A Catalog of Basque Dialectal Resources: Online Collections and Standard-to-Dialectal Adaptations

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

Abstract

Recent research on dialectal NLP has identified data scarcity as a primary limitation. To address this limitation, this paper presents a catalog of contemporary Basque dialectal data and resources, offering a systematic and comprehensive compilation of the dialectal data currently available in Basque. Two types of data sources have been distinguished: online data originally written in some dialect, and standard-to-dialect adapted data. The former includes all dialectal data that can be found online, such as news and radio sites, informal tweets, as well as online resources such as dictionaries, atlases, grammar rules, or videos. The latter consists of data that has been adapted from the standard variety to dialectal varieties, either manually or automatically. Regarding the manual adaptation, the test split of the XNLI Natural Language Inference dataset was manually adapted into three Basque dialects: Western, Central, and Navarrese-Lapurdian, yielding a high-quality parallel gold standard evaluation dataset. With respect to the automatic dialectal adaptation, the automatically adapted physical commonsense dataset (BasPhyCo$_{west}$) underwent additional manual evaluation by native speakers to assess its quality and determine whether it could serve as a viable substitute for full manual adaptation (i.e., silver data creation).

Keywords: Basque, dialects, low-resource, data-collection

1. Introduction

Dialectal variation is a core feature of all natural languages. However, up until now, Natural Language Processing (NLP) research has almost exclusively focused on tailoring data and resources for the standard forms of each language.

In recent years, this trend has slowly started to shift, with some studies increasingly focusing on dialects. However, the range of tasks and languages addressed remains fairly limited. Several surveys review recent developments and research directions in dialectal NLP, such as Joshi et al. (2025) and Zampieri et al. (2020).

Therefore, a major limitation identified by these works is the scarcity of data and resources for non-standard varieties. This presents a significant challenge, since recent advances in state-of-the-art NLP have reinforced the importance of data quantity in developing high-performing language technology tools, especially in low-resource scenarios (Artetxe et al., 2022). Given the importance of data quantity, the field of dialectal NLP could be regarded as a low-resource research scenario.

The lack of modern dialectal data is especially pronounced for Basque NLP technologies. The majority of resources have been developed with a high focus on Standard Basque, such as spell-checkers (Agirre et al., 1992), Neural Machine Translators[1] or, more recently, text representations (Agerri et al., 2020), and instruction fine-tuned Large Language Models (LLMs) trained for Basque (Etxaniz et al., 2024; Sainz et al., 2025).

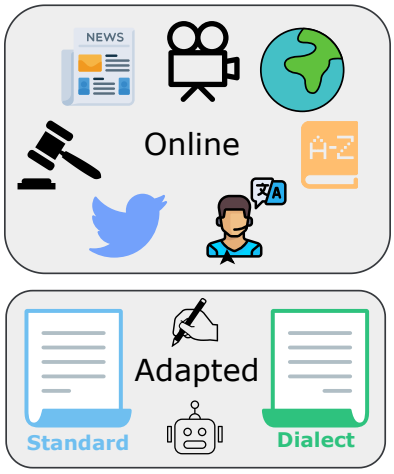

Figure 1: The Basque dialectal catalog consists of two different sources: **online** dialectal data and standard-to-dialect **adapted** data, either manually or automatically.

In this context, this paper aims to present a thorough compilation of Basque dialectal data and resources in order to facilitate potential future work on Basque dialects. We hope this will be useful not only in the field of NLP, but also in other related areas such as sociolinguistics or variationist linguistics.

This work groups Basque dialectal data and resources into two main categories, presented in Section 3 and Section 4, respectively:

[1] Elia or Itzuli.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 153–164
16 May 2026.

- **Online dialectal data and resources.** This category includes online sites such as local news and radio stations, tweets, linguistic atlases, dictionaries, and videos.
- **Standard-to-dialect adapted data.** This includes standard data adapted into dialects, either manually or automatically. For the former, we present **Parallel XNLIvar**, an expansion of XNLIvar (Bengoetxea et al., 2025) that includes the manual adaptation of the 5000-instance test set for the task of Natural Language Inference (NLI). The latter describes **BasPhyCo**, a physical commonsense dataset automatically adapted into dialectal Basque using Large Language Models (LLMS) (Bengoetxea et al., 2026). A novel manual evaluation has been conducted by native dialect speakers on the automatically obtained data.

These resources and adapted datasets are collected and publicly available[2].

2. Related Work

Dialectal Data Collection Many recent studies have focused their research on the systematic collection of dialectal data as a strategy to address the problem of data scarcity. For instance, Sun et al. (2025) introduced a gamified dialect data collection system, where native speakers could access an interface and either rewrite sentences into their dialect or match dialectal sentences to their geographical location. Their work has been found to efficiently increase user engagement in data collection processes.

However, given the growing demand to acquire large volumes of data, manual collection is frequently impractical. Therefore, several works have resorted to silver data creation, such as MultiVALUE (Ziems et al., 2023), a rule-based translation system for 50 different English dialects and 189 linguistic features, which performs as a successful data augmentation method.

Although the critical role of dialectal data is widely recognized, no study currently provides a comprehensive dialectal database. There are, however, some benchmark works that do not explicitly focus on data, but provide an invaluable source of dialectal data. For instance, Faisal et al. (2024) presented an extensive dialectal benchmark, evaluating 10 different tasks in 281 varieties. Additionally, Alam et al. (2024) focused their dialectal benchmark on Machine Translation (MT), as they evaluated several varieties from 12 different languages.

Both of these benchmarks included Basyque (Uria and Etxepare, 2012), a dataset of Northern Basque dialects used to evaluate MT. No other task was evaluated in Basque in these works, as there was no other dialectal dataset for modern Basque dialects at the time of these benchmarks.

Basque Dialects The classification of Basque dialects has been up for debate for many years. Bonaparte (1869) proposed a classification of eight dialects and 25 subdialects, a distribution that highlights the remarkable degree of variation within Basque, despite its comparatively small geographic area.

This classification has been considered canonical until a recent Basque dialectology work by Zuazu (2008), which established an extensive and comprehensive categorization of five Basque dialects, and provided a broad archive of the most representative features of each variation.

This paper will adopt the dialectal classification proposed by Zuazu (2008), distinguishing the following Basque dialects: Western, Central, Navarrese, Navarrese-Lapurdian and Zuberoan. For the purposes of this work, the Navarrese-Lapurdian and Zuberoan dialects will additionally be referred to as Northern dialects. A map illustrating this classification by Zuazu (2008) is provided in the Appendix.

Basque Standardization While extensive dialectal diversity may constitute a source of linguistic richness, it can also increase the risk of language endangerment. This concern led many Basque linguists to push for the necessity of a standard variation (Garabide, 2010).

The process for the creation of a standard variation was a lengthy one. Many meetings were held by contemporary linguists, until Koldo Mitxelena's proposal in the 60s. He suggested using the Central dialects as the foundation of the Standard, mainly due to its practicality: it was the dialect with the most literary prestige to date, as well as being understandable by all Basque speakers. This proposal was discussed, modified, and accepted in the Congress of Arantzazu in 1968 (Garabide, 2010).

The emergence of the standard variety was followed by growing movements to preserve and revitalize dialects. For instance, the Standard Western variety, which was developed to formalize and thus promote its use in written form as well as other registers and use cases (Labayru and Kutxafundazioa, 2001).

Overall, Basque standardization and dialect preservation have developed side by side, both playing an important role in shaping the language today. This work aims to contribute to this movement by providing a collection of current dialectal resources.

[2] https://github.com/hitz-zentroa/Catalog-of-Basque-Dialects

Source	Type	Dialect	Register	Modality	License
Bizkaia Irratia	Text & Audio	Western	Formal	News & Radio	cc-by-sa
Bizkaie!	Text & Audio	Western	Formal	News & Radio	CC-BY-SA
Xiberoko Botza	Text & Audio	Zuberoan	Formal	News & Radio	CC-BY-SA
Irulegiko Irratia	Text & Audio	Nav-Lap	Formal	News & Radio	CC-BY-SA
Gure Irratia	Text & Audio	Nav-Lap	Formal	News & Radio	CC-BY-SA
General Assemblies	Text	Western	Specialized	Minutes	N/A
BSM	Text	Mixed	Informal	Tweets	CC-BY-SA
Linguistic Atlas	Resource	Mixed	-	Atlas	N/A
Euskaltzaindia	Resource	Mixed	-	Dict.	N/A
LabayruHiztegia	Resource	Western	-	Dict.	N/A
LabayruGramatika	Resource	Western	-	Grammar	N/A
Basyque	Resource	Northern	-	Grammar	CC-BY-SA
Ahotsak	Video	Mixed	-	Speech	CC-BY-SA
Mintzoak	Video	Northern	-	Speech	CC-BY-NC-SA
Euskalkiak.eus	Video	Mixed	-	Speech	CC-BY-SA

Table 1: Summary of online Basque dialectal resources. N/A = Not Available.

3. Online Dialectal Data

In this section, we introduce the dialectal data available within the Basque online community, categorized according to their modality. All sources and their characteristics are summarized in Table 1.

3.1. News and Radio Sites

Several Basque news sites write articles in their local dialectal variety. For instance, news sites that have been written in the Western dialect include **Bizkaia Irratia** and **Bizkaie!**. Inside the news domain, these texts are written in a formal, not technical register.

These sites provide up-to-date news articles in the Standard Western dialect. This variety was first formalized in Labayru and Kutxafundazioa (2001), where orthography, morphology (including verb declination), and syntax issues were established. More recently, the Labayru foundation has digitalized this grammar and provided Standard Western Basque information on their website (more information on Section 3.6).

Apart from the Western dialect, several online news sites are also available in other varieties, including the Northern dialects. In fact, the **Euskal Irratiak** association consists of several independent radios that collect news from different locations, in their local variation of Basque. These include **Xiberoko Botza**, **Irulegiko Irratia** and **Gure Irratia**.

3.2. Legal Documents

The website of the **Biscayan General Assemblies** contains the minutes of the highest organizational body representing the citizens of Biscay, i.e., the Biscayan parliament. This organization exercises regulatory power in the region and approves the budgets for the historical territory.

Some of the legal documents on their website are written in the Western dialect, providing dialectal texts not only in a formal register, but also in the specialized legal domain. As is the case for the news site texts, these technical documents are written in the Standard Western dialect.

3.3. Basque Social Media Corpus

Some previous work has been done on the comparison of different registers in Basque, where informal speech often includes strong dialectal variations. These works include Fernandez de Landa et al. (2019) and Fernandez de Landa and Agerri (2021), where they work on real-world data collected from Twitter, which includes dialectal, slang, informal, and code-switched data. The **Basque Social Media (BSM)** corpus consists of approximately 11 million posts produced by more than 13,000 Basque-speaking users, with a total of around 188 million words.

3.4. Linguistic Atlas

Euskaltzaindia is the academic regulatory institution for the Basque language. In their many efforts to research and support the language, they have conducted some invaluable work on dialects, including the **Linguistic Atlas of Basque**.

This atlas provides linguistic maps for many Basque words according to their geographical location, thus illustrating the big lexical variation that could be encountered through the several Basque-

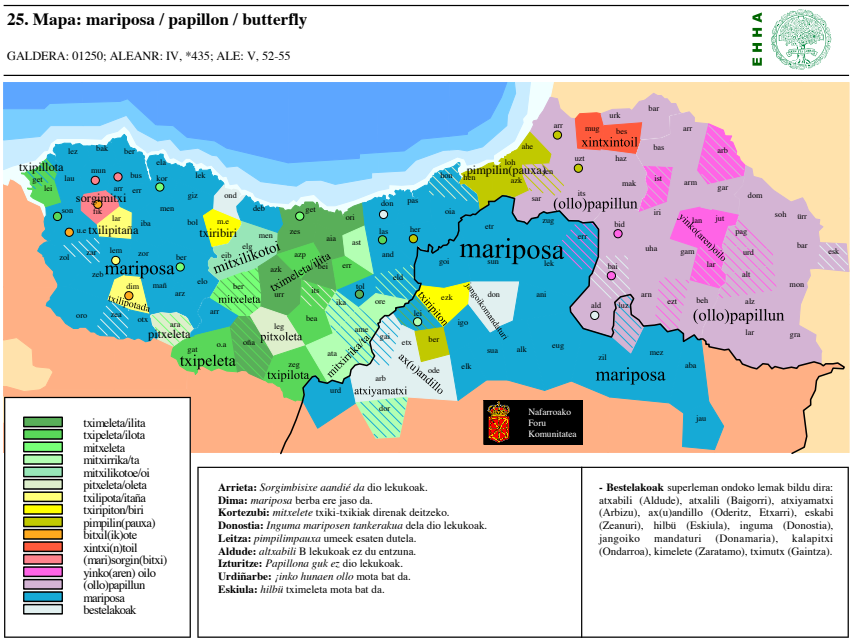

Figure 2: Example of a linguistic atlas.

speaking regions. An example map is provided in Figure 2.

3.5. Dictionaries

Some online dictionaries often include dialectal information when searching for specific words, such as Elhuyar or **Euskaltzaindia**. The Elhuyar dictionary provides dialectal information about some words. The Euskaltzaindia dictionary provides not only geographical information for selected words, indicating their potential dialectal affiliation, but also enables searches to be filtered by dialect.

However, these standard dictionaries only provide information on dialectal words that are accepted in the standard, while other variations of dialectal words are absent. For instance, *berba* (word) is accepted in the standard form and appears in dictionaries, while its possible variations, such as *berbie* or *berbi*, do not.

Additionally, some dictionaries are designed to focus on a particular dialect, for example **LabayruHiztegia**. This is a Western-focused dictionary, with Standard-Western word pairs, as well as information on some multi-word expressions particular to this dialect.

3.6. Grammar

Some efforts have been made to formalize the grammar rules from several Basque dialects. For example, the Labayru foundation has recently launched **LabayruGramatika**, a compilation of the grammatical rules from the Western dialect, with explanations as well as examples for every grammatical phenomenon.

Furthermore, Uria and Etxepare (2012) provided **Basyque**, an online resource to store, organize, manage and search for all the information concerning dialectal variation in, specifically, the North-Eastern Basque dialects, providing information that enables the syntactic analysis of these dialects.

3.7. Dialectal Videos

In an attempt to observe, preserve and analyze diachronic variation, some speakers have been interviewed, and the recordings have been uploaded to several web pages. These videos are annotated with the speakers' geographical origin.

For instance, **Ahotsak** provides thousands of videos of interviews with people from different generations, genders, backgrounds, and geographical locations. Some of these interviews have been transcribed and could act as a great resource for the analysis of oral Basque dialects. However, the videos with missing transcriptions still pose a great challenge.

Similar to Ahotsak, **Mintzoak** compiles video interviews of Northern Basque speakers, in order to keep the collective memory alive. Contrary to Ahotsak, this site does not contain transcriptions of the interviews.

Additionally, **Euskalkiak.eus** also contains some videos based on the geographical location of the speakers, also with no transcriptions.

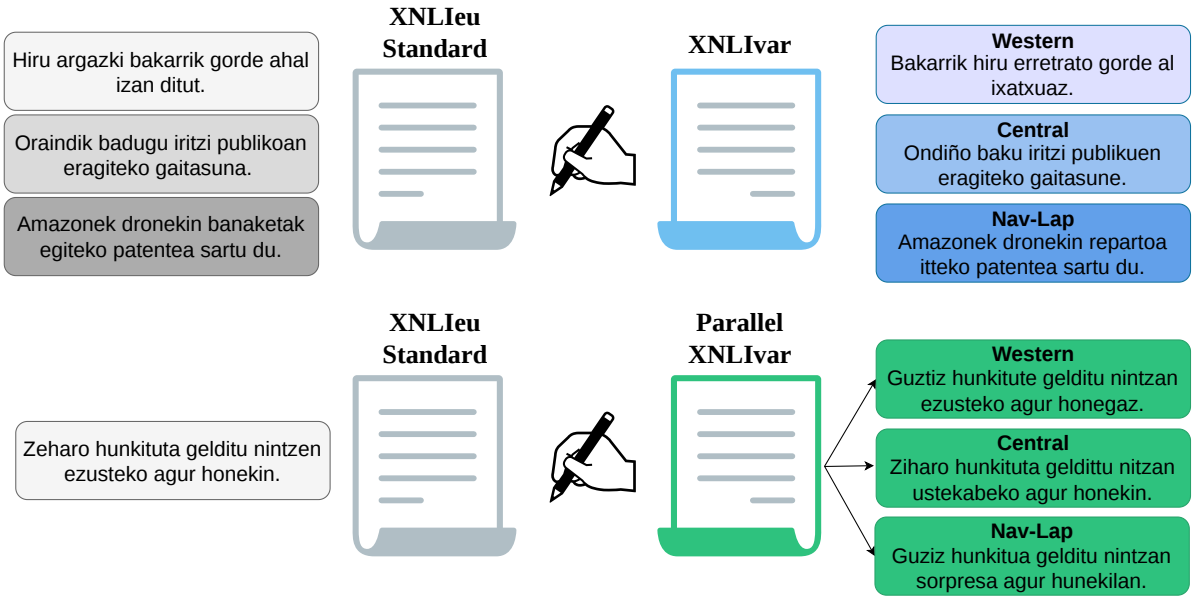

Figure 3: Illustration of XNLIeu dialectal adaptation. XNLIvar was a compilation of different instances in three different dialects. Parallel XNLIvar provides the same instances in three Basque dialects, offering completely parallel data.

4. Standard-to-Dialect Adapted Data

In addition to these online dialectal resources, both manual and automatic adaptations from standard Basque into dialects have recently been developed. Standard-dialect parallel data constitute an important resource not only for NLP but also for studies in variationist linguistics and sociolinguistics.

4.1. Manual Dialectal Adaptation

Manually adapting standard texts into dialects is to this day the most effective way to obtain gold standard data. However, despite its importance, relatively few studies have addressed this topic in Basque. This section describes the first manually adapted dataset into Basque dialects (Bengoetxea et al., 2025). Additionally, this paper presents a newly developed expansion of it, namely Parallel XNLIvar, also collected through manual adaptation.

XNLIvar Bengoetxea et al. (2025) presented XNLIvar, a Natural Language Inference (NLI) dataset which contained data in three Basque dialects. This dataset was manually adapted from XNLIeu (Heredia et al., 2024), an NLI dataset in Basque translated from the multilingual XNLI dataset (Conneau et al., 2018). This dataset consists of Premise-Hypothesis pairs with entailment, neutral or contradiction relations.

For the dialectal adaptation, Bengoetxea et al. (2025) adapted the native partition of XNLIeu, i.e., a 621-instance test set manually created in Basque. Although the dataset contained material from three Basque dialects (Western, Central, and Navarrese), it did not include fully parallel versions of the same content across dialects.

Dialect	Birthplace	Age	Gender	Studies
Western	Elorrio	58	Male	Translation
Central	Arroa/Zumaia	34	Female	Translation
Nav-Lap	Donibane-Lohizune	33	Female	Translation

Table 2: Metadata from native Basque speakers who adapted the test partition of XNLIeu into dialects.

Parallel XNLIvar This paper presents an expansion of XNLIvar by adapting the original XNLI test set of around 5000 instances into three different Basque dialects (Western, Central, and Navarrese-Lapurdian). This expansion was created in parallel, providing three fully dialectal versions of the XNLIeu test set, each corresponding to a different Basque dialect. This adaptation process is illustrated in Figure 3.

The adaptation from standard to dialectal Basque has been manually carried out by native speakers of each dialect. Metadata from each native speaker is illustrated in Table 2, such as birthplace, age, gender and previous studies.

All in all, this novel resource allows for a more thorough assessment of dialectal effects in NLI and facilitates per-dialect analysis through the inclusion of parallel data.

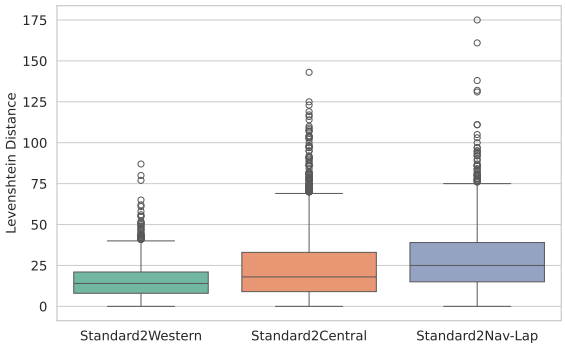

Figure 4: Levenshtein distance distribution for the three Parallel XNLIvar datasets.

4.1.1. Analysis of Parallel XNLIvar

Further analysis has been conducted to explore how different the dialectal datasets are from the standard. To do so, the Levenshtein distance between standard and dialectal sentences was calculated. This metric measures the minimum number of single-character insertions, deletions, or substitutions required to transform one string into another. The average distance results are presented in Table 3.

The results suggest that Navarrese-Lapurdian is the dialect most distant from the standard variety. This is consistent with Basque dialectological research, which has shown that peripheral dialects typically exhibit greater divergence (Mitxelena, 1981).

Under this theoretical framework, the Western dialect should not be the closest to the standard variety, given that it is likewise considered a peripheral dialect. However, as we can see in Figure 4, the Central dialect seems to be further from the standard than the Western. This could be due to the distance distribution, as the Central dialect shows many sentences with big modifications, whereas the distances in the Western dataset remain consistently low across all sentences.

Further analysis of the two datasets shows that the linguist who did the Central adaptations frequently modified sentence word order. That could be why the Levenshtein distance for the Central dataset is higher than for the Western dataset, in which word order was largely preserved. This highlights the importance of robust metrics for measuring dialectal variation, as this distance metric estimates the transformation distance between sentences, but its effectiveness in capturing dialectalness seems to remain unclear, especially in free word-order languages like Basque.

4.2. Automatic Dialectal Adaptation

Given the high cost and time demands of manual dialectal adaptation, recent work has focused on automatically converting standard data into dialects, which is discussed in this section.

BasPhyCo Bengoetxea et al. (2026) presented a physical commonsense reasoning dataset consisting of 356 instances of 5-sentence stories, which could be plausible or implausible. BasPhyCo was manually translated into standard Basque from its original Italian version (Pensa et al., 2024). Additionally, the standard Basque dataset was automatically adapted into the Western dialect through few-shot prompting of Latxa-It-70B (Sainz et al., 2025).

Consequently, Bengoetxea et al. (2026) provide two parallel versions of the same dataset: BasPhyCo and BasPhyCo$_{\text{west}}$. The availability of standard–dialect parallel data allows to examine the impact of dialectal variation on physical commonsense reasoning.

4.2.1. Evaluation of BasPhyCo$_{\text{west}}$

The automatic adaptation of BasPhyCo$_{\text{west}}$ was validated by a professional Basque linguist (Bengoetxea et al., 2026). However, we sought to extend this evaluation to the adapted dataset, thereby measuring the actual dialectal value of the automatic adaptation for native Western dialect speakers. To do so, we have outlined the following manual evaluation framework.

Evaluated datasets The original prompt of Bengoetxea et al. (2026) for the adaptation of BasPhyCo$_{\text{west}}$ explicitly allowed for non-standard orthographic modifications. We have adapted this prompt by eliminating the possibility of orthographic changes and obtained a new version, BasPhyCo$_{\text{west-new}}$. Both prompts used for the adaptations are provided in the Appendix.

Therefore, our evaluation was done on three versions of the dataset: the **Standard** (BasPhyCo), the Western **Orthographic** (BasPhyCo$_{\text{west}}$), and a novel **Western**, non-standard orthographic version (BasPhyCo$_{\text{west-new}}$). The following examples illustrate two dialectal sentences, both with standard and non-standard orthography.

Dialect	Distance
Western	15.56
Central	24.73
Nav-Lap	29.36

Table 3: The average Levenshtein **Distance** of the Parallel XNLIvar dialectal datasets to the Standard.

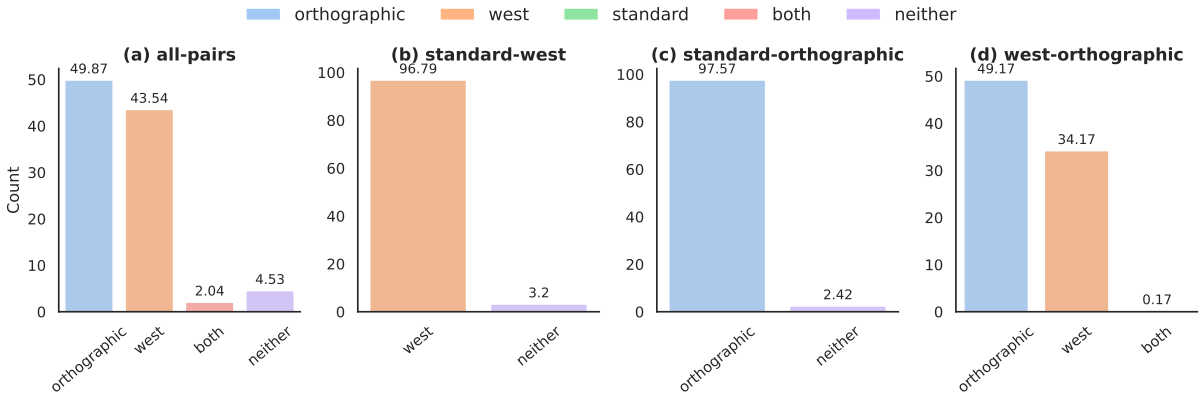

Figure 5: Manual evaluation results. From left to right, results for all sentence pairs (**all-pairs**), as well as results for different sentence pair combinations (**standard-west**, **standard-orthographic**, and **west-orthographic**).

Standard Teknikaria ez da oraindino etorri

Non-Standard Teknikarixa ez dau oraindiño etorri

We can observe that the non-standard version includes the letter *ñ* in *oraindiño*, while the standard version follows standard orthographic rules and favors the use of the letter *n*.

Description of the task The evaluation was proposed as a pairwise comparison task. The pairs of sentences to be evaluated were constructed as a Cartesian product, i.e., every sentence from dataset A was paired with every sentence in dataset B. Thus, the combination of datasets A $\times$ B is the following:

$$A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}. \quad (1)$$

As we have three different datasets (A, B and C), the Cartesian product for all three dataset combinations was conducted and subsequently concatenated. Given that $\times$ is the Cartesian product, the pairs to be evaluated were constructed as:

$$(A \times B) \cup (A \times C) \cup (B \times C) \quad (2)$$

For the manual evaluation, 180 sentence pairs from each dataset combination were randomly sampled, obtaining a total of 540 sentence pairs for evaluation.

The task consisted of annotating which sentence from each pair was closer to the Western dialect of Basque. The evaluators were given four options: *sentence A* is more Western, *sentence B* is more Western, they are *both* Western, or *neither* is Western.

Dialect	Birthplace	Age	Gender	Studies
Western	Oñati	33	Female	Translation
Western	Elorrio	59	Male	Translation

Table 4: Metadata from the native Western Basque speakers for the manual evaluation of the automatic adaptation.

Annotators The evaluation was made by two different native Western speakers, with the same educational background, but different age, gender, and birthplace. The metadata for each evaluator is presented in Table 4.

The Inter-Annotator Agreement (IAA) between the two evaluators has been computed, with a Cohen's Kappa of 0.71, which constitutes a substantial agreement score.

Quantitative results The general results, as well as the results for different sentence-pair types, are illustrated in Figure 5.

The results for all sentence types (5a) show that the orthographic dataset has the most Western features, although it is closely followed by the Western dataset. No Standard sentence was marked as dialectal, which consolidates the confidence of the evaluation quality. Evaluators marked some doubtful sentence pairs as *Neither* or *Both*, which are examined in the following analysis section.

Regarding the results for the different sentence-pair types, we can observe that when the dialectal sentences were paired with a standard sentence (i.e., standard-west and standard-orthographic sentence pairs), the evaluators always chose the adapted sentences over the standard (Figure 5b and 5c). Additionally, when having to choose between the two adapted sentence types (west-orthographic), the evaluators deemed the Orthographic dataset more dialectal, but still closely fol-

lowed by Western sentences (5d).

Consequently, this manual evaluation has shown that both automatically adapted datasets (Orthographic and Western) contain considerable dialectal features compared to the standard version. Additionally, explicitly stating in the adaptation prompt that non-standard orthographic changes are possible seems to generate even more dialectal adaptations, according to this evaluation.

Analysis During the pairwise comparison, evaluators were also given two extra options: *Both* and *Neither*. It can be observed in Figure 5b and 5c that all *Neither* labeled pairs seem to occur when one of the sentences in the pair is from the standard dataset. This reveals that the automatic adaptation sometimes failed to transform the standard sentences into their dialectal form. For example, the following sentences from the adapted datasets do not contain dialectal features:

Ortho Koldo esnatu da. (*Koldo has woken up*)

West Izotz-ontziak urtu dira. (*The ice-cubes have melted*)

Similarly, all *Both* instances seem to occur when comparing the two automatically adapted datasets. This highlights that the Standard dataset is not biased towards dialectal language, as evaluators have not once considered a standard sentence to be dialectal.

Furthermore, disagreements between annotators have also been examined. The majority of these instances occur on difficult sentence pairs, i.e., sentences that contain little to no dialectal features. Further analysis of these instances has revealed a slight annotator bias: one annotator considered *heldu* (arrive) a Western marker, whereas the other did not.

Ortho Ane berandu **helduko** da etxera. (*Ane arrived late home*)

Standard Ane berandu **iritsiko** da etxera. (*Ane arrived late home*)

Finally, although error identification was not an annotation requirement, evaluators noted several errors in the dialectal adaptations. This points to clear limitations in the current dialectal adaptation capabilities of LLMs, which require further investigation.

5. Conclusion

This work presents a comprehensive collection of Basque dialectal data, categorized into two groups. First, online dialectal data and resources have been presented, grouped according to their domain, such as news, legal documents, informal tweets, dictionaries, grammar collections or even audiovisual resources. Secondly, standard datasets that were adapted into dialects have been described, both manually and automatically.

This paper has introduced two main contributions: (i) Parallel XNLIvar, a novel manually adapted parallel dialectal dataset for NLI. (ii) A manual evaluation of the automatically adapted $\text{BasPhyCo}_{\text{west}}$ dataset, outlined as a pairwise comparison task and performed by native dialectal speakers.

Limitations

The main limitation regarding online sources is that some data cannot be used due to licensing issues.

Additionally, automatic dialectal adaptation has been evaluated for only one Basque dialect, while other dialects remain unexamined.

Finally, the manual evaluation of the automatic dialectal adaptation has revealed some errors, which present an evident limitation in the possibility of silver data creation through automatic adaptation.

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 project: (i) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR and (ii) DeepThought (PID2024-159202OB-C21) funded by ERDF, EU.

Bibliographical References

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.

Eneko Agirre, Iñaki Alegria, Xabier Arregi, Xabier Artola, Arantza Diaz de Ilarraza, Montserrat Maritxalar, Kepa Sarasola, and Miriam Urkia. 1992. XUXEN: A spelling checker/corrector for Basque based on two-level morphology. In *Third Conference on Applied Natural Language Processing*, pages 119–125, Trento, Italy. Association for Computational Linguistics.

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.

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.

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

Jaione Bengoetxea, Itziar Gonzalez-Dios, and Rodrigo Agerri. 2026. Physical commonsense reasoning for lower-resourced languages and dialects: a study on basque.

Louis-Lucien Bonaparte. 1869. Le verbe basque en tableaux. *Berrargitaratua:[JA Arana Martija, arg.], Opera omnia vasconice*.

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.

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

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

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

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

Elkartea Garabide. 2010. Hizkuntzaren estandarizazioa. *Euskararen berreskuratzea ii*.

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.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.*, 57(6).

Ikastegia Labayru and Bilbao Bizkaia Kutxafundazioa. 2001. Bizkai euskeraren jarraibide liburua. *Lehenengo Pausuak. Labayru Ikastegia-BBK-Eusko Jaurlaritza-Bizkaiko Foru Aldundia-Bizkaia Irratia*.

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

Giulia Pensa, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. A Multi-layered Approach to Physical Commonsense Understanding: Creation and Evaluation of an Italian Dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 819–831.

Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. Instructing large language models for low-resource languages: A systematic study for Basque. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29136–29160, Suzhou, China. Association for Computational Linguistics.

Jiugeng Sun, Rita Sevastjanova, Sina Ahmadi, Rico Sennrich, and Mennatallah El-Assady. 2025. Dia-lingle: A gamified interface for dialectal data

collection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 148–158, Vienna, Austria. Association for Computational Linguistics.

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.

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.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

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

A. A Map of Basque Dialects

The map in Figure 6 shows the Basque dialect classification according to Zuazu (2008).

B. Dialectal Adaptation Prompts

The prompts used to prompt Latxa-3.1-70B-It for the dialectal adaptation are provided below, both for the non-standard orthographic changes (Figure 7) and for the standard orthographic changes (Figure 8).

Figure 6: The map of the classification of Basque dialects according to Zuazu (2008).

I will give you three versions of a story. Each version has five sentences. Some sentences are identical across versions. You need to adapt this text so that it includes Bizkaian dialectal features. **You can use non-standard orthography. Try to make it as similar as possible to oral language.**

Task:

1. First, list all unique sentences across all three stories.
2. Adapt each unique sentence exactly once into the Bizkaian dialect.
3. Then reconstruct the three stories with the translations, making sure that any identical source sentence always has the identical translation.
4. If there are more than three stories, repeat the same process for all of them.

Format:

This is an example of an standard (INPUT) instance and an example of the dialectal (OUTPUT) adaptation that you need to do:

Standard:

STORY1: ['Mikel lanera joan da', 'Mikelek ordenagailua piztu du', 'Mikelek mezuak irakurri ditu', 'Mikelek mezuak erantzun ditu', 'Mikel etxera joan da']
STORY2: ['Mikel lanera joan da', 'Mikelek mezuak erantzun ditu', 'Mikelek mezuak irakurri ditu', 'Mikelek ordenagailua piztu du', 'Mikel etxera joan da']
STORY3: ['Mikel lanera joan da', 'Mikelek ordenagailua itzali du', 'Mikelek mezuak irakurri ditu', 'Mikelek mezuak erantzun ditu', 'Mikel etxera joan da']

Dialectal:

STORY1: ['Mikel lanera jun de', 'Mikelek ordenagaillua piztu dau', 'Mikelek mesuek irakurri dauz', 'Mikelek mesuek erantzun dauz', 'Mikel etxera jun de']
STORY2: ['Mikel lanera jun de', 'Mikelek mesuek erantzun ditu', 'Mikelek mesuek irakurri dauz', 'Mikelek ordenagaillua piztu dau', 'Mikel etxera jun de']
STORY3: ['Mikel lanera jun de', 'Mikelek ordenagaillua amatatu dau', 'Mikelek mesuek irakurri dauz', 'Mikelek mesuek erantzun dauz', 'Mikel etxera jun de']

Output only the reconstructed stories in the exact same format as the input. Do not output explanations, steps, or commentary.

Figure 7: Dialectal adaptation prompt for the non-orthographic dataset version ($\text{BasPhyCo}_{\text{west}}$).

I will give you three versions of a story. Each version has five sentences. Some sentences are identical across versions. **You need to adapt this text so that it includes Bizkaian dialectal features.**

Task:

1. First, list all unique sentences across all three stories.
2. Adapt each unique sentence exactly once into the Bizkaian dialect.
3. Then reconstruct the three stories with the translations, making sure that any identical source sentence always has the identical translation.
4. If there are more than three stories, repeat the same process for all of them.

Format:

This is an example of an standard (INPUT) instance and an example of the dialectal (OUTPUT) adaptation that you need to do:

Standard:

STORY1: ['Mikel lanera joan da', 'Mikelek ordenagailua piztu du', 'Mikelek mezuak irakurri ditu', 'Mikelek mezuak erantzun ditu', 'Mikel etxera joan da']
STORY2: ['Mikel lanera joan da', 'Mikelek mezuak erantzun ditu', 'Mikelek mezuak irakurri ditu', 'Mikelek ordenagailua piztu du', 'Mikel etxera joan da']
STORY3: ['Mikel lanera joan da', 'Mikelek ordenagailua itzali du', 'Mikelek mezuak irakurri ditu', 'Mikelek mezuak erantzun ditu', 'Mikel etxera joan da']

Dialectal:

STORY1: ['Mikel lanera jun de', 'Mikelek ordenagaillua piztu dau', 'Mikelek mesuek irakurri dauz', 'Mikelek mesuek erantzun dauz', 'Mikel etxera jun de']
STORY2: ['Mikel lanera jun de', 'Mikelek mesuek erantzun ditu', 'Mikelek mesuek irakurri dauz', 'Mikelek ordenagaillua piztu dau', 'Mikel etxera jun de']
STORY3: ['Mikel lanera jun de', 'Mikelek ordenagaillua amatatu dau', 'Mikelek mesuek irakurri dauz', 'Mikelek mesuek erantzun dauz', 'Mikel etxera jun de']

Output only the reconstructed stories in the exact same format as the input. Do not output explanations, steps, or commentary.

Figure 8: Dialectal adaptation prompt for the orthographic dataset version (BasPhyCo$_{\text{west-new}}$).