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

Anonymous ACL submission

Abstract

In this paper, we evaluate the capacity of current language technologies to understand Basque and Spanish language varieties. We use NLI as a pivot task and introduce a novel, manually-curated parallel dataset in Basque and Spanish and their corresponding variants. Empirical analysis of comprehensive crosslingual and in-context learning experiments with, respectively, encoder-only and decoder-based Large Language Models (LLMs), reveals a performance drop when processing linguistic variations, with more pronounced effects observed in Basque. Error analysis indicates that lexical overlap plays no role, suggesting that linguistic variation represents the primary reason for the lower results. All data and code are publicly available[1] under Attribution-NonCommercial 4.0 International license.

1 Introduction

Variation is an intrinsic characteristic of language, influenced by multiple factors. Trask and Stockwell (2007) noted that "variation, far from being peripheral and inconsequential, is a vital part of ordinary linguistic behavior". In sociolinguistics, three types of variability are considered (Coseriu, 1956): (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. In this paper, we focused on geographical variation in Basque, a low-resource language isolate with around 1 million speakers that is still undergoing the normalisation process (started in 1968), and in Spanish, a higher resourced language whose standardization process started in the XVIII century with around 600 million speakers.

With the rise of Artificial Intelligence (AI) and Natural Language Processing (NLP), Hovy and Yang (2021) have highlighted the importance of social factors of language in NLP systems. Thus, developing NLP resources that not only process standard language but also variations is crucial, as this would alleviate any potential linguistic discrimination by providing more linguistically-inclusive resources.

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 context, although previous work on NLP has primarily focused on standard language, recent research has slightly shifted attention to the exploration of language variation, e.g., Zampieri et al. (2020) or Joshi et al. (2024) present a thorough outline of variation-inclusive research. However, due to a lack of data on linguistic variants, most of the NLP research has focused on a narrow list of languages and their variants, such as Arabic, Indic languages, or German. Furthermore, other larger efforts are either based on automatically obtained data or do not provide fine-grained variant distinctions for some large languages, such as Spanish (Faisal et al., 2024; Alam et al., 2024).

Regarding Basque, the scarce work that has been done has 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

[1] https://anonymous.4open.science/r/XNLIvar

through geolocation techniques (España-Bonet and Barrón-Cedeño, 2024; Valentini et al., 2024).

Consequently, the objective of this paper is to provide the first manually curated variation dataset for Basque and Spanish. To do so, we introduce XNLIvar, the first variation-inclusive NLI dataset in Basque and Spanish. An example of an instance from our dataset is in Table 1. Additionally, we aim to evaluate its performance with current state-of-the-art language models. We make the following contributions:

1. The first publicly available manually-curated NLI dataset of Basque and Spanish with geographic language variations.
2. A comprehensive evaluation of encoder-only and decoder-based Large Language Models (LLMs) demonstrates substantially worse performance when processing language variations, particularly in Basque. Detailed error analysis shows that lexical overlap has no impact on the performance drop, which means that linguistic variation must be considered the most significant factor.
3. Empirical results suggest that the performance of LLMs with Spanish variants may be attributed to the substantial representation of Spanish-language content in pre-training corpora.

To the best of our knowledge, no work has extensively dealt with the automatic processing of current language variation of western and central Basque dialects in the task on NLI.

2 Related Work

This section presents previous work on language variation in the field of NLP, with specific focus on up-to-now research 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 variation 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: 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.

The lack of comprehensive dialectal data has led to research on linguistic variation to be limited to certain languages. The Arabic dialect family, due to its relative data availability, has received the most attention, followed by languages such as Indic languages, Chinese and German.

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, some works, such as 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) also collected tweets from European and Latin American geolo-

cations and annotated them specifically for hate speech.

3 Data

In this work, we introduce a novel dataset, **XNLIvar**, that expands the XNLI framework by generating various dialectal variations for both Basque and Spanish languages. We choose Natural Language Inference (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). NLI, or textual entailment, is a fundamental NLP task that involves determining the semantic relationship between two text segments: a premise and a hypothesis. The objective is to classify whether the hypothesis can be inferred from the premise into 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), which has contributed significantly to advancing research in this field.

XNLI provides an extension of MultiNLI in 15 languages, among them Spanish (**XNLIes**). The authors collected 7,500 English examples via crowdsourcing, which were then professionally translated to create XNLI in two splits, the development (2500 instances) and the test set (5K instances). 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 Natural Language Inference (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 **XNLIvar** novel 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 languages, 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 selected annotators, 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 the data collection a single annotator generated two types of variants of 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$_{\text{var}}$** exhibits a higher instance count (894) compared to the original **XNLIeu$_{\text{native}}$** 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$_{\text{var}}$ XNLIeu$_{\text{native}}$ was automatically translated into Spanish using the LLM Claude 3.5 Sonnet[2], generating the **XNLIeu2es$_{\text{native}}$** 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. The annotators 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.

It is worth mentioning that some annotators experienced difficulties during the adaptation. 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.

As it was the case in 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). Consequently, the XNLIes$_{\text{var}}$ dataset contains 666 examples, representing a marginally higher count than the base dataset.

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

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

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 conducted utilizing the aforementioned datasets to evaluate the impact of dialectal variation incorporation on NLI performance. The experimental methodology consisted of both discriminative and generative modeling approaches.

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.
- **Translate-train:** The MNLI training is automatically translated into Basque and Spanish (MNLI$_{\text{eu}}$ and MNLI$_{\text{es}}$); 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 ($\text{XNLIeu}_{\text{native}}$, $\text{XNLIeu2es}_{\text{native}}$) and, (iii) the native datasets adapted to different variations for each of the target languages ($\text{XNLIeu}_{\text{var}}$, $\text{XNLIes}_{\text{var}}$).

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

Alternative prompting methodologies were evaluated, specifically few-shot and Chain of Thought (CoT) approaches. The few-shot prompt implemented single examples 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) framework. In this experimental configuration, the input prompt was restructured as a question, wherein the LLM was given a question to answer given three labels. The three previously established prompting strategies were maintained across this methodological adaptation. The complete set of prompt templates used across all task formulations has been documented and is available for reference in the 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).

5 Results

We first report the results obtained in the discriminative settings, while in Section 5.2, we discuss the results obtained by applying 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 $\text{XNLIeu}_{\text{var}}$ (73.21), while for $\text{XNLIeu}_{\text{native}}$ (75.85), the train-test provides slightly better scores. Overall, the empirical results demonstrate that the translate-train approach with XLM-RoBERTa yielded the best overall performance for Spanish and Basque. This finding suggests that conducting both training and evaluation in the target language constitutes the optimal method, irrespective of whether the data includes standard or variation-inclusive linguistic content in either target language.

Looking at the *native* and *variant* results, the analysis reveals a consistent pattern in which Spanish performances exceed those of Basque across

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

	Basque								
	Model transfer			Translate-train			Translate-test		
	XNLIeu	XNLIeu$_{native}$	XNLIeu$_{var}$	XNLIeu	XNLIeu$_{native}$	XNLIeu$_{var}$	XNLIeu	XNLIeu$_{native}$	XNLIeu$_{var}$
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$_{native}$	XNLIes$_{var}$	XNLIes	XNLIeu2es$_{native}$	XNLIes$_{var}$	XNLIes	XNLIeu2es$_{native}$	XNLIes$_{var}$
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.

	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$_{native}$	38.81	56.68	39.61	58.61	**60.71**	64.90	66.67	65.70	**68.28**	66.99
XNLIeu$_{var}$	33.78	48.66	31.54	**50.11**	49.22	57.61	60.96	57.49	**61.52**	58.05

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

	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$_{native}$	62.96	62.48	62.16	70.69	**77.29**	71.50	72.62	73.91	73.43	**76.97**
XNLIes$_{var}$	59.42	62.32	54.27	69.24	**75.52**	70.37	72.30	72.79	72.14	**74.56**

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

Table 5: Results with LLMs.

all experimental settings and evaluation datasets. The data indicates that Spanish exhibits greater resilience to linguistic variation, as evidenced by the minimal performance degradation observed between standard data (XNLIeu2es$_{native}$) and variation datasets (XNLIes$_{var}$). Quantitative analysis of Basque performance shows substantial degradation, with model transfer and translate-test approaches experiencing approximately 4 percentage points decrease (from XNLIeu$_{native}$ to XNLIeu$_{var}$), while the translate-train methodology exhibits a more moderate reduction of 2.5 percentage points. In contrast, the drop in performance for Spanish remains minimal, with model-transfer and translate-train approaches showing less than 1 percentage point reduction, although the translate-test methodology demonstrates a decrease of 2 percentage points.

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).

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, 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. We limit our discussion to the results obtained by 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$_{native}$ and XNLIeu2es$_{native}$) to their variant counterparts (XNLIeu$_{var}$ and XNLIes$_{var}$). This suggests a substantial limitation in the capacity of LLMs to process and comprehend linguistic variations within the task. The data indicates that the incorporation of examples in the prompt engineer-

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

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

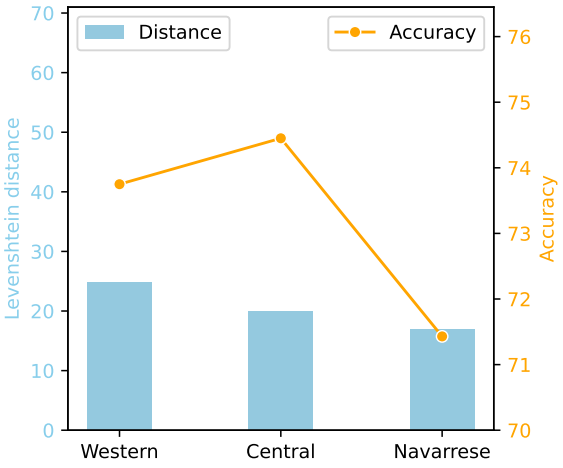
Regarding Spanish, Table 5b shows that CoT prompting seems to generally obtain the highest accuracy results. Moreover, the variation-inclusive evaluation dataset (XNLIes$_{var}$) obtains very close results to XNLIes$_{native}$, with 75.77 and 77.29 for Llama 3.1 and 74.56 and 76.56 in Gemma 2, respectively. However, variation still slightly hinders accuracy results in Spanish. In any case, Llama 3.1 performs slightly better than Gemma 2, although 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.

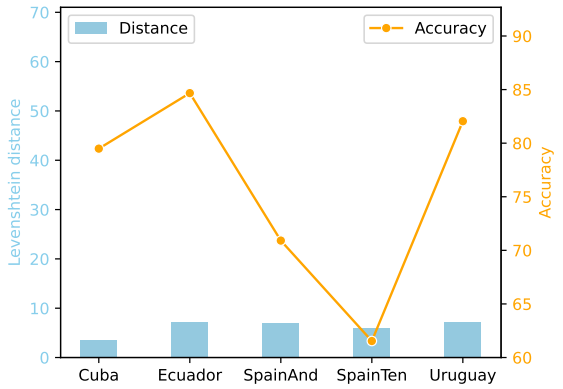
6 Error analysis

This section presents a quantitative error analysis to evaluate the 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. An analysis in Figure 1 of the resultant scores demonstrates that Basque dialectal variants exhibit significantly greater divergence from the standard form compared to Spanish variants, 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.



(a) Basque results

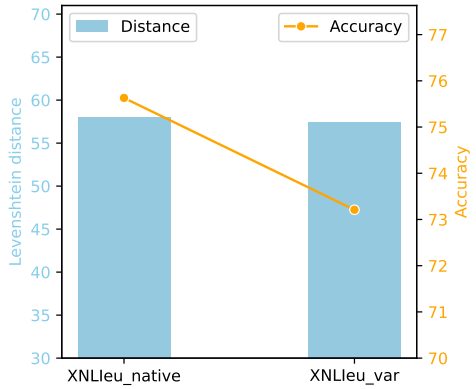


(b) Spanish results

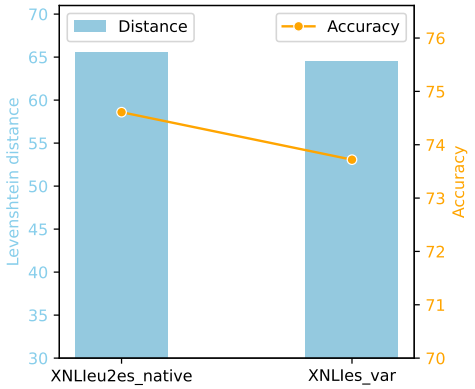
Figure 1: Standard to dialectal Levenshtein distance vs accuracy of best discriminative models.

Accuracy per dialect We analyzed the accuracy results for each individual dialect level, 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 language. This could be because of its under-representation in our dataset, as Navarrese examples comprise only 7% of our data (Appendix B). When focusing



(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

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 are harder to process.

Analysis of Spanish dialectal variations revealed no significant correlation between accuracy and edit distance. This observation may be attributed to two primary factors. First, differences in the manifestation of linguistic variations, with certain languages demonstrating greater inherent variability than others. Second, the extensive training data available for Spanish, a high-resource language, may have exposed the models to a broader spectrum of Spanish varieties, thereby affecting their performance characteristics. Detailed accuracy metrics for individual dialect classifications are presented in Appendix E.

Premise and hypothesis lexical overlap To investigate the potential correlation between lexical overlap and accuracy results, we calculated the Levenshtein distance between premises and hypotheses. The analysis of the data in Figure 2 indicates that the degree of lexical overlap between premise and hypothesis remains relatively constant across standard and dialectal language varieties. However, 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 a 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 methodologies.

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 Natural Language Inference (NLI). The investigation involved the empirical evaluation of both discriminative and generative language models across various NLI task configurations. Results demonstrate a significant inverse correlation between model performance and the inclusion of linguistic variation. 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. Finally, the lexical overlap between premises and hypotheses appears to have minimal impact, suggesting that lower performance is due to linguistic variation.

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 kind of speakers, but we are aware that all the speakers have linguistic and NLP background and laypeople could contribute differently.

We have empirically shown that in our variation dataset the accuracy drops in NLI, but to generalyse this we should include more data from other variants and test it in other NLP tasks.

In the future, we plan to augment the dataset with more geographical variants of Basque and Spanish and add other languages. We are looking for more speakers with different backgrounds that could enrich the dataset with their variations. We also plan to test the performance of NLP tools and LLMs with dialectal and register information in other tasks. Research on variation-inclusive monolingual models could be an interesting future direction.

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.

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, Austen Gregerson, Ava Spataru, Bap tiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Laurens Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen ley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Babu Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Ro main Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Ben Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Her-

moso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, 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).

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. *arXiv*, 2502.09387.

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.

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, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clé ment Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Kather ine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeffrey Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

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.

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.

Robert Lawrence Trask and Peter Stockwell. 2007. *Language and linguistics: The key concepts*. Routledge.

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.