BERnaT: Basque Encoders for Representing Natural Textual Diversity

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}@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, reduce model robustness 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 Basque, a morphologically rich and low-resource language, we construct new corpora combining standard, social media, and historical sources, and pre-train 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 consistently outperform those trained on standard corpora, improving performance across all task types without compromising standard benchmark accuracy. These findings highlight the importance of linguistic diversity in building inclusive, generalizable language models.

Keywords: Training, Fine-tuning, Adaptation, Alignment, and Representation Learning; Less-Resourced/Endangered Languages; Evaluation Methodologies; Corpus (Creation, Annotation, etc.); Social Media Processing

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 robust and 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 et al., 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. We select Basque for our study, a low-resource and morphologically rich language, and conduct our experiments using encoder-based architectures. 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

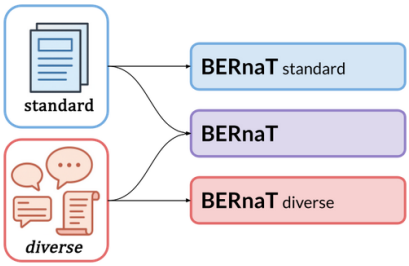

Figure 1: Summary of corpora combinations used to train BERnaT models. *Standard Corpus*: high-quality standard Basque from sources like News or Wikipedia. *Diverse Corpora*: social media and historical texts capturing informal, dialectal, and pre-standard Basque.

much larger (Weller et al., 2025; Gisserot-Boukhlef et al., 2025). Consequently, our research's aim is 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 pre-train a collection of models on three different corpora combinations: a widely used standard corpus is used to train the BERnaT$_{\text{standard}}$ models; the BERnaT$_{\text{diverse}}$ models are trained using a newly introduced diverse corpus that covers different varieties of Basque; and, finally, the BERnaT models are trained with both the standard and the diverse corpora.

To evaluate the models, we propose a new configuration of NLU tasks, 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. In addition, we include other pre-existing tasks and even create new ones. 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.

We obtain these main findings from our experiments, which could be helpful when creating data and models for other languages:

- The inclusion of non-standard text in model pre-training is helpful for tasks that deal with non-standard language, such as text from social media, dialectal variations and historical texts. Moreover, the benefit of including diversity in the pre-training is highest when the fine-tuning training dataset is small.
- Including diverse text doesn't degrade the results of traditional Natural Language Understanding (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 results when combining 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.

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

2. Related work

Approaches to language diversity. The exclusion of varieties outside of the standard can be motivated by underlying language ideologies about what constitutes clean language (Walsh, 2021) or the convenience of a more standardized and homogeneous language in language modeling and evaluation (Jørgensen et al., 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 et al., 2025; Srirag et al., 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 et al., 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. Encoder models have been generally trained on standard corpora, such as Wikipedia or journalistic texts, both in English (Liu et al., 2019) and multilingual settings (Conneau et al., 2020). Several efforts have also focused on training encoder models on diverse corpora from social media in English (Nguyen et al., 2020) and other languages (Barbieri et al., 2022), and on evaluating them on that specific source (Barbieri et al., 2020). Recently, works have centered on training encoder models with modern

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	13k	188M	$1.4e^{-1} \pm 1.7e^{-1}$
EKC	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.

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 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 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

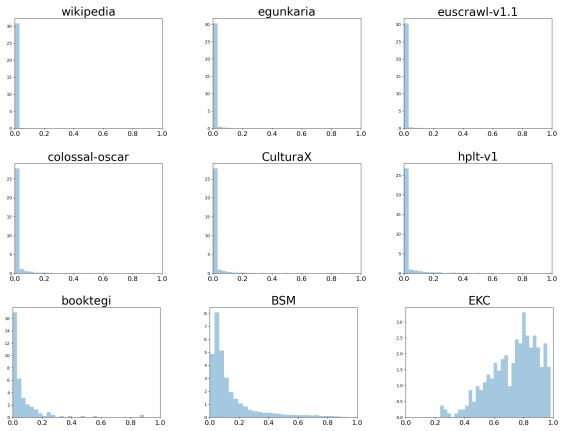

Figure 2: Visualization of diversity distribution for latxa standard corpora (*wikipedia, egunkaria, euscrawl-v1.1, colossal-oscar, CulturaX, hplt-v1, booktegi*) and newly added *BSM* and *EKC* non-standard corpora.

language use. In the following, we describe the corpora used to train the models, as well as the language diversity analysis performed.

3.1. Standard Language Corpus

We utilize the largest standard Basque corpus currently available, latxa-corpus-v1.1 (Etxaniz et al., 2024). To ensure linguistic consistency and quality, the corpus was systematically normalized, deduplicated, and exhaustively filtered. 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 HLPT v1 (de Gibert 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).

Dataset	Source	Task	Train/Dev	Test	Domain
BTHC	(Urbizu et al., 2022)	Topic classification	10,442	1,854	News
Korref	(Soraluze et al., 2012)	Correference resolution	1,306	587	News
NERCid	(Alegría et al., 2006)	NERC	64,475	35,855	News
NERCod	(Urbizu et al., 2022)	NERC	79,420	14,462	News, Wikipedia
QNLI	(Urbizu et al., 2022)	QA/NLI	1,994	238	Wikipedia
WiC	(Urbizu et al., 2022)	Word sense disambiguation	409,159	1,400	News
XNLIeu-nat**	(Heredia et al., 2024)	NLI	392,000	621	News
POSud**	(de Marneffe et al., 2021)	POS tagging	7,194	1,799	News, literary texts
POShis-nor**	**[new]**	POS tagging	5,095	526	Normalized historical texts
BEC	(Urbizu et al., 2022)	Sentiment Analysis	7,380	1,302	Twitter
Intent	(López de Lacalle et al., 2020)	Intent classification	5,322	1,087	Facebook
Slot	(López de Lacalle et al., 2020)	Slot filling	30,443	5,633	Facebook
Vaxx	(Agerri et al., 2021)	Stance detection	1,070	312	Twitter
XNLIeu-var**	(Bengoetxea et al., 2025)	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.

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 et al., 2019). Building on the same techniques and extending previously collected resources (Fernandez de Landa et al., 2024), it 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 two complementary subsets containing the same textual content but organized differently: 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 perform a language diversity analysis on the entire pre-training data to ensure that the sources are consistent 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 re-

sources for modeling the full spectrum of Basque linguistic variation.

4. Evaluation datasets

We use Natural Language Understanding (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.

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 et al., 2019). To analyze performance under different linguistic variations, we split the benchmark into two evaluation blocks: standard and diverse. The standard block includes topic classification (BHTC), coreference resolution (Korref), question answering framed as NLI (QNLI), word sense disambiguation in context (WiC), and named entity recognition/classification in both in-domain (NERCid) and out-of-domain (NERCod) settings. The diverse block, 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. For standard language evaluation, we use the so-called XNLIeu-nat part of the dataset, a test set created entirely from scratch by native Basque speakers. We also include the NLI dataset presented in (Bengoetxea et al., 2025), where XNLIeu-nat was adapted into three Basque dialects (Western, Central, and Navarrese), therefore allowing us to evaluate NLI in both standard and diverse language settings. Both datasets, however, are relatively small and thus used exclusively for evaluation purposes. For training, we rely on the automatically translated train set from XNLIeu, 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. For standard language, we employ 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) primarily drawn from literary and journalistic sources. To capture language diversity, we additionally introduce POShis, a newly developed dataset specifically targeting historical language. This dataset is 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). Therefore, we present a selection from this morphosyntactically annotated diachronic corpus, consisting of 5,621 instances (693,414 tokens), making it approximately five times larger than POSud. The manual annotations in the dataset have been converted to follow the same criteria as those used in POSud. The dataset includes 5,095 instances for the training and development sets and 526 instances for evaluation. Additionally, a parallel normalized version of this dataset, POShis-nor, is also provided 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 pre-training on NLU tasks. To that end, we start by pre-training the models with the corpora described in Section 3, which are then finetuned 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) 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), as depicted in Figure 1, therefore generating $\text{BERnaT}_{\text{standard}}$, $\text{BERnaT}_{\text{diverse}}$ and BERnaT models respectively. We follow Artetxe et al. (2022) and pre-train each model with 1 million tokens with a sequence length of 512 tokens. Shorter documents are concate-

		BERnaT$_{\text{standard}}$			BERnaT$_{\text{diverse}}$			BERnaT		
		med	base	large	med	base	large	med	base	large
Standard tasks	BHTC	74.85	75.37	77.53	73.37	73.97	76.70	75.06	75.56	**77.83**
	Korref	64.45	58.66	58.04	57.13	59.00	61.61	60.70	63.54	**69.73**
	NERCid	83.35	83.83	**86.48**	80.19	75.50	83.34	81.33	83.39	84.97
	NERCod	73.29	75.14	75.04	65.99	67.77	70.63	72.60	75.67	**76.06**
	QNLI	69.61	73.25	**74.23**	70.73	70.73	71.01	69.19	71.15	72.96
	WiC	67.69	70.57	70.07	68.86	70.33	70.14	69.07	69.55	**70.86**
	XNLIeu-nat	67.31	71.28	**74.93**	61.89	66.72	64.52	66.51	67.95	72.79
	POSud	94.85	95.61	**96.27**	94.06	95.42	95.38	94.46	95.86	95.97
	POShis-nor	71.48	74.29	78.92	72.73	72.55	76.96	73.10	76.12	**79.74**
	AVG standard tasks	74.10	75.33	76.83	71.66	72.44	74.48	73.56	75.42	**77.88**
Diverse tasks	BEC	69.87	66.05	68.10	69.43	69.82	**70.10**	68.87	70.02	69.40
	Intent	75.31	77.46	77.80	77.34	75.50	**79.70**	75.22	76.84	78.04
	Slot	76.73	**78.95**	77.01	76.93	78.26	76.01	78.02	75.75	78.61
	Vaxx	61.60	63.06	65.97	64.03	66.67	**68.46**	66.13	65.77	67.44
	XNLIeu-var	65.21	68.05	**74.05**	58.50	65.14	63.68	62.30	64.91	72.82
	POShis	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.

nated 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 pre-training, 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 adopt the hyperparameter configuration proposed by Artetxe et al. (2022). However, when applying the same configuration to the more heterogeneous corpora, we observe frequent occurrences of exploding gradients. This instability is likely attributable to the increased diversity of the training data and the corresponding sensitivity of the model when optimizing over such a distribution. To mitigate this issue, we systematically reduce the learning rate until stable convergence is achieved. Preliminary experiments reveal a clear correlation between corpus diversity and the magnitude of the learning rate required to maintain training stability. Specifically, as data diversity increases, proportionally smaller learning rates are necessary to prevent gradient explosion. Consequently, we identify optimal learning rates of $8e^{-4}$, $4e^{-4}$ and $1e^{-4}$ for the medium, base, and large models, respectively, to ensure stable and complete training. We maintain these learning rates across all BERnaT variants to enable a fair comparison of results. For the remaining hyperparameters, we follow the configuration proposed by Artetxe et al. (2022).

Tokenizers. In order to accurately represent the data on our corpora, given that the target language is morphologically rich and exhibits substantial lexical variation across domains, 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$_{\text{Both}}$) on all corpus configurations. Token fertility quantifies the average number of subword tokens required to represent a single 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. As expected, tokenizers trained on specific corpora obtain the lowest token fertility on that corpora, 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 et al., 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 (BERnaT$_{\text{standard}}$, 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), BERnaT$_{\text{standard}}$ unsurprisingly achieves strong results, with the large version still below 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 BERnaT$_{\text{diverse}}$ models surpass their standard counterparts on several tasks, including Intent, Vaxx, and BEC, showing a clear correlation between pre-training data composition and task-specific performance, except for the NLI tasks, where BERnaT$_{\text{diverse}}$ yields results up to 13 points lower than BERnaT$_{\text{standard}}$, affecting its average scores. This stark contrast 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. Hence, we confirm that data diversity, contrary to being harmful, can enhance performance if enough standard language corpora are included.

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. Small models (med) do not benefit from 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 medium models (base), 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.

7. Analysis and Discussion

The impact of fine-tuning data size. We want to analyze whether the benefits of including diverse data in pre-training diminish with the size of the annotated data used to finetune 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, BERnaT$_{\text{standard}}$ outperforms BERnaT$_{\text{diverse}}$ in standard tasks, whereas the opposite holds for diverse tasks. This behavior aligns with our expectations: when the pre-training 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 BERnaT$_{\text{standard}}$ achieves the best results in almost all cases, regardless of diversity. The gains from combining standard and diverse data during pre-training 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' pre-training models becomes for downstream performance. In any case, the full BERnaT is a good compromise that obtains the best results on average.

NLI Dialect Results. We analyze NLI results with more fine-grained dialect classes. Table 5 shows fine-grained results for the large models of the BERnaT family. Similar to the results by Bengoetxea et al. (2025), all BERnaT models have a preference for the central variety, as it is closer to the standard, followed by western and navarrese.

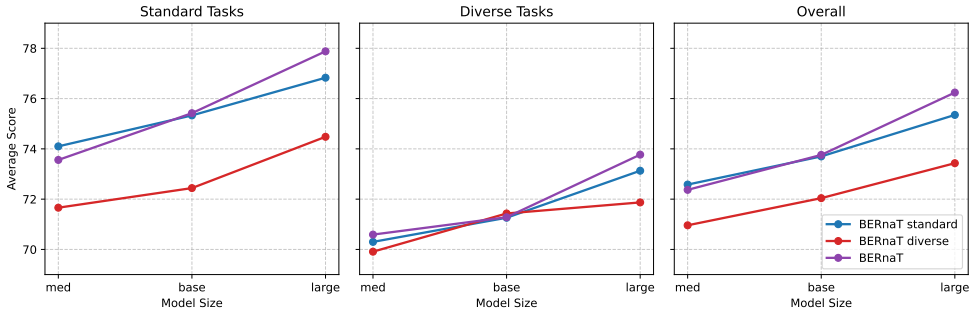

Figure 3: Average performance of Diverse, Standard and combined models in three sizes on Standard, Diverse and all tasks.

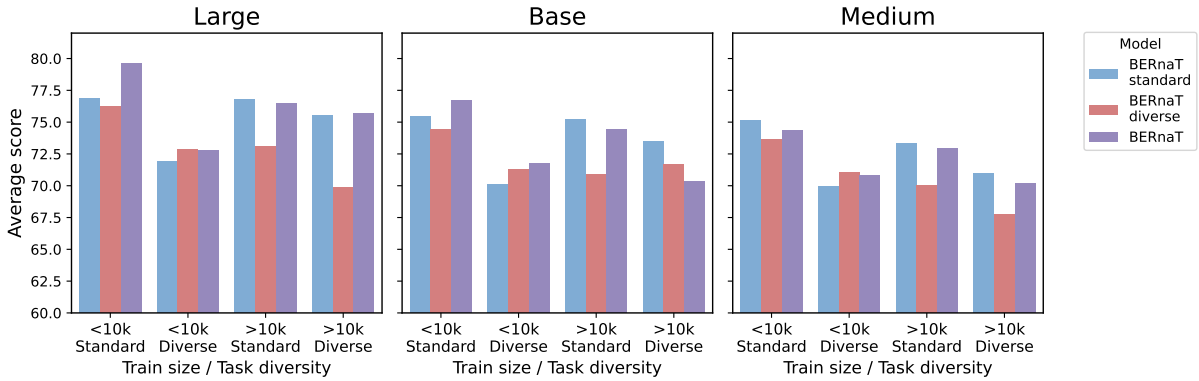

Figure 4: Average performance of Diverse, Standard and combined models, with results grouped according to training size and diversity of the task as presented in Table 2 (standard/diverse). Each plot corresponds to the size of the models (medium/base/large).

	central	western	navarrese
BERnaT	75.13	72.08	66.67
BERnaT$_{\text{standard}}$	**75.47**	**72.50**	**69.84**
BERnaT$_{\text{diverse}}$	64.64	58.75	61.90

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

Surprisingly, BERnaT$_{\text{standard}}$ shows slightly higher results than BERnaT, and 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.

8. Conclusion

This study examines the impact of linguistic diversity on language model pre-training 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 Natural Language Understanding (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 pre-training, 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 robust 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 multilin-

gual 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), the Spanish Ministry for Digital Transformation and Civil Service, and the EU-funded NextGenerationEU Recovery, Transformation and Resilience Plan (ILENIA project, 2022/TL22/00215335; and ALIA project). 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 and DeepKnowledge (PID2021-127777OB-C21) 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_2024_2_0028 and PRE_2024_1_0035, 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.

Bibliographical References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Rodrigo Agerri, Roberto Centeno, María Espinosa, Joseba Fernandez de Landa, and Alvaro 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.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

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

Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhl Eryani, Houda Bouamor, and Nizar Habash. 2024. Exploiting dialect identification in automatic dialectal text normalization. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 42–54, Bangkok, Thailand. Association for Computational Linguistics.

Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, and Larraitz 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.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Jaione Bengoetxea, Itziar Gonzalez-Dios, and Rodrigo Agerri. 2025. Lost in variation? evaluating NLI performance in Basque and Spanish geographical variants. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 452–468, Vienna, Austria. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. 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.

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

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

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

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

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

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

Ainara Estarrona, Izaskun Etxeberria, Ander Soraluze, Ricardo Etxepare, and Manuel Padilla-Moyano. 2021. The first annotated corpus of historical basque. *Digital Scholarship in the Humanities*, 37(2):391–404.

Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.

EHU Euskara Institutua. 2013. Euskal klasikoen corpusa (ekc).

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

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

Joseba Fernandez de Landa, Iker García-Ferrero, Ander Salaberria, and Jon Ander Campos. 2024. Uncovering social changes of the Basque speaking Twitter community during COVID-19 pandemic. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages @ LREC-COLING 2024*, pages 363–371, Torino, Italia. ELRA and ICCL.

Hippolyte Gisserot-Boukhlef, Nicolas Boizard, Manuel Faysse, Duarte M. Alves, Emmanuel Malherbe, André F. T. Martins, Céline Hudelot, and Pierre Colombo. 2025. Should we still pretrain encoders with masked language modeling?

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

Maite Heredia, Julen Etxaniz, Muitze Zulaika, Xabier Saralegi, Jeremy Barnes, and Aitor Soroa. 2024. XNLIeu: a dataset for cross-lingual NLI in Basque. In *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. Association for Computational Linguistics.

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

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China. Association for Computational Linguistics.

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine 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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Maddalen López de Lacalle, Xabier Saralegi, and Iñaki San Vicente. 2020. Building a task-oriented dialog system for languages with no training data: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2796–2802, Marseille, France. European Language Resources Association.

Nina Markl and Stephen Joseph McNulty. 2022. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6328–6339, Marseille, France. European Language Resources Association.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu 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.

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

Jorge Palomar-Giner, Jose Javier Saiz, Ferran Espuña, Mario Mina, Severino Da Dalt, Joan Llop, Malte Ostendorff, Pedro Ortiz Suarez, Georg Rehm, Aitor Gonzalez-Agirre, and Marta Villegas. 2024. A CURATEd CATalog: Rethinking the extraction of pretraining corpora for mid-resourced languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 335–349, Torino, Italia. ELRA and ICCL.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Iñaki San Vicente, Gorka Urbizu, Ander Corral, Zuhaitz Beloki, and Xabier Saralegi. 2024. Zelaihandi: A large collection of basque texts.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Ander Soraluze, Olatz Arregi, Xabier Arregi, and Klara Ceberio. 2012. Mention detection: First steps in the development of a basque coreference resolution system. In *KONVENS*, pages 128–136.

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

Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya 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. Association for Computational Linguistics.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Gorka Urbizu, Muitze Zulaika, Xabier Saralegi, and Ander Corral. 2024. How well can BERT learn the grammar of an agglutinative and flexible-order language? the case of Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8334–8348, Torino, Italia. ELRA and ICCL.

Olivia Walsh. 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.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.