News
Korref [2]	Coreference resolution	1,306	587	News
NERCid [3]	NERC	64,475	35,855	News
NERCod [1]	NERC	79,420	14,462	News, Wikipedia
QNLI [1]	QA/NLI	1,994	238	Wikipedia
WiC [1]	WSD	409,159	1,400	News
XNLIeu-nat [4] †	NLI	392,000	621	News
POSud [5] †	POS tagging	7,194	1,799	News, literary texts
POShis-nor [new] †	POS tagging	5,095	526	Normalized historical texts
BEC [1]	Sentiment Analysis	7,380	1,302	Twitter
Intent [6]	Intent classification	5,322	1,087	Facebook
Slot [6]	Slot filling	30,443	5,633	Facebook
Vaxx [7]	Stance detection	1,070	312	Twitter
XNLIeu-var [8] †	NLI	392,000	894	Dialectal language
POShis [new] †	POS tagging	5,095	526	Historical texts

Table 2: Dataset characteristics and sizes for Basque benchmarks. The upper block includes standard language datasets, while the lower block includes diverse language datasets (non-standard, dialectal, or historical). Datasets marked with † are not part of BasqueGLUE. Sources: [1](Urbizu et al., 2022), [2](Soraluze et al., 2012),[3](Alegría et al., 2006),[4](Heredia et al., 2024),[5](de Marneffe et al., 2021),[6](López de Lacalle, Saralegi, and San Vicente, 2020),[7](Agerri et al., 2021),[8](Bengoetxea, Gonzalez-Dios, and Agerri, 2025).

two complementary subsets containing the same textual content but differ in organization: i) BSMtime, where posts are ordered chronologically by publication time, independent of author identity (11M posts). ii) BSMauthor, where posts are grouped by author, forming 13K complete individual documents representing their timelines. This dual organization enables the analysis of both diachronic language variation and author-specific linguistic behavior, providing a flexible resource for a range of sociolinguistic and computational studies.

EKC: The Corpus of Basque classical writers (*Euskal Klasikoen Corpusa*, EKC) (Euskara Institutua, 2013) aims to serve as the repository for nearly all classical texts up to the 20th century. It contains 338 documents or books ranging from the 16th century to 1975, comprising a total of 21 million words. This corpus covers various literary genres, including poetry, narrative, theater, essays, and religious texts. The texts originate from different dialects, as the standard Basque language was not established until the late 20th century. Due to the historical span and dialectal variety represented in the collection, the corpus is expected to be highly diverse.

3.3 Language diversity analysis

We conduct a language diversity analysis on the entire pretraining data to verify that the sources align with the theoretical distinction between standard and diverse language. To that end, we employ the method presented in (Fernandez de Landa and Agerri, 2021), which automatically classifies each sentence in a document as either standard or non-standard. Using these classifications, the diversity of each document is computed as the proportion of non-standard sentences relative to the total number of sentences. Subsequently, the language diversity of each source in the corpus is determined by calculating the mean diversity across all documents within that source.

Table 1 and Fig. 2 present the results of our language diversity analysis across the standard corpora of *latxa-corpus-v1.1* corpus as well as our newly proposed *Diverse Corpora*. The results show that, on the one hand, standard language sources such as Wikipedia, Egunkaria, and EusCrawl exhibit very low levels of non-standard usage, indicating low linguistic diversity. Although CulturaX and Colossal-oscar are both large-scale multilingual corpora derived from Common Crawl, their diversity values are slightly below those of HPLT-v1, likely due to the more extensive preprocessing applied to produce cleaner, more standardized text. Booktegi, though small, contains literary genres (fiction, essays, and poetry) that are inherently diverse in syntax and vocabulary, which explains its slightly higher diversity. On the other hand, the BSM corpus, representing social media text, reaches diversity levels more than double those of the

most heterogeneous standard corpus. Finally, the EKC corpus stands out with a mean diversity of 0.733, an order of magnitude greater than any other source. This high diversity observed in the BSM and especially the EKC corpora confirms their role as essential complements to standardized resources for modeling the full spectrum of Basque linguistic variation.

4 Evaluation datasets

We use NLU tasks to assess the effect of language diversity on Basque. Hence, we evaluate the models on all tasks from BasqueGLUE (Urbizu et al., 2022), the standard benchmark for Basque NLU, as well as Natural Language Inference (NLI) and Part-of-Speech (POS) tagging tasks. We divide each dataset into *standard* and *diverse* subsets, which allows us to analyze how linguistic diversity influences model performance. Details of each dataset, including size and domain, are described in Table 2. We apply the same linguistic diversity analysis to the evaluation datasets as performed on the pretraining data; these results are detailed in Appendix A.

BasqueGLUE. This dataset collection consists of ten tasks that cover a wide spectrum of Basque NLU challenges. The benchmark includes datasets from standard language domains such as news and literature, as well as collections drawn from more diverse contexts, particularly social media, where dialectal and non-standard Basque are widely used (Fernandez de Landa, Agerri, and Alegria, 2019). The standard subset of BasqueGLUE includes Topic Classification (BHTC), Coreference Resolution (Korref), Question Answering framed as NLI (QNLI), Word Sense Disambiguation (WSD) in Context (WiC), and Named Entity Recognition/Classification in both in-domain (NERCid) and out-of-domain (NERCod) settings. The diverse subset, in contrast, covers Sentiment Analysis (BEC), Stance Detection (Vaxx), Intent Classification (Intent), and Slot Filling (Slot).

Natural Language Inference. We use the Basque NLI dataset introduced in (Heredia et al., 2024), which extends the XNLI benchmark (Conneau et al., 2018) to this language. The standard subset consists of the so-called XNLIeu-nat part of the dataset, a test set created entirely from scratch by native Basque speakers. The diverse subset consists of the NLI dataset in (Bengoetxea, Gonzalez-Dios, and Agerri, 2025), where XNLIeu-nat was adapted into three Basque dialects (Western, Central, and Navarrese), therefore allowing NLI evaluation in both standard and diverse language settings. For training, we rely on the automatically translated XNLIeu training set, derived from the English XNLI corpus, which naturally retains certain domain characteristics from the original English data.

Tokenizer	Corpora		
	Standard	**Diverse**	**Both**
$\text{Tok}_{\text{Standard}}$	1.29	1.62	1.53
$\text{Tok}_{\text{Diverse}}$	1.37	1.42	1.41
Tok_{Both}	1.30	1.45	1.41

Table 3: Token fertility for different tokenizers. Rows correspond to tokenizer's training corpora, and columns to evaluation corpora (lower is better).

POS tagging. The standard subset consists of the Universal Dependencies POS (POSud) dataset (de Marneffe et al., 2021), based on the Basque UD treebank (Aranzabe et al., 2015), which contains 8,993 sentences (121,443 tokens) drawn primarily from literary and journalistic sources. The diverse subset is POShis, a dataset derived from the BIM corpus (Estarrona et al., 2021), which comprises historical texts dating from the 15th to the mid-18th century, spanning multiple historical dialects (Western, Central, Labourdin, Souletin). We select 5,621 instances (693,414 tokens) from this morphosyntactically annotated diachronic corpus, making the resulting POShis subset approximately five times larger than POSud. The manual annotations have been converted to follow the same criteria as POSud. Additionally, we provide a parallel normalized version, POShis-nor, to facilitate evaluation under standard conditions.

5 Training Models

The main aim of the experiments is to evaluate the effect of including diverse language data in pretraining on NLU tasks. To that end, we start by pretraining the models with the corpora described in Section 3, which are then fine-tuned to each task and dataset (c.f. Section 4). In this section, we describe these steps in turn.

Pretraining. We use encoder-only models that follow the RoBERTa (Liu et al., 2019)

		BERnaT$_{\text{standard}}$			BERnaT$_{\text{diverse}}$			BERnaT		
		med	base	large	med	base	large	med	base	large
Standard tasks	BHTC (F_1)	74.85	75.37	77.53	73.37	73.97	76.70	75.06	75.56	**77.83**
	Korref (Acc.)	56.45	62.24	57.69	56.27	56.67	52.13	63.09	**64.68**	62.07
	NERCid (F_1)	83.35	83.83	**86.48**	80.19	75.50	83.34	81.33	83.39	84.97
	NERCod (F_1)	73.29	75.14	75.04	65.99	67.77	70.63	72.60	75.67	**76.06**
	QNLI (Acc.)	69.61	73.25	**74.23**	70.73	70.73	71.01	69.19	71.15	72.96
	WiC (Acc.)	67.69	70.57	70.07	68.86	70.33	70.14	69.07	69.55	**70.86**
	XNLIeu-nat (Acc.)	67.31	71.28	**74.93**	61.89	66.72	64.52	66.51	67.95	72.79
	POSud (F_1)	94.85	95.61	**96.27**	94.06	95.42	95.38	94.46	95.86	95.97
	POShis-nor (F_1)	71.48	74.29	78.92	72.73	72.55	76.96	73.10	76.12	**79.74**
	AVG standard tasks	73,21	75,73	76,79	71,57	72,18	73,42	73,82	75,55	**77,03**
Diverse tasks	BEC (F_1)	69.87	66.05	68.10	69.43	69.82	**70.10**	68.87	70.02	69.40
	Intent (F_1)	75.31	77.46	77.80	77.34	75.50	**79.70**	75.22	76.84	78.04
	Slot (F_1)	76.73	**78.95**	77.01	76.93	78.26	76.01	78.02	75.75	78.61
	Vaxx (MF_1)	61.60	63.06	65.97	64.03	66.67	**68.46**	66.13	65.77	67.44
	XNLIeu-var (Acc.)	65.21	68.05	**74.05**	58.50	65.14	63.68	62.30	64.91	72.82
	POShis (F_1)	73.06	73.96	75.84	73.22	73.17	73.27	73.00	74.36	**76.33**
	AVG diverse tasks	70.30	71.26	73.13	69.91	71.43	71.87	70.59	71.28	**73.77**
AVG overall		72.58	73.70	75.35	70.96	72.04	73.43	72.37	73.76	**76.24**

Table 4: Results by task for all BERnaT model variants and sizes, divided by standard and diverse tasks. Metrics based on the original implementation of the tasks as explained in the respective papers (F_1 = micro-average F_1-score , Acc. = Accuracy, MF_1 = macro-average F_1-score).

architecture for several model sizes: medium (51M), base (124M) and large (355M). We train the models from scratch, using three corpora combinations (standard, diverse and both) therefore generating BERnaT$_{\text{standard}}$, BERnaT$_{\text{diverse}}$ and BERnaT models, respectively (see Figure 1). We follow (Artetxe et al., 2022) and pretrain each model with a batch size of 1 million tokens and a sequence length of 512 tokens. We use packing, i.e., shorter documents are concatenated or packed into a single sequence to fully utilize the sequence length and maximize the efficiency of each training iteration. Documents that exceed the sequence length, particularly those from Booktegi and EKC, are divided into smaller chunks. We train the models up to 100 epochs and choose the best checkpoint based on the validation loss. To speed up pretraining, we use Flash-Attention 2 (Dao, 2023) and mixed-precision (FP16), as preliminary experiments have shown no difference in performance compared to full precision and regular attention implementation.

Handling diversity. During the initial phase of training the standard models, we adopted the hyperparameter configuration proposed by (Artetxe et al., 2022). Applying the same configuration to the more heterogeneous corpora, however, frequently yielded exploding gradients. This instability is likely due to the increased data diversity, and the corresponding sensitivity of the model under such a distribution. To mitigate this issue, we progressively reduced the learning rate until stable convergence was achieved. Preliminary experiments showed a correlation between corpus diversity and the learning rate required for stability: as diversity increases, smaller learning rates are needed to prevent exploding gradients. Accordingly, we identified learning rates of $8e^{-4}$, $4e^{-4}$ and $1e^{-4}$ as good candidates for stable training of medium, base, and large models, respectively. We maintain these learning rates across all BERnaT variants to enable fair comparison. For the remaining hyperparameters, we follow the configuration proposed by (Artetxe et al., 2022).

Tokenizers. To accurately represent the data on our corpora, and given that the target language is morphologically rich and exhibits substantial cross-domain lexical variation, we train a specialized BPE tokenizer for each corpus combination, with a vocabulary size of 50k (Liu et al., 2019). Table 3 reports the token fertility (Rust et al., 2021) of each tokenizer (denoted Tok$_{\text{Standard}}$, Tok$_{\text{Diverse}}$, and Tok$_{\text{Both}}$) across all corpus configurations. Token fertility measures the average number of subword tokens required to represent a single

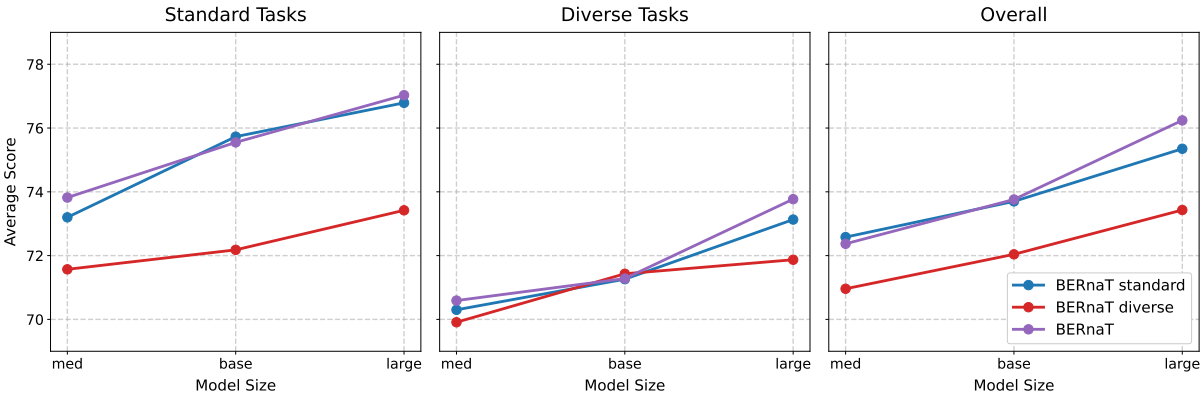

Figure 3: Average performance of Diverse, Standard and combined models in three sizes on Standard, Diverse and all tasks. X axis represents the model size and Y axis the average scores for each task group.

word, providing insight into the tokenizer's efficiency and its alignment with the linguistic properties of the data. Lower fertility values indicate more compact tokenization, reflecting a closer match between the tokenizer's learned vocabulary and the underlying lexical structure of the corpora. Table 3 shows that tokenizers trained on specific corpora obtain the lowest token fertility on that corpora, as expected, with the tokenizer trained with both standard and diverse corpora obtaining the second-best token fertility values.

Fine-tuning. For the tasks included in BasqueGLUE, we use the same training configuration and hyperparameters as in (Artetxe et al., 2022; Urbizu et al., 2022), except for the large models, where we halve the learning rate to reduce excessive variation across random seeds. For POSud, POShis-nor, and POShis, we use their respective training sets for training and evaluate on the corresponding test sets. All POS tasks use the same hyperparameters as those employed to fine-tune the NERC tasks in BasqueGLUE. For the NLI tasks, we use XNLIeu-nat and XNLIeu-var as test sets, while training on the XNLIeu training set, following the same hyperparameters as in (Bengoetxea, Gonzalez-Dios, and Agerri, 2025). We perform three runs for each task and report the average performance along with the variation across runs.

6 Results

Table 4 summarizes the performance of all BERnaT model variants ($\text{BERnaT}_{\text{standard}}$, $\text{BERnaT}_{\text{diverse}}$, and BERnaT) on the proposed set of standard and diverse NLU tasks. Models are evaluated in three sizes (medium, base, and large), enabling analysis of the effects of both training data composition and model scale.

Including diverse corpora is beneficial. The BERnaT models trained with both standard and diverse corpora achieve the best results overall on both standard and diverse datasets. This shows that including non-standard text in pretraining improves model performance not only on NLU tasks involving informal, dialectal, or historical texts, but also on tasks dealing with standard Basque texts.

In the standard subset (Table 4, upper half), $\text{BERnaT}_{\text{standard}}$ unsurprisingly achieves strong results. However, on average, these results are still below those achieved by the combined BERnaT model. While standard models perform well on tasks such as NERCid, QNLI, XNLIeu-nat, and POSud, the combined model achieves higher average performance across all large-model tasks. This indicates that exposure to diverse data does not harm performance on conventional benchmarks; on the contrary, the mixed-corpus BERnaT retains or improves accuracy on most tasks.

The diverse subset (Table 4, lower half) highlights the advantage of training with linguistically heterogeneous data. The $\text{BERnaT}_{\text{diverse}}$ models surpass their standard counterparts on several tasks, including Intent, Vaxx, and BEC, showing a clear correlation between pretraining data composition and task-specific performance, except for the NLI tasks, where $\text{BERnaT}_{\text{diverse}}$ yields results up to 13 points lower than $\text{BERnaT}_{\text{standard}}$, affecting its average scores. This stark contrast

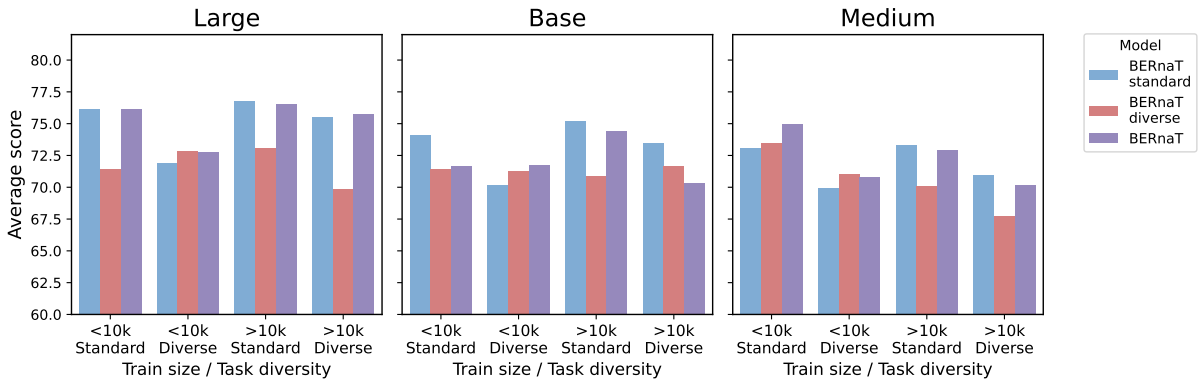

Figure 4: Average performance of Diverse, Standard and combined models, with results grouped according to training size and diversity of the task as presented in Table 2 (standard/diverse). Each plot corresponds to the size of the models (medium/base/large).

appears in both XNLIeu-nat and XNLIeu-var, which suggests that it is not a consequence of the diversity of the tasks, but rather of the task itself. Nonetheless, the combined BERnaT again delivers the best average performance on diverse tasks, slightly above both the standard and diverse variants, confirming that mixing standard and diverse sources leads to a more balanced and generalizable representation space. All in all, these results show that data diversity is particularly beneficial when balanced with standard language data: overly diverse corpora may introduce stylistic or lexical noise, but when paired with curated standard text, they help models capture a wider linguistic spectrum (variation across dialects, registers, social groups or even time) without losing precision on standard tasks.

Model size matters. The increase in performance resulting from including diverse corpora in pretraining varies according to the model size, as shown in Figure 3. Smaller models (med) do not benefit when including diverse text, and for most tasks (6 out of 9 standard tasks and 4 out of 6 diverse tasks), it is more effective to train specialized models using either standard or non-standard texts. With base models, using both standard and diverse corpora slightly underperforms compared to specialized models. As previously mentioned, large models gain the most from the inclusion of the entire corpus. Moreover, the performance boost increases with model size, not only on diverse tasks but also on standard benchmarks.

The impact of fine-tuning data size. We analyze whether the benefits of including diverse data in pretraining diminish with the size of the annotated data used to fine-tune the models for downstream tasks. In Figure 4, the evaluation tasks are grouped into two main categories, depending on the size of the training split: those with fewer than 10K instances and those with more than 10K instances. While minor differences can be found depending on the model size, a distinct trend can also be perceived. For models trained on smaller training sets, $\text{BERnaT}_{\text{standard}}$ outperforms $\text{BERnaT}_{\text{diverse}}$ in standard tasks, whereas the opposite holds for diverse tasks. This behavior aligns with our expectations: when the pretraining and the fine-tuning data match in diversity, the performance of the model improves. Nevertheless, with an increase in the size of fine-tuning data, this match is no longer necessary, as we can see that $\text{BERnaT}_{\text{standard}}$ achieves the best results in almost all cases, regardless of diversity. The gains from combining standard and diverse data during pretraining are therefore most notable when fine-tuning data are scarce, which suggests that the smaller the training dataset for fine-tuning, the more important the diversity of models' pretraining models becomes for downstream performance.

Model	Central	Western	Navarrese
BERnaT	75.13	72.08	66.67
$\text{BERnaT}_{\text{standard}}$	**75.47**	**72.50**	**69.84**
$\text{BERnaT}_{\text{diverse}}$	64.64	58.75	61.90

Table 5: Accuracy per dialect obtained by the large models of the BERnaT family.

NLI Dialect Results. Finally, we analyze NLI performance of BERnaT models for dif-

ferent Basque dialects. Table 5 shows the results of the large BERnaT models on three dialects, central, western, and navarrese. Similar to the results by (Bengoetxea, Gonzalez-Dios, and Agerri, 2025), all BERnaT models show a clear preference for the central variety, as it is closer to standard Basque, followed by western and navarrese. Surprisingly, $\text{BERnaT}_{\text{standard}}$ shows slightly higher results than BERnaT, and $\text{BERnaT}_{\text{diverse}}$ lags behind both by a large margin. We attribute these results to the effect of the size of fine-tuning data as discussed earlier, given the considerable size of the training data used for the XNLIeu-var task.

7 Conclusion

This study examines the impact of linguistic diversity on language model pretraining through the lens of Basque, a low-resource language. We introduce the BERnaT family of encoder-only models, each trained on different types of corpora: standard, diverse, and combined. To assess their performance, we introduce a novel evaluation framework comprising a comprehensive suite of NLU tasks, organized into two distinct subsets: standard and diverse evaluation benchmarks.

Our findings demonstrate that combining standard and diverse data yields the best overall performance. Models trained only on diverse data do not outperform standard models, but the integration of both leads to substantial gains in generalization across linguistic contexts. Importantly, exposure to diverse linguistic data does not degrade performance on standard benchmarks, indicating that diversity and quality are not mutually exclusive.

We also observe that larger models benefit more from data diversity, suggesting that increased capacity enhances the model's ability to internalize complex linguistic variation. Furthermore, tasks involving social media or historical texts benefit most from diverse pretraining, confirming the practical value of incorporating diverse language data in low-resource settings.

In conclusion, our results underscore the importance of linguistic diversity in building equitable and reliable language technologies. The balanced inclusion of standard and clean texts with informal, dialectal, and pre-standard language variations is key not only for fair representation but also for improved performance and generalization in multilingual and low-resource contexts.

Limitations

Despite the advances presented in our paper, our approach remains constrained by the availability of labeled data for diverse language tasks and the limitations of encoder-only architectures. We will therefore explore the use of both larger model scales and generative architectures in few-shot or zero-shot settings. This direction is promising for capturing complex linguistic variation with limited labeled training data. We will also integrate generation-based evaluations to assess linguistic adaptability beyond fine-tuning.

Acknowledgments

This work has been partially supported by the Basque Government (Research group funding IT1570-22 and IKER-GAITU project) as well as ALIA Models Development Project, Resolution of the SEDIA 19/08/2024, within the framework of the National Language Technologies Plan – ENIA 2024, funded by MTDFP, PRTR, and the EU – NextGenerationEU. The project also received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No 101135724, Topic HORIZON-CL4-2023-HUMAN-01-21, DeepKnowledge (PID2021-127777OB-C21) and HumanAIze (AIA2025-163322-C61) founded by MCIN/AEI/10.13039/501100011033 and FEDER. Jaione Bengoetxea, Julen Etxaniz and Ekhi Azurmendi hold a PhD grant from the Basque Government (PRE_2024_1_0028, PRE_2025_2_0101 and PRE_2025_2_0229, respectively). Maite Heredia and Mikel Zubillaga hold a PhD grant from the University of the Basque Country UPV/EHU (PIF23/218 and PIF24/04, respectively). The models were trained on the Leonardo supercomputer at CINECA under the EuroHPC Joint Undertaking, project EHPC-EXT-2024E01-042 and EHPC-AIF-2026LS03-009.

References

Abadji, J., P. Ortiz Suarez, L. Romary, and B. Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and

S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France, June. European Language Resources Association.

Agerri, R., R. Centeno, M. Espinosa, J. Fernandez de Landa, and A. Rodrigo. 2021. Vaxxstance@ iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del lenguaje natural*, 67:173–181.

Agerri, R., I. San Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, and E. Agirre. 2020. Give your text representation models some love: the case for Basque. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France, May. European Language Resources Association.

Alegría, I., O. Arregi, N. Ezeiza, and I. Fernandez. 2006. Lessons from the development of a named entity recognizer for basque. *Procesamiento del Lenguaje Natural*, 36.

Alhafni, B., S. Al-Towaity, Z. Fawzy, F. Nassar, F. Eryani, H. Bouamor, and N. Habash. 2024. Exploiting dialect identification in automatic dialectal text normalization. In N. Habash, H. Bouamor, R. Eskander, N. Tomeh, I. Abu Farha, A. Abdelali, S. Touileb, I. Hamed, Y. Onaizan, B. Alhafni, W. Antoun, S. Khalifa, H. Haddad, I. Zitouni, B. AlKhamissi, R. Almatham, and K. Mrini, editors, *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 42–54, Bangkok, Thailand, August. Association for Computational Linguistics.

Aranzabe, M. J., A. Atutxa, K. Bengoetxea, A. D. de Ilarraza, I. Goenaga, K. Gojenola, and L. Uria. 2015. Automatic conversion of the basque dependency treebank to universal dependencies. In *Proceedings of the fourteenth international workshop on treebanks an linguistic theories (TLT14)*, pages 233–241.

Artetxe, M., I. Aldabe, R. Agerri, O. Perez-de Viñaspre, and A. Soroa. 2022. Does corpus quality really matter for low-resource languages? In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Barbieri, F., J. Camacho-Collados, L. Espinosa Anke, and L. Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November. Association for Computational Linguistics.

Barbieri, F., L. Espinosa Anke, and J. Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France, June. European Language Resources Association.

Bengoetxea, J., I. Gonzalez-Dios, and R. Agerri. 2025. Lost in variation? evaluating NLI performance in Basque and Spanish geographical variants. In G. Boleda and M. Roth, editors, *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 452–468, Vienna, Austria, July. Association for Computational Linguistics.

Blodgett, S. L., S. Barocas, H. Daumé III, and H. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July. Association for Computational Linguistics.

Blodgett, S. L., L. Green, and B. O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in*

Natural Language Processing, pages 1119–1130, Austin, Texas, November. Association for Computational Linguistics.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Conneau, A., R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November. Association for Computational Linguistics.

Dao, T. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning.

de Gibert, O., G. Nail, N. Arefyev, M. Bañón, J. van der Linde, S. Ji, J. Zaragoza-Bernabeu, M. Aulamo, G. Ramírez-Sánchez, A. Kutuzov, S. Pyysalo, S. Oepen, and J. Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia, May. ELRA and ICCL.

de Marneffe, M.-C., C. D. Manning, J. Nivre, and D. Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, June.

Eisenstein, J. 2013. What to do about bad language on the internet. In L. Vanderwende, H. Daumé III, and K. Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia, June. Association for Computational Linguistics.

Elordui, A. and J. Aiestaran. 2022. Basque in instagram: A scalar approach to vernacularisation and normativity. *Journal of Sociolinguistics*.

Estarrona, A., I. Etxeberria, A. Soraluze, R. Etxepare, and M. Padilla-Moyano. 2021. The first annotated corpus of historical basque. *Digital Scholarship in the Humanities*, 37(2):391–404, 10.

Etxaniz, J., O. Sainz, N. Miguel, I. Aldabe, G. Rigau, E. Agirre, A. Ormazabal, M. Artetxe, and A. Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand, August. Association for Computational Linguistics.

Euskara Institutua, E. 2013. Euskal klasikoen corpusa (ekc).

Fernandez de Landa, J. and R. Agerri. 2021. Social analysis of young basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, 0(0):1–15.

Fernandez de Landa, J., R. Agerri, and I. Alegria. 2019. Large scale linguistic processing of tweets to understand social interactions among speakers of less resourced languages: The basque case. *Information*, 10(6).

Fernandez de Landa, J., I. García-Ferrero, A. Salaberria, and J. A. Campos. 2024. Uncovering social changes of the Basque speaking Twitter community during COVID-19 pandemic. In M. Melero, S. Sakti, and C. Soria, editors, *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 363–371, Torino, Italia, May. ELRA and ICCL.

Gisserot-Boukhlef, H., N. Boizard, M. Faysse, D. M. Alves, E. Malherbe, A. F. T. Martins, C. Hudelot, and P. Colombo. 2025. Should we still pretrain encoders with masked language modeling?

Gururangan, S., D. Card, S. Dreier, E. Gade, L. Wang, Z. Wang, L. Zettlemoyer, and N. A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Han, B. and T. Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In D. Lin, Y. Matsumoto, and R. Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA, June. Association for Computational Linguistics.

Heredia, M., J. Etxaniz, M. Zulaika, X. Saralegi, J. Barnes, and A. Soroa. 2024. XNLIeu: a dataset for cross-lingual NLI in Basque. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4177–4188, Mexico City, Mexico, June. Association for Computational Linguistics.

Hualde, J. and K. Zuazo. 2007. The standardization of the basque language. *Language Problems & Language Planning*, 31:142–168, 07.

Jørgensen, A., D. Hovy, and A. Søgaard. 2015. Challenges of studying and processing dialects in social media. In W. Xu, B. Han, and A. Ritter, editors, *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China, July. Association for Computational Linguistics.

Kuparinen, O., A. Miletić, and Y. Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore, December. Association for Computational Linguistics.

Laurençon, H., L. Saulnier, T. Wang, C. Akiki, A. V. del Moral, T. L. Scao, L. V. Werra, C. Mou, E. G. Ponferrada, H. Nguyen, J. Frohberg, M. Šaško, Q. Lhoest, A. McMillan-Major, G. Dupont, S. Biderman, A. Rogers, L. B. allal, F. D. Toni, G. Pistilli, O. Nguyen, S. Nikpoor, M. Masoud, P. Colombo, J. de la Rosa, P. Villegas, T. Thrush, S. Longpre, S. Nagel, L. Weber, M. R. Muñoz, J. Zhu, D. V. Strien, Z. Alyafeai, K. Almubarak, V. M. Chien, I. Gonzalez-Dios, A. Soroa, K. Lo, M. Dey, P. O. Suarez, A. Gokaslan, S. Bose, D. I. Adelani, L. Phan, H. Tran, I. Yu, S. Pai, J. Chim, V. Lepercq, S. Ilic, M. Mitchell, S. Luccioni, and Y. Jernite. 2022. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

López de Lacalle, M., X. Saralegi, and I. San Vicente. 2020. Building a task-oriented dialog system for languages with no training data: the case for Basque. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2796–2802, Marseille, France, May. European Language Resources Association.

Markl, N. and S. J. McNulty. 2022. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6328–6339, Marseille, France, June. European Language Resources Association.

Marone, M., O. Weller, W. Fleshman, E. Yang, D. Lawrie, and B. V. Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning.

Nguyen, D. Q., T. Vu, and A. Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online, October. Association for Computational Linguistics.

Nguyen, T., C. Van Nguyen, V. D. Lai, H. Man, N. T. Ngo, F. Dernoncourt, R. A. Rossi, and T. H. Nguyen. 2024. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237.

Oñederra, M. L. 2016. Standardisation of basque: From grammar (1968) to pronunciation (1998). *Sociolinguistica*, 30(1):125–144.

Palomar-Giner, J., J. J. Saiz, F. Espuña, M. Mina, S. Da Dalt, J. Llop, M. Ostendorff, P. Ortiz Suarez, G. Rehm, A. Gonzalez-Agirre, and M. Villegas. 2024. A CURATEd CATalog: Rethinking the extraction of pretraining corpora for mid-resourced languages. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 335–349, Torino, Italia, May. ELRA and ICCL.

Rust, P., J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online, August. Association for Computational Linguistics.

San Vicente, I., G. Urbizu, A. Corral, Z. Beloki, and X. Saralegi. 2024. Zelaihandi: A large collection of basque texts.

Soldaini, L., R. Kinney, A. Bhagia, D. Schwenk, D. Atkinson, R. Authur, B. Bogin, K. Chandu, J. Dumas, Y. Elazar, V. Hofmann, A. Jha, S. Kumar, L. Lucy, X. Lyu, N. Lambert, I. Magnusson, J. Morrison, N. Muennighoff, A. Naik, C. Nam, M. Peters, A. Ravichander, K. Richardson, Z. Shen, E. Strubell, N. Subramani, O. Tafjord, E. Walsh, L. Zettlemoyer, N. Smith, H. Hajishirzi, I. Beltagy, D. Groeneveld, J. Dodge, and K. Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand, August. Association for Computational Linguistics.

Soraluze, A., O. Arregi, X. Arregi, and K. Ceberio. 2012. Mention detection: First steps in the development of a basque coreference resolution system. In *KONVENS*, pages 128–136.

Soraluze Irureta, A., M. Padilla Moyano, A. Estarrona Ibarloza, and I. Etxeberria Uztarroz. 2019. Spelling normalisation of basque historical texts. *Procesamiento del lenguaje natural*, 63:59–66.

Srirag, D., N. R. Sahoo, and A. Joshi. 2025. Evaluating dialect robustness of language models via conversation understanding. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 24–38, Abu Dhabi, January. Association for Computational Linguistics.

Urbizu, G., I. San Vicente, X. Saralegi, R. Agerri, and A. Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France, June. European Language Resources Association.

Urbizu, G., M. Zulaika, X. Saralegi, and A. Corral. 2024. How well can BERT learn the grammar of an agglutinative

and flexible-order language? the case of Basque. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8334–8348, Torino, Italia, May. ELRA and ICCL.

Walsh, O. 2021. Introduction: in the shadow of the standard. standard language ideology and attitudes towards 'non-standard' varieties and usages. *Journal of Multilingual and Multicultural Development*, 42(9):773–782.

Warner, B., A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Weller, O., K. Ricci, M. Marone, A. Chaffin, D. Lawrie, and B. V. Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders.

Wenzek, G., M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May. European Language Resources Association.

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

A Evaluation datasets language diversity analysis

As with the pretraining data, we conducted a language diversity analysis on the evaluation datasets. For each task, we computed diversity by combining the train, validation, and test splits. The exceptions to this were XNLIeu-var and POShis-nor, where we used only the test splits, as their training data is directly extracted from the XNLIeu-nat and POShis tasks, respectively. Table 6 presents the diversity scores for each task in descending order. We observed a correlation between domain and diversity, the most diverse datasets typically originate from broader domains (such as social media or historical texts). An exception is POSud, which yielded diversity scores similar to those of the Slot or Intent tasks.

Task	**Diversity**
POShis	$6.58e^{-1}$
XNLIeu-var*	$3.97e^{-1}$
POShis-nor*	$1.68e^{-1}$
BEC	$9.13e^{-2}$
Vaxx	$2.94e^{-2}$
Slot	$1.65e^{-2}$
Intent	$1.58e^{-2}$
POSud	$1.56e^{-2}$
WiC	$9.50e^{-3}$
NERCid	$8.82e^{-3}$
Korref	$6.69e^{-3}$
XNLIeu-nat*	$5.60e^{-3}$
NERCod	$5.52e^{-3}$
BHTC	$3.65e^{-3}$
QNLI	$3.47e^{-3}$

Table 6: Diversity scores per evaluation dataset. Scores are computed across all combined splits, except for datasets marked with (*), which use only the test split.