Lost in Variation?
Evaluating NLI Performance in Basque and Spanish Geographical Variants

Jaione Bengoetxea and Itziar Gonzalez-Dios and Rodrigo Agerri
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{jaione.bengoetxea,itziar.gonzalezd,rodrigo.agerri}@ehu.eus

Abstract

In this paper, we evaluate the capacity of current language technologies to understand Basque and Spanish language varieties. We use Natural Language Inference (NLI) as a pivot task and introduce a novel, manually-curated parallel dataset in Basque and Spanish, along with their respective variants. Our empirical analysis of crosslingual and in-context learning experiments using encoder-only and decoder-based Large Language Models (LLMs) shows a performance drop when handling linguistic variation, especially in Basque. Error analysis suggests that this decline is not due to lexical overlap, but rather to the linguistic variation itself. Further ablation experiments indicate that encoder-only models particularly struggle with Western Basque, which aligns with linguistic theory that identifies peripheral dialects (e.g., Western) as more distant from the standard. All data and code are publicly available.[1]

1 Introduction

Sociolinguistics examines language variation in relation to various regional, contextual, or social factors. During the 70s and 80s, sociolinguist William Labov highlighted the social aspect of language, and his work on rule-governed language variation has thereby legitimized non-standard language and transformed the study of sociolinguistics. For example, Labov (2006) noted that “the linguistic behavior of individuals cannot be understood without knowledge of the communities that they belong to”. Thus, *variation* is an intrinsic characteristic of language, influenced by factors such as gender, age, socio-economics, or geographical location. In fact, humans make no distinction when processing their own dialect or the standard variant.

In this regard, Coseriu (1956) offered a systematic typology of language variation, based on the following three types: (i) diatopic variation, or geographical variation such as dialects, (ii) diastratic variation, or speech of different societal groups, and (iii) diaphasic variation, or speech changes depending on the communicative environment.

Basque	
Standard	Variation
Zeharo hunkituta gelditu nintzen ezusteko agur honekin	Asko emoziona nintzen ezusteko agur horregaz
Spanish	
Me quedé completamente conmovido con esta despedida inesperada	Me quedé completamente conmovío con ehta dehpedía inehperá
English	
I was completely surprised by that unexpected goodbye	

Table 1: Example from standard to variation sentences in Basque and Spanish.

In this paper, we focus on geographical variation in Basque, a low-resource language isolate with around 1 million speakers that is still undergoing a normalization process (started in 1968), and in Spanish, a higher-resourced language whose standardization process started in the 18th century with around 600 million speakers worldwide.

Recent developments in Artificial Intelligence (AI) and Natural Language Processing (NLP) have underscored the significance of social factors in language for NLP systems, as noted by Hovy and Yang (2021). This indicates the importance of developing NLP technology that not only processes standard language but also variations, as this would alleviate any potential language-based discrimination by providing more linguistically-inclusive resources.

Although previous work on NLP has primarily focused on standard language, recent research has slightly shifted its attention to the exploration of

[1] https://huggingface.co/datasets/HiTZ/XNLIvar

language variation. For instance, Zampieri et al. (2020) or Joshi et al. (2024) present thorough outlines of variation-inclusive research. However, due to the lack of data on linguistic variation, most NLP research has focused on a narrow list of languages and their variants, such as Arabic, Indic languages, or German (Joshi et al., 2024). Furthermore, other larger efforts are either based on automatically obtained data or do not provide fine-grained variation distinctions for some widely-spread languages, such as Spanish (Faisal et al., 2024; Alam et al., 2024).

Regarding Basque, the few available works have focused on historical dialects (Estarrona et al., 2020) or northern Basque dialects (Uria and Etxepare, 2012). In Spanish, all datasets with linguistic variation have been automatically collected through geolocation techniques (España-Bonet and Barrón-Cedeño, 2024; Valentini et al., 2024).

In this context, the objective of this paper is to provide the first manually curated variation dataset for Basque and Spanish that captures language variation in real-world usage. To do so, we introduce XNLIvar, the first variation-inclusive Natural Language Inference (NLI) dataset in Basque and Spanish. An example of an instance from our dataset can be found in Table 1, which will be used to evaluate current state-of-the-art language models. The main contributions are the following:

1. The first publicly available manually-curated NLI dataset for Basque and Spanish geographic language variations.
2. A comprehensive evaluation of encoder-only and decoder-based Large Language Models (LLMs) demonstrates substantially worse performance when processing language variation, particularly in Basque. Detailed error analysis shows that lexical overlap between premise and hypothesis has no impact on the performance drop, which indicates that linguistic variation could be the primary factor for this decrease in accuracy.
3. Empirical results suggest that LLM performance with Spanish variants may be attributed to the substantial representation of Spanish-language content in pre-training corpora. Further error analysis suggests that orthographic changes have a substantially negative effect on Spanish language variation processing.

To the best of our knowledge, no work has extensively addressed the automatic processing of language variation of Western and Central Basque dialects in the task of NLI.

2 Related Work

This section presents previous work on language variation in the field of NLP, with a specific focus on Basque language variation.

Language variation in NLP In recent years, there has been an increasing interest in dialects in several fields of NLP, such as dialect identification (Ramponi and Casula, 2023), sentiment analysis (Ball-Burack et al., 2021), Machine Translation (MT) (Kuparinen et al., 2023), and dialogue systems (Alshareef and Siddiqui, 2020).

Aepli and Sennrich (2022) explored cross-lingual transfer between closely related varieties by adding character-level noise to high-resource data to improve generalization. Moreover, Ramponi and Casula (2023) pretrained LLMs for geographic variations of Italian tweets. Finally, Demszky et al. (2021) showed that BERT models trained on annotated corpora obtained high accuracy for Indian English feature detection.

One of the primary limitations of these studies is the scarcity of available dialectal data. Therefore, research has largely focused on developing resources such as lexicons and dialectal datasets on a small subset of languages: Artemova and Plank (2023) propose a bilingual lexicon induction method for German dialects using LLMs, while Hassan et al. (2017) introduce a synthetic data creation method through embeddings by transforming input data into its dialectic variant. With respect to language coverage, the Arabic family, due to its relative data availability, has received the most attention, followed by Indic languages, Chinese, and German (Joshi et al., 2024).

Basque language variation In dialectology, Zuazu (2008) established an extensive and comprehensive descriptive representation of features of modern Basque dialects. In NLP, Estarrona et al. (2020) worked on a morpho-syntactically annotated corpus of Basque historical texts as an aid in the normalization process. Moreover, Uria and Etxepare (2012) introduced a corpus of syntactic variation in northern Basque dialects.

Additionally, some dialectal benchmark works have included Basque in their experimentation: both Alam et al. (2024) and Faisal et al. (2024) presented benchmarks for MT with northern Basque dialects.

Spanish language variation Several works have dealt with Spanish varieties. For instance, España-Bonet and Barrón-Cedeño (2024) automatically filtered Open Super-large Crawled Aggregated coRpus (OSCAR) by geolocation into different Spanish variants and performed a stylistic analysis. Valentini et al. (2024) automatically collected Google queries from several Spanish-speaking countries and provided an Information Retrieval baseline for Spanish varieties.

Additionally, Lopetegui et al. (2025) introduced a Cuban Spanish dataset by collecting geolocated tweets from Twitter. They focused their study on common examples, i.e., instances that can be valid across several dialects. They performed a manual annotation of tweets into *Cuban dialect*, *other dialect*, or *common example*. Similarly, Castillo-lópez et al. (2023) collected tweets from European and Latin American geolocations and annotated them for hate speech.

3 Data

In this work, we introduce a novel dataset, **XNLI-var**, that expands the XNLI framework by generating various dialectal variations for both Basque and Spanish languages. We choose NLI as an evaluation framework because it is considered to be a general benchmark for evaluating language understanding, which requires dealing with semantic relationships, logical implications, world knowledge, and contextual nuances (Williams et al., 2018; Conneau et al., 2018; Artetxe et al., 2020), including figurative language (Naik et al., 2018; Stowe et al., 2022; Liu et al., 2022; Sanchez-Bayona and Agerri, 2024). NLI is a fundamental NLP task that involves classifying the logical relationship between two segments (a premise and a hypothesis) as one of three categories: entailment (the hypothesis logically follows from the premise), contradiction (the hypothesis contradicts the premise), or neutral (the hypothesis neither follows from nor contradicts the premise). The most popular dataset is the English **MultiNLI** (Williams et al., 2018).

XNLI (Conneau et al., 2018) provides an extension of MultiNLI in 15 languages, among them Spanish (**XNLIes**). The training set for each language was created by translating the original MNLI data. However, as the test partition of MNLI is not public, Conneau et al. (2018) collected 7,500 English examples via crowdsourcing, which were then professionally translated to create the development (2500 instances) and the test (5K instances) splits of XNLI. This parallel multilingual corpus has facilitated crosslingual NLI research beyond English-centric approaches by exploring model-transfer, translate-train, and translate-test techniques to alleviate the lack of annotated training data in a given target language (Artetxe et al., 2020, 2022).

XNLIeu is a professionally translated version of the English XNLI set into Basque (Heredia et al., 2024), a language not included in the original XNLI dataset. Additionally, we also use **XNLIeu$_{\text{native}}$**, an NLI dataset generated by collecting native Basque premises and hiring Basque annotators to create three hypotheses per premise (Heredia et al., 2024). The experimental results from XNLIeu demonstrate that NLI systems exhibit significant performance sensitivity to disparities between training and testing data distributions, highlighting the critical role of data provenance (Artetxe et al., 2020; Volansky et al., 2013).

3.1 XNLI with Geographic Variants

To investigate the impact of language variation via evaluation in NLI, we developed two novel **XNLI variants** datasets encompassing Basque and Spanish geographic-based linguistic variations, namely, **XNLIeu$_{\text{var}}$** and **XNLIes$_{\text{var}}$**. The methodology involved a language adaptation phase to ensure the incorporation of variant diversity within the data. These two variant datasets were developed taking **XNLIeu$_{\text{native}}$** as a starting point for dialectal augmentation due to its authentic representation of Basque language patterns and its suitable scale for manual paraphrasing.

The adaptation process was the same for Basque and Spanish, including native speakers as linguistic informants for variant transformation. We wanted to analyze the variation that naturally occurs among native speakers, employing minimally restrictive parameters to capture authentic dialectal features. Thus, informants were instructed to perform dialectal adaptations of source sentences, with allowance for modifications across multiple linguistic dimensions, including lexical, grammatical, phonetic, and orthographic alterations. The full adaptation guidelines are detailed in Appendix A.

XNLIeu$_{\text{var}}$ Twelve native Basque speakers were recruited from diverse geographical regions. All participants possessed expertise in NLP and held university degrees in either Linguistics, Computer

Science, or Engineering. Each participant was tasked with reformulating approximately 20 brief sentences, with the resulting adaptations categorized according to three major dialectal variants: Western, Central, and Navarrese. To facilitate cross-dialectal comparison, a subset of 10 identical sentences was assigned to more than one annotator, enabling parallel dialectal representations. The demographic and professional characteristics of the annotators, including age, gender, and educational background, are detailed in Appendix B.

It should be noted that during data collection, a single annotator generated two types of variants for each sentence, including both dialectal variations and allocutive agreement forms in Basque. The allocutive system in Basque requires morphological marking of the addressee's gender (masculine/feminine) within the verbal form. Consequently, **XNLIeu$_{\mathbf{var}}$** exhibits a higher instance count (894) compared to the original **XNLIeu$_{\mathbf{native}}$** dataset (621), as shown in Table 2.

In terms of dialect distribution, 592 instances correspond to the Central dialect, usually associated with the province of Gipuzkoa, 240 instances to the Western dialect (West Gipuzkoa and Biscay), and just 63 instances to the Navarrese dialect, comprising 7% of the data. Thus, the Navarrese dialect is clearly under-represented in our data.

XNLIes$_{\mathbf{var}}$ XNLIeu$_{\text{native}}$ was automatically translated into Spanish using Claude 3.5 Sonnet[2], generating the **XNLIeu2es$_{\mathbf{native}}$** dataset and facilitating the creation of a parallel corpus for Basque and Spanish texts with their respective variants. Quality verification was conducted through manual review of the machine-generated translations, making sure that they constituted an authentic representation of Spanish language patterns. Finally, the translated corpus was provided to Spanish-language annotators for variant-specific adaptation.

The adaptation task involved six independent annotators, each assigned a set of 50 sentences for dialectal adaptation into their respective Spanish variants. They represented four distinct geographical locations: Cuba, Ecuador, Spain, and Uruguay. Two annotators from Spain performed adaptations into separate dialectal variants (Andalusian and Tenerife), resulting in a total of five Spanish dialectal variations in the final dataset. The demographic and professional characteristics of the annotators, including age, gender, and educational background, are documented in Appendix B.

Train	
Dataset	Instances
MNLI	392k
MNLIeu	392k
MNLIes	392k
Test	
XNLIeu	5010
XNLIes	5010
XNLIeu$_{\text{native}}$	621
XNLIeu2es$_{\text{native}}$	621
XNLIeu$_{\text{var}}$	894
XNLIes$_{\text{var}}$	666

Table 2: Datasets used for training and testing.

It is worth noting that some annotators found it difficult to add dialectal features to the standard sentences. This could be due to the high number of common examples in Spanish varieties (Lopetegui et al., 2025; Zampieri et al., 2024). In other words, the distinctions between Spanish varieties tend to be more homogeneous and thus contain less variation compared to Basque (Section 6).

Similar to the Basque adaptation, multiple dialectal variants were documented by some annotators. These variants exhibited phonological phenomena such as word-final /s/ deletion (e.g., *digamos* → *digamo*) and /s/ to /j/ substitution in word-final position (resulting in *digamoj*). Thus, XNLIes$_{\text{var}}$ contains 666 examples, representing a marginally higher count than the base dataset.

Table 2 provides an overview of the datasets used for experimentation, including our newly generated **XNLIvar**, consisting of XNLIeu$_{\text{var}}$ and XNLIes$_{\text{var}}$.

4 Experimental settings

Empirical research was based on the aforementioned datasets to evaluate the impact of dialectal variation on NLI performance.

Discriminative experiments Table 3 illustrates the experiments performed using encoder-only Transformer models and the datasets specified in Table 2.

- **Model transfer:** The train split of the original MNLI (English) is used to fine-tune multilingual encoder models. Evaluation is performed on the test sets for Basque and Spanish specified in Table 2.

[2] https://www.anthropic.com/news/claude-3-5-sonnet

- **Translate-train:** The MNLI training is automatically translated into Basque and Spanish (MNLI$_{eu}$ and MNLI$_{es}$); multilingual and monolingual encoders are then fine-tuned using the translated training data and evaluated in each of the target languages.
- **Translate-test:** Tests in the target languages are translated into English and evaluated using the MNLI fine-tuned encoders (in English).

Configuration	Train	Test
Model transfer	English	Target language
Translate-train	Target language	Target language
Translate-test	English	Target → English

Table 3: Discriminative model configurations and data. **→:** Translated to.

Summarizing, training is always done with MNLI, either in its original English form or using the automatically translated versions to Basque and Spanish. Moreover, there are three different test data types: (i) XNLI test data professionally translated into the target languages (XNLIeu, XNLIes) (ii) the manually created native Basque data and its translation to Spanish (XNLIeu$_{native}$, XNLIeu2es$_{native}$) and, (iii) the native datasets adapted to different variations for each of the target languages (XNLIeu$_{var}$, XNLIes$_{var}$).

We employed two multilingual encoder-only language models for our target languages: XLM-RoBERTa large (Conneau et al., 2020) and mDeBERTa (He et al., 2021). The hyperparameter configuration followed Heredia et al. (2024), implementing differential learning rates of 5e-5 and 10e-6 for BERT and RoBERTa architectures, respectively. All other parameters were maintained at their default values. The training process consisted of 10 epochs across all model configurations.

Generative experiments We experimented with generative LLMs to evaluate the decoders' ability to perform NLI when language variation is present. We started with a zero-shot setting, where we prompt LLMs to identify the NLI relation.

We also evaluated alternative prompting methodologies, specifically, few-shot and Chain of Thought (CoT) approaches. The few-shot prompt implemented a single example for each classification category. The CoT methodology incorporated detailed task-specific contextual information alongside a single example for each label.

To further evaluate the linguistic comprehension capabilities of LLMs with respect to Basque and Spanish variants, we implemented an alternative methodological approach by transforming the NLI task into a Question-Answering (QA) setting. In this experimental configuration, the input prompt was restructured as a question to be answered by the LLM, with the three possible answers based on the NLI inference labels. Zero-shot and few-shot prompting strategies kept the same. The complete set of prompt templates used across all task formulations is available in Appendix C.

We selected Llama-3.1-Instruct (8B and 70B versions) (Dubey et al., 2024) and Gemma 2 instruct (9B and 27B versions) (Mesnard et al., 2024) due to their strong performance in both Basque and Spanish languages[3] (Etxaniz et al., 2024; Figueras et al., 2025). In the next section we focus on the results obtained by the larger LLMs (performances with smaller LLMs in Appendix E).

5 Results

We first report the results obtained in the discriminative settings, while in Section 5.2, we discuss the results of in-context learning with LLMs.

5.1 Discriminative Experiments

By looking at the results reported in Table 4, the empirical results demonstrate a significant performance degradation when comparing XNLIeu and XNLIes against the native and variation datasets. This observation aligns with existing literature documenting the adverse effects of train-test distribution shifts in cross-lingual settings (Artetxe et al., 2020; Volansky et al., 2013). When comparing native and variation data results, where the only difference is the presence of dialectal data, we see a decrease in results. Therefore, results show that language models perform worse when variants are included in the NLI task.

By doing a cross-configuration analysis, we see that for Basque, the best results are obtained with XLM-RoBERTa in the translate-train for XNLIeu (83.42) and XNLIeu$_{var}$ (73.21), while for XNLIeu$_{native}$ (75.85), the train-test is superior. Overall, the empirical results demonstrate that the translate-train approach with XLM-RoBERTa yielded the best overall performance for Spanish and Basque. This suggests that training and evalu-

[3] https://hf.co/spaces/la-leaderboard/la-leaderboard

Basque									
	Model transfer			Translate-train			Translate-test		
	XNLIeu	XNLIeu$_{\text{native}}$	XNLIeu$_{\text{var}}$	XNLIeu	XNLIeu$_{\text{native}}$	XNLIeu$_{\text{var}}$	XNLIeu	XNLIeu$_{\text{native}}$	XNLIeu$_{\text{var}}$
XLM-RoBERTa large	80.00	72.09	68.24	**83.42**	75.63	**73.21**	-	**75.85**	71.63
mDeBERTa	78.95	70.21	67.26	81.42	72.14	69.77	-	72.68	70.28
Spanish									
	XNLIes	XNLIeu2es$_{\text{native}}$	XNLIes$_{\text{var}}$	XNLIes	XNLIeu2es$_{\text{native}}$	XNLIes$_{\text{var}}$	XNLIes	XNLIeu2es$_{\text{native}}$	XNLIes$_{\text{var}}$
XLM-RoBERTa large	83.05	74.02	73.07	**84.69**	**74.61**	**73.72**	-	73.86	71.77
mDeBERTa	82.02	74.13	71.57	83.27	72.25	70.77	-	72.30	69.89

Table 4: Accuracy results for Basque and Spanish discriminative experiments.

Basque										
	Llama-3.1-Instruct-70B					**Gemma-2-it-27B**				
	nli-zero	nli-few	qa-zero	qa-few	chain	nli-zero	nli-few	qa-zero	qa-few	chain
XNLIeu	33.65	53.17	33.31	54.89	**55.25**	61.10	62.81	61.84	**65.27**	58.28
XNLIeu$_{\text{native}}$	38.81	56.68	39.61	58.61	**60.71**	64.90	66.67	65.70	**68.28**	66.99
XNLIeu$_{\text{var}}$	33.78	48.66	31.54	**50.11**	49.22	57.61	60.96	57.49	**61.52**	58.05
Spanish										
	Llama-3.1-Instruct-70B					**Gemma-2-it-27B**				
	nli-zero	nli-few	qa-zero	qa-few	chain	nli-zero	nli-few	qa-zero	qa-few	chain
XNLIes	54.65	62.18	51.54	65.69	**73.97**	66.75	71.28	70.52	**73.05**	68.88
XNLIeu2es$_{\text{native}}$	62.96	62.48	62.16	70.69	**77.29**	71.50	72.62	73.91	73.43	**76.97**
XNLIes$_{\text{var}}$	59.42	62.32	54.27	69.24	**75.52**	70.37	72.30	72.79	72.14	**74.56**

Table 5: Results with generative LLMs.

ating in the target language constitutes the optimal method, irrespective of whether the data includes standard or variation-inclusive linguistic content.

Regarding *native* and *variant* results, analysis reveals that Spanish consistently outperforms Basque across all settings and evaluation datasets, demonstrating greater resilience to linguistic variation. In fact, while Spanish accuracy drops minimally (less than 1 percentage point in most cases), Basque performance suffers a higher decrease, with model-transfer and translate-test approaches showing an approximately 4-point drop and translate-train a 2.5-point drop. This highlights a sharper impact of variation on Basque performance.

These results show that when English is the source training data, model-transfer provides competitive results for a high-resource, structurally similar language such as Spanish, while for a low-resource and morphologically different language such as Basque, the data-transfer (translate-train) strategy remains preferable (Agerri et al., 2020; Artetxe et al., 2020; García-Ferrero et al., 2022).

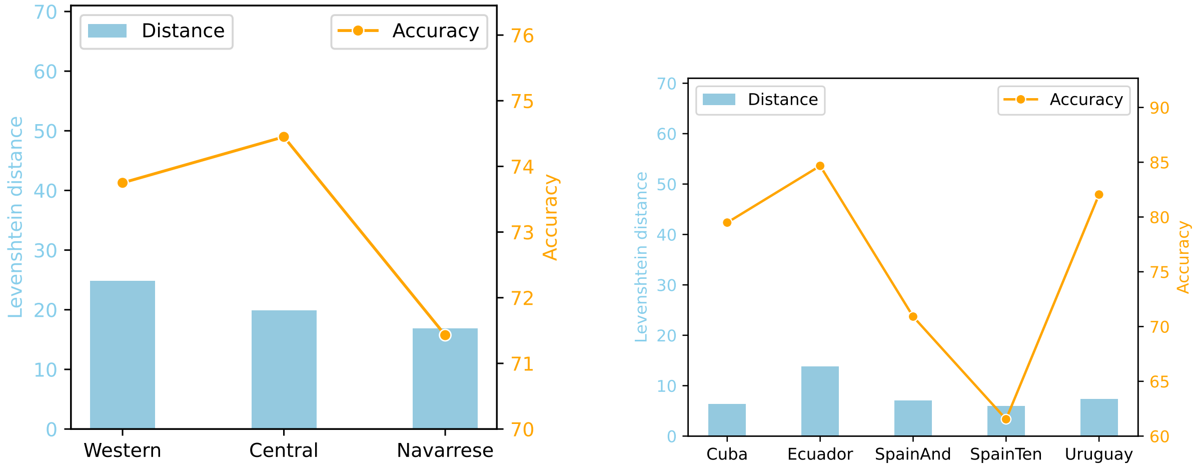
Overall, the consistently lower performance observed in Basque relative to Spanish across all evaluation conditions can be attributed to three key factors: (i) Basque's agglutinative morphological structure, (ii) its classification as a language isolate, and (ii) reduced Basque language representation in the models' pre-training data (Agerri et al., 2020; Etxaniz et al., 2024).

Finally, we also experimented with two Basque monolingual models, RoBERTa-Euscrawl (Artetxe et al., 2022) and BERTeus (Agerri et al., 2020), in the translate-train setting. However, while competitive, their results did not outperform those obtained by XLM-RoBERTa large. Further details can be found in Appendix D.

5.2 Generative Experiments

Table 5 presents the evaluation results for LLMs in the task on variation-inclusive NLI with the largest LLMs tested, namely, Llama-3.1-Instruct-70B and Gemma-2-it-27B.

A first observation reveals a significant performance degradation across all evaluated LLMs when transitioning from standard datasets (XNLIeu$_{\text{native}}$ and XNLIeu2es$_{\text{native}}$) to their variant counterparts (XNLIeu$_{\text{var}}$ and XNLIes$_{\text{var}}$). This suggests a substantial limitation in the capacity of LLMs to process and comprehend linguistic variations within the task. The results also indicate that including examples in the prompt engineering pro-



(a) Basque results

(b) Spanish results

Figure 1: Standard to dialectal Levenshtein distance vs accuracy of best discriminative models.

cess yields positive effects (in the qa-few and CoT methodologies). Notably, for Spanish, the CoT approach demonstrates superior performance compared to XLM-RoBERTa large on XNLIeu2es$_{\text{var}}$ and XNLIes$_{\text{var}}$ datasets.

Concerning Basque, the experimental results demonstrate that Gemma-2 exhibits better performance compared to Llama-3.1. Moreover, for XNLIeu$_{\text{var}}$ Gemma's optimal performance (61.52) experiences a reduction of 6.5 percentage points relative to the standard XNLIeu$_{\text{native}}$ (68.28). In contrast, Llama-3.1 exhibits a more substantial decline of 10 percentage points in XNLIeu$_{\text{var}}$ performance. These findings indicate that Gemma maintains greater robustness against linguistic variation compared to Llama-3.1.

For Spanish, CoT prompting generally yields the highest accuracy. The variation-inclusive evaluation dataset (XNLIes$_{\text{var}}$) produces results very close to those of XNLIes$_{\text{native}}$, with Llama 3.1 achieving 75.77 and 77.29, and Gemma 2 reaching 74.56 and 76.56, respectively. Despite this closeness, linguistic variation still causes a drop in accuracy. Overall, Llama 3.1 performs slightly better than Gemma 2, though the difference is minimal.

The empirical evidence obtained from these analyses of Basque and Spanish language understanding indicates that LLMs exhibit significant limitations in their capacity to comprehend linguistic content when confronted with dialectal and geographical variations.

6 Error analysis

This section presents a quantitative error analysis to evaluate XLM-RoBERTa large's performance with respect to variation-inclusive evaluation data.

Dialect to standard distance The Levenshtein distance metric, which quantifies the minimum number of single-character operations (insertions, deletions, or substitutions) necessary for string transformation, was computed between dialectal and standard sentences. The analysis of distance results demonstrates that Basque dialectal variants (Figure 1a) exhibit significantly greater divergence from the standard form compared to Spanish variants (Figure 1b), which display higher proximity to their standardized counterpart. The observed inter-dialectal variation patterns suggest a more pronounced linguistic differentiation within Basque dialectal systems relative to Spanish dialectal varieties. This emphasizes the difference in variation between languages and highlights the importance of language-specific analysis in the field of language variation processing in NLP.

Accuracy per dialect We analyzed the accuracy results for each individual dialect class, in order to see if some dialects are more difficult to process than others. The relation between the accuracy for each dialect and the distance from standard to dialect is illustrated in Figure 1.

In the case of Basque (Figure 1a), we see that, in terms of string distance, the Western dialect is the one that is the most different from the standard, followed by the Central and Navarrese dialects. However, the lowest accuracy is accounted for in the Navarrese dialect, which is the dialect label that seems to be closest to the standard form of language. This could be because of its underrepresentation in our dataset, as Navarrese examples comprise only 7% of our data (Appendix B). When focusing on Western and Central dialects, it can be observed that, as the distance from standard to dialectal gets higher, accuracy gets lower, suggesting that dialects further from the standard (in our case, Western) are harder to process. The Central dialect being closer to the standard is expected, as it served as the main foundation for the current standard form of Basque.

In fact, according to research in Basque dialectology, peripheral dialects have been found to be more distant from the rest (Mitxelena, 1981). This fact has also been corroborated by NLP studies analyzing Basque historical dialects, where Biscayan (Western) and Souletin display the greatest difference (Estarrona et al., 2023). Additionally, research has documented the Bizkaian dialect's historical tendency toward linguistic divergence, both from other Basque dialects and from its own earlier forms (Zuazu, 2015).

In Spanish variants, Figure 1b shows Ecuador and Uruguay displaying the highest distance values and accuracy scores. Further analysis has shown that adaptations into these two variants mostly include replacing lexical words with alternatives that are more commonly used in those varieties (e.g., *construccion futura* > *nuevos edificios*), as well as grammatical structures typical of those dialects (e.g., *he podido* > *pude*). However, standard orthography has been preserved throughout.

In turn, adaptations into Cuban, SpainAnd and SpainTen variants mostly include phonological or orthography changes (e.g. *misma* > *mihma, fuerza* > *fuersa*), which have resulted in lower distance to the standard form of Spanish, but a decrease in accuracy compared to the variants written in standard orthography (Ecuador and Uruguay). This reveals a correlation between standard orthography and high accuracy, and highlights the difficulties of discriminative models to deal with data which includes non-standard orthography. This analysis is illustrated in Appendix G. These results match those observed in earlier studies, where orthography variations have also been found to be problematic (De la Rosa et al., 2024). Additional results of per-dialect accuracy results are presented in Appendix F.

Dataset	Instances	Discrimin.	Genera.
Basque			
XNLIeu$_{\text{native}}$	621	75.63	68.28
XNLIeu$_{\text{var}}$	894	73.21	61.52
Less-western	834	73.14	60.79
Less-central	834	72.70	61.03
No repetitions	621	71.77	60.39
Spanish			
XNLIeu2es$_{\text{native}}$	621	74.61	77.29
XNLIes$_{\text{var}}$	666	73.72	75.52
No repetitions	621	73.00	77.13

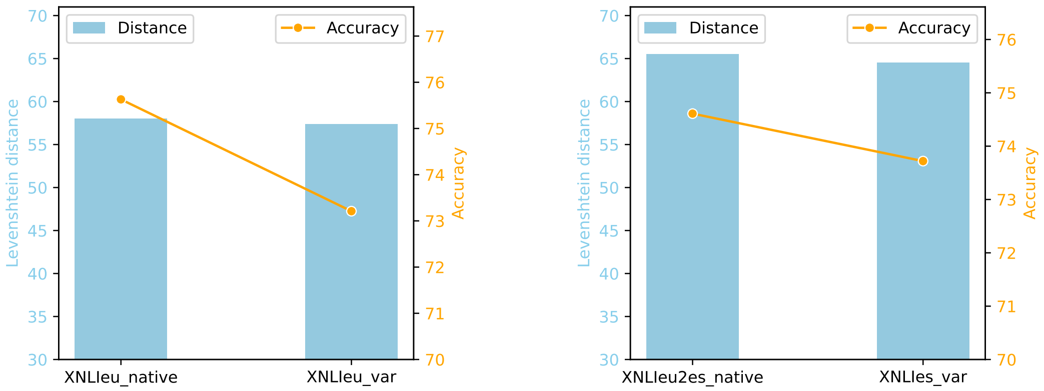
Table 6: Ablation experiments on Basque and Spanish variation data (XNLIeu$_{\text{var}}$ and XNLIes$_{\text{var}}$, respectively). Results obtained using the best discriminative setting (Translate-train XLM-RoBERTa large in Table 4) as well as best generative results for Basque (Gemma-2 qa-few) and Spanish (Llama-3.1 chain) in Table 5.

Ablation Tests As explained in Section 3.1 and illustrated in Table 2, test data in XNLIeu$_{\text{var}}$ and XNLIes$_{\text{var}}$ contains duplicated instances in different dialects. In order to see the effect that different types of variation have on accuracy, we have performed some ablation experiments.

Four different Basque speakers (two Western and two Central) adapted the same 10 sentences, providing us with four distinct versions of those 10 sentences. We used these instances to create two new versions of the dataset, one by removing the repeated sentences from Western variants (*Less-western*), and another one without the repeated instances from the Central one (*Less-central*). Table 6 presents accuracy results with these datasets.

The results show that accuracy is higher when Western-dialect instances are removed (73.14) than when Central instances are excluded (72.70).

Additionally, we removed all duplicated variant instances from XNLIvar$_{\text{eu}}$, resulting in a completely parallel variation dataset to XNLIeu$_{\text{native}}$ (*No repetitions*), which allows us to calculate whether the results between the standard and the variant versions are statistically significant. As reported in Table 6, accuracy between *No repetitions* and the *standard* substantially decreases (71.77 vs 75.63) for the Basque discriminative experiments. According to a chi-square test of independence, this difference is highly statistically significant (p < .001, df=1). Similar to Basque, all the repeated variant instances from the Spanish



(a) Premise-Hypothesis distance and accuracy for Basque

(b) Premise-Hypothesis distance and accuracy for Spanish

Figure 2: Levenshtein distance from premise to hypothesis and accuracy of discriminative models.

variation dataset were removed, obtaining a parallel dataset to $\text{XNLIes}_{\text{native}}$. Using the *No repetitions* split, a chi-square test of independence establishes that differences with results on $\text{XNLIeu2es}_{\text{native}}$ are highly statistically significant ($p < .001$, df=1).

Generative LLMs follow the same trend, with differences in performance being highly statistically significant in both languages ($p < .001$, df=1).

Premise and hypothesis lexical overlap To investigate the potential correlation between lexical overlap and accuracy, we measured the Levenshtein distance between premises and hypotheses. The analysis of the data in Figure 2 indicates that lexical overlap remains consistent across standard and dialectal varieties, while a substantial decrease in accuracy was observed in both Basque and Spanish datasets. These findings suggest that while lexical overlap appears to have minimal impact on accuracy metrics, linguistic variation emerges as the significant factor affecting performance. Therefore, the observed pattern implies that dialectal variations, rather than lexical similarities, may be the primary factor of accuracy degradation in this context.

In fact, Figure 2a demonstrates a more pronounced decrease in accuracy for Basque compared to Spanish, underscoring both the critical need to improve Basque representation in multilingual discriminative models and the necessity for additional investigation into language variation processing.

7 Concluding Remarks

This paper presents a novel dataset that includes geographical variants of Basque and Spanish. The dataset represents the first documented instance of a manually-curated, variation-inclusive corpus for these languages, facilitating research and evaluation on linguistic variants via NLI. Additional speaker metadata expands its value as a resource for sociolinguistic research on generational and geographical differences in Basque and Spanish. Our investigation involved the empirical evaluation of both discriminative and generative language models across various NLI task configurations.

Results indicate that language models' performance drops when linguistic variation is present. This performance degradation is particularly pronounced in Basque variants, where linguistic variation is higher compared to Spanish variants. Furthermore, the performance drop intensifies proportionally with the linguistic distance between dialectal variants and their respective standardized forms for Basque, with a higher impact in the Western dialect. This coincides with previously established linguistic theory, which states that some Basque dialects (such as Western) have a historical tendency to distance themselves from the standard. In the case of Spanish, variants with non-standard orthography have shown a significant accuracy drop. Finally, the lexical overlap between premises and hypotheses appears to have minimal impact, suggesting that lower performance is due to linguistic variation.

Future work will involve expanding the dataset to include additional geographical variants of both Basque and Spanish, as well as incorporating other languages. Investigation of variation-inclusive monolingual models represents a promising avenue for future research.

Limitations

In this paper, we have focused on geographic variants of language due to their low representation in NLP. We conducted our experiments for a lesser-resourced language, Basque, and a higher-resourced language, Spanish. However, we have only represented some of the variations of these languages, and our variation datasets have been created by 12 speakers for Basque and 6 speakers for Spanish. We tried to include the most representative dialects with different kinds of speakers, but we are aware that all the speakers have linguistic and NLP backgrounds, and laypeople could contribute differently.

Our empirical findings demonstrate decreased accuracy in natural language inference tasks within our variants dataset. However, generalization of these results requires expansion to include additional linguistic variants and evaluation across a broader range of NLP tasks.

To augment the dataset, we are recruiting speakers from diverse linguistic backgrounds to contribute additional variation data. We further intend to evaluate the performance of NLP tools and LLMs on tasks incorporating dialectal and register variation.

Acknowledgments

This work has been supported by the HiTZ center and the Basque Government (Research group funding IT-1805-22). Jaione Bengoetxea is funded by the Basque Government pre-doctoral grant (PRE_2024_1_0028). We also acknowledge the following MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and by FEDER, EU; (ii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR.

We acknowledge the contributions of the following people to generate the Basque and Spanish linguistic variants: Ainara Estarrona, Alex Muñoz Alvarado, Ekhi Azurmendi, Irune Zubiaga, Itziar Aldabe, Izaskun Aldezabal, Jon Alkorta, Jose Maria Arriola, Leonel Cosme, María Teresa Martín-Valdibia, Marta Torres Martínez, Mikel Zubillaga, Naiara Perez, Nancy Cristina Álamo Suárez, Oier Ijurco, Yanisbel Ríos Laborde and Yoandra Chuen Gómez.

References

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

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.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. CODET: A benchmark for contrastive dialectal evaluation of machine translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

Tahani Alshareef and Muazzam Ahmed Siddiqui. 2020. A seq2seq neural network based conversational agent for gulf arabic dialect. In *2020 21st International Arab Conference on Information Technology (ACIT)*, pages 1–7.

Ekaterina Artemova and Barbara Plank. 2023. Low-resource bilingual dialect lexicon induction with large language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 371–385, Tórshavn, Faroe Islands. University of Tartu Library.

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 116–128, New York, NY, USA. Association for Computing Machinery.

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

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.

Eugenio Coseriu. 1956. *La geografía lingüística*, volume 11. Universidad de la República, Facultad de Humanidades y Ciencias.

Javier De la Rosa, Álvaro Cuéllar, and Jörg Lehmann. 2024. The modernifa project: orthographic modernization of spanish golden age dramas with language models. *Anuario Lope de Vega Texto literatura cultura*, 30:410–425.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, (...), and Zhiwei Zhao. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2024. Elote, choclo and mazorca: on the varieties of Spanish. 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 3689–3711, Mexico City, Mexico. Association for Computational Linguistics.

Ainara Estarrona, Izaskun Etxeberria, Ricardo Etxepare, Manuel Padilla-Moyano, and Ander Soraluze. 2020. Dealing with dialectal variation in the construction of the Basque historical corpus. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 79–89, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Ainara Estarrona, Izaskun Etxeberria, Manuel Padilla-Moyano, and Ander Soraluze. 2023. Measuring language distance for historical texts in basque. *Procesamiento del Lenguaje Natural*, 70:53–61.

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.

Fahin Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. *ArXiv*, abs/2403.11009.

Blanca Calvo Figueras, Eneko Sagarzazu, Julen Etxaniz, Jeremy Barnes, Pablo Gamallo, Iria De Dios Flores, and Rodrigo Agerri. 2025. Truth knows no language: Evaluating truthfulness beyond english. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416.

Hany Hassan, Mostafa Elaraby, and Ahmed Y. Tawfik. 2017. Synthetic data for neural machine translation of spoken-dialects. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 82–89, Tokyo, Japan. International Workshop on Spoken Language Translation.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

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.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *Preprint*, arXiv:2401.05632.

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.

William Labov. 2006. *The Social Stratification of English in New York City*, 2 edition. Cambridge University Press.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452.

Javier A. Lopetegui, Arij Riabi, and Djamé Seddah. 2025. Common ground, diverse roots: The difficulty of classifying common examples in Spanish varieties. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 168–181, Abu Dhabi, UAE. Association for Computational Linguistics.

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Rivière, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, Anna Bulanova, (...), and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

Luis Mitxelena. 1981. Lengua común y dialectos vascos. *Anuario del Seminario de Filología Vasca" Julio de Urquijo"*, 15:289–313.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

Alan Ramponi and Camilla Casula. 2023. DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 187–199, Dubrovnik, Croatia. Association for Computational Linguistics.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2024. Meta4xnli: A crosslingual parallel corpus for metaphor detection and interpretation. *arXiv*, 2404.07053.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388.

Larraitz Uria and Ricardo Etxepare. 2012. Hizkeren arteko aldakortasun sintaktikoa aztertzeko metodologiaren nondik norakoak: Basyque aplikazioa. *Lapurdum. Euskal ikerketen aldizkaria| Revue d'études basques| Revista de estudios vascos| Basque studies review*, (16):117–135.

Francisco Valentini, Viviana Cotik, Damián Ariel Furman, Ivan Bercovich, Edgar Altszyler, and Juan Manuel P'erez. 2024. Messirve: A large-scale spanish information retrieval dataset. *ArXiv*, abs/2409.05994.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Mahesh Bangera. 2024. Language variety identification with true labels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10100–10109, Torino, Italia. ELRA and ICCL.

Koldo Zuazu. 2008. *Euskalkiak. Euskararen dialektoak*. Elkar.

Koldo Zuazu. 2015. The unification of the basque language. `http://euskalkiak.eus/en/euskararen_batasuna.php`. Accessed: 09-06-2025.

A Guidelines for Variation Adaptation

Language	Adaptation guidelines
Basque	Ataza honetan testu motz batzuk hizkuntza formal/estandarretik hizkuntza informalagora/euskalkietara berridatzi behar dira. Bakoitzak bere hizkuntza informal/dialektalean esango lukeen bezala idaztea da helburua. Hau horrela, ondorengo aldaketak proposatzen ditugu: • Esamolde edo hizkuntza informalagoa bilakatu. • Ezaugarri dialektalak gehitu, bai lexiko aldetik eta bai gramatika edo fonetika aldetik. • Hika. Erregistroa edo dialektoak barne hartzen dituen beste edozein aldaketa ongietorria da. Adibidez: **Jatorrizkoa:** Bi dantzari horiek dantza hunkigarria eskaini zuten herriko frontoian. **Berridatzia:** Bi dantsari hoiek dantza emozionantia eskeiñi zuten herriko frontoien.
Spanish	En esta tarea se deben reescribir algunos textos cortos del lenguaje formal/estándar a un lenguaje más informal/dialectal. El objetivo es adaptar las frases como cada persona lo diría en su propio lenguaje dialectal. De esta manera, se propone hacer los siguientes cambios: • A nivel de registro: Más informal, reescribiéndola de manera más coloquial • Con rasgos dialectales, sean léxicos, gramaticales o fonéticos • Adaptar la ortografía para que refleje vuestra pronunciación, dialecto Cualquier otro cambio que refleje un cambio de registro o dialecto es bienvenido. Por ejemplo: **Frase original:** El amigo se quedó sin opciones cuando le dijeron que el autobús no pasaría más. **Frase adaptada:** El socio se quedo botao cuando le dijeron que la guagua no pasaba ma.
English	This task involves rewriting short texts from formal/standard language to a more informal/dialectal language. The objective is to rewrite the sentences as each person would say them in their own dialectal language. The following changes are proposed: • At the register level: More informal, rewriting in a more colloquial manner • With dialectal features, whether lexical, grammatical, or phonetic • Adapting spelling to reflect your pronunciation, dialect Any other changes that reflect a change in register or dialect are welcome. For example: **Original phrase:** Everyone, hurry up now, dinner is about to get cold. **Adapted phrase:** Y'all better hurry up now, supper's fixin' to get cold.

Table 7: Guidelines for standard to dialectal adaptations, both in Basque and Spanish, and an English translation

B Adaptation Process Information

B.1 Annotator Metadata

Variable	Category	N	%
Location	Gipuzkoa	7	58.34
	Biscay	4	33.34
	Navarre	1	8.34
Age	20-30	5	41.67
	30-40	3	25.00
	40+	4	33.34
Gender	Male	5	41.67
	Female	7	58.34
Background	Linguist	8	66.67
	Non-linguist	4	33.34

(a) Demographic metadata of annotators. **N** = Count; % = Percentage

Variable	Category	N	%
Location	Cuba	2	33.33
	Ecuador	1	16.67
	SpainAndalusia	1	16.67
	SpainTenerife	1	16.67
	Uruguay	1	16.67
Age	20-30	1	16.67
	30-40	3	50.00
	40+	2	33.33
Gender	Male	2	33.34
	Female	4	66.67
Background	Linguist	3	50.00
	Non-linguist	3	50.00

(b) Demographic metadata of annotators. **N** = Count; % = Percentage

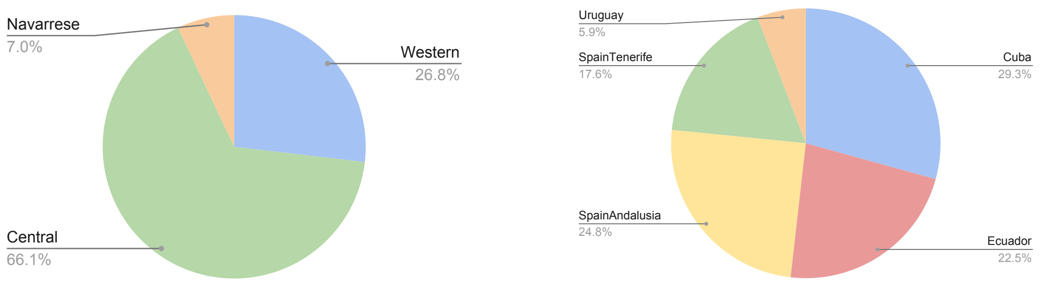
Table 8: Annotator metadata

B.2 Adaptation Type

	Basque		**Spanish**	
Change type	**N**	**%**	**N**	**%**
re-write	18	6.04	61	27.48
dialectal	223	74.83	161	72.52
Allocutive_masc	37	12.41	-	-
allocutive_fem	20	6.71	-	-
Total	298		222	

Table 9: Number and percentage of change types in Basque and Spanish data. **N**: Count of examples; **%**: Percentage

B.3 Geographical Variants Distribution in Data



(a) Number of examples per geographical variants in Basque

(b) Number of examples per geographical variants in Spanish

Figure 3: Geographical variant label representation in XNLIvar

C Prompts

Task formulation	Prompt
nli-zero	Please, answer in one word, with one of the following labels: <entailment>, <contradiction> or <neutral> Use exactly one of these three labels.
nli-few	"Say which is the inference relationship between these two sentences. Please, answer in one word, with one of the following labels: <entailment>, <contradiction> or <neutral> Use exactly one of these three labels. Here you have some examples: Postal Service were to reduce delivery frequency -> The postal service could deliver less frequently: <entailment>. This elegant spa town on the edge of the Lac du Bourget has offered cures for rheumatism and other ailments for centuries -> The town was only established in the past fifty years: <contradiction>. And while we allow people to give a kidney to their child, we do not allow them to donate their heart -> You can't always donate organs to your child: <neutral>.
qa-zero	Are these two sentences entailed, contradicted or undetermined to each other? Please, answer in one word, with one of the following labels: <entailment>, <contradiction> or <neutral> Use exactly one of these three labels.
qa-few	Are these two sentences entailed, contradicted or undetermined to each other? Please, answer in one word, with one of the following labels: <entailment>, <contradiction> or <neutral> Use exactly one of these three labels. Here you have some examples: Postal Service were to reduce delivery frequency -> The postal service could deliver less frequently: <entailment>. This elegant spa town on the edge of the Lac du Bourget has offered cures for rheumatism and other ailments for centuries -> The town was only established in the past fifty years: <contradiction>. And while we allow people to give a kidney to their child , we do not allow them to donate their heart -> You can't always donate organs to your child: <neutral>.
chain	You are an expert linguist and your task is to annotate sentences for the task of Natural Language Inference. This task consists in determining if a first sentence (premise) entails, contradicts or does not entail nor contradict the second sentence (hypothesis). Please, answer in one word, with one of the following labels: <entailment>, <contradiction> or <neutral> \n Use exactly one of these three labels \n Here you have a few examples:\n Premise: Postal Service were to reduce delivery frequency. \n Hypothesis: The postal service could deliver less frequently. \n Answer: <entailment> \n Premise: This elegant spa town on the edge of the Lac du Bourget has offered cures for rheumatism and other ailments for centuries. \n Hypothesis: The town was only established in the past fifty years. \n Answer: <contradiction> \n Premise: And while we allow people to give a kidney to their child , we do not allow them to donate their heart. \n Hypothesis: You can't always donate organs to your child. \n Answer: <neutral>

Table 10: Different task formulation prompts for generative model prompting

D Basque Monolingual Discriminative Results

	Translate-train		
	XNLIeu	XNLIeu$_{\text{native}}$	XNLIeu$_{\text{var}}$
RoBERTa-Euscrawl	82.63	73.43	72.24
BERTeus	78.15	68.81	63.67

Table 11: Accuracy results for Basque monolingual discriminative experiments

E Additional Generative Results

	Llama-3.1-Instruct-8B					Gemma-2-it-9B				
	nli-zero	nli-few	qa-zero	qa-few	chain	nli-zero	nli-few	qa-zero	qa-few	chain
XNLIeu	20.30	16.63	09.16	38.50	51.76	55.61	38.66	37.96	44.51	48.88
XNLIeu$_{\text{native}}$	21.36	17.90	07.05	36.24	41.83	61.19	39.94	41.55	49.11	54.43
XNLIeu$_{\text{var}}$	20.45	13.04	14.33	37.04	46.22	53.47	39.04	36.13	41.72	45.53

(a) Accuracy results with generative LLMs on Basque data.

	Llama-3.1-Instruct-8B					Gemma-2-it-9B				
	nli-zero	nli-few	qa-zero	qa-few	chain	nli-zero	nli-few	qa-zero	qa-few	chain
XNLIes	27.96	22.87	16.43	49.28	57.78	64.11	55.57	44.73	57.41	66.83
XNLIeu2es$_{\text{native}}$	28.34	15.62	23.19	48.79	62.80	69.24	55.72	50.24	59.42	71.82
XNLIes$_{\text{var}}$	26.73	21.62	19.37	48.35	56.46	67.63	53.14	44.28	55.72	68.92

(b) Accuracy results with generative LLMs on Spanish data.

Table 12: Results with 8B and 9B LLMs.

F Per-dialect Accuracy Results

	Model transfer			Translate-train			Translate-test		
	Western	Central	Navarrese	Western	Central	Navarrese	Western	Central	Navarrese
XLM-RoBERTa large	71.25	67.17	71.43	73.75	74.45	71.43	71.67	72.42	74.60
mDeBERTa	62.08	70.90	60.32	66.25	72.59	66.67	69.58	71.40	73.02

Table 13: Accuracy results for discriminative models in Basque dialects

	Model transfer				
	Cuba	Ecuador	SpainAndalusia	SpainTenerife	Uruguay
XLM-RoBERTa large	79.49	78.67	69.70	58.97	82.05
mDeBERTa	76.92	78.67	70.30	58.97	79.49
	Translate-train				
	Cuba	Ecuador	SpainAndalusia	SpainTenerife	Uruguay
XLM-RoBERTa large	79.49	84.67	70.91	61.54	82.05
mDeBERTa	73.33	81.33	69.09	53.85	87.18
	Translate-test				
	Cuba	Ecuador	SpainAndalusia	SpainTenerife	Uruguay
XLM-RoBERTa large	72.82	76.67	72.12	64.10	76.92
mDeBERTa	69.23	77.33	67.88	62.39	79.49

Table 14: Accuracy results for discriminative modes in Spanish variants

G Spanish Correlation Between Adaptation Types and Accuracy

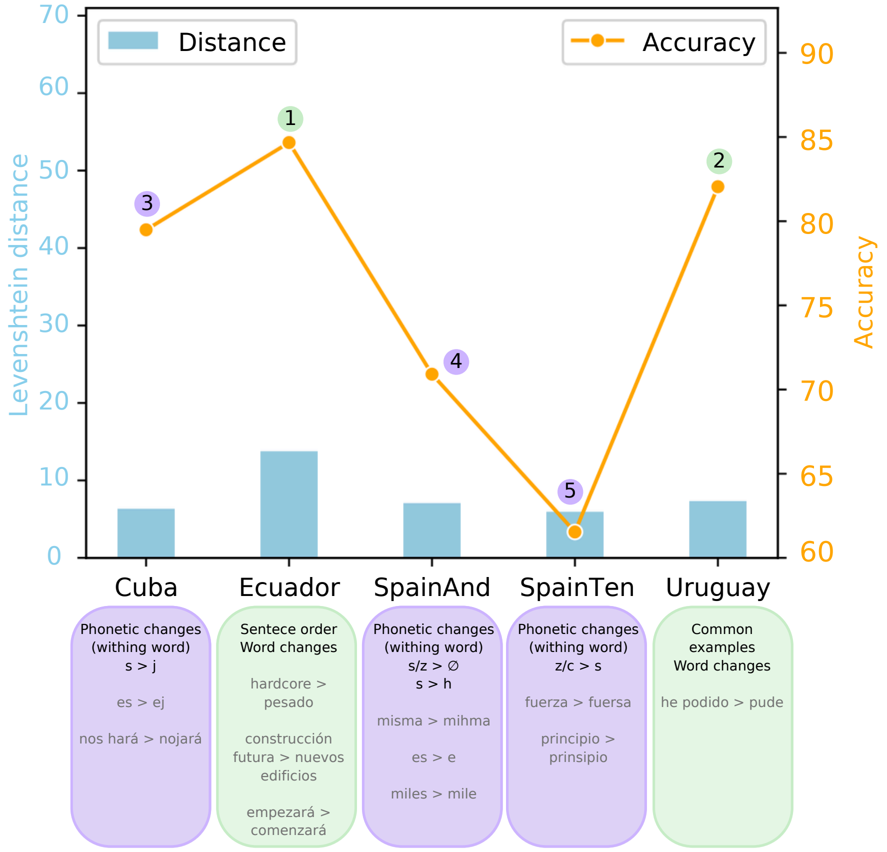


Figure 4: Spanish accuracy results and its correlation to types of linguistic adaptations. 1 and 2 have the highest accuracies, but changes usually involve word changes. For 3 , 4 and 5 , the accuracy decreases respectively, as variations majorly involve phonetic changes.