Language Variability in Basque: Data Creation, Evaluation and Systems for Dialect and Register Awareness in Natural Language Processing

Jaione Bengoetxea

HiTZ Basque Center for Language Technology - Ixa, University of the Basque Country UPV/EHU

Abstract

Works regarding language variation in Natural Language Processing (NLP) are scarce and often focus on languages for which data is easily available, such as English. Thus, no works have dealt with the automatic processing of linguistic variability in Basque, and the few works that exist on variability in NLP have focused on the linguistic theory or feature theoretical characterization. In this context, the objective of this thesis will be to explore the effect of language variability via automatic techniques using Large Language Models (LLM) in Basque. In order to do so, we will generate the first datasets with manually annotated linguistic variability in Basque, data that will serve as a base for the development of Basque variability-aware NLP systems. Prior to this, a careful linguistic analysis of Basque language variability will first be carried out. Additionally, to overcome the limitations of data scarcity, data collection and augmentation methods will be worked on, thus reducing the time, effort and cost that collecting and generating data involves. Finally, specific NLP tasks such as Question-Answering (QA) and Natural Language Inference (NLI) will be evaluated in terms of variability awareness, and the improvement of those tasks when variation is present will be investigated with the aim of developing robust variability-aware NLP systems in Basque.

Keywords

Variation, Basque, evaluation, low resource

1. Reason for the Proposed Research

Linguistic variability is an inherent characteristic of natural language use. A given sentence may be expressed in multiple forms while preserving approximately the same semantic content. In fact, sociolinguistics identifies three types of language variability [1]: (i) diatopic variation, or geographical variation such as dialects (ii) diastratic variation or speech of different societal groups, like the language of the youth and (iii) diaphasic variation or speech changes depending on the communicative environment, such as registers. This thesis will focus on the geographical variation, as well as register changes. Examples for both variability types are presented in Table 1.

Table 1
Examples of different linguistic variation

Register		Geographical	
Formal	Informal	Standard	Dialectal
To socialize with other teenagers is of utmost importance during the adolescent years.	Hanging out with friends their age is super important for teenagers.	She often spends her own money.	She be spendin' her own money.

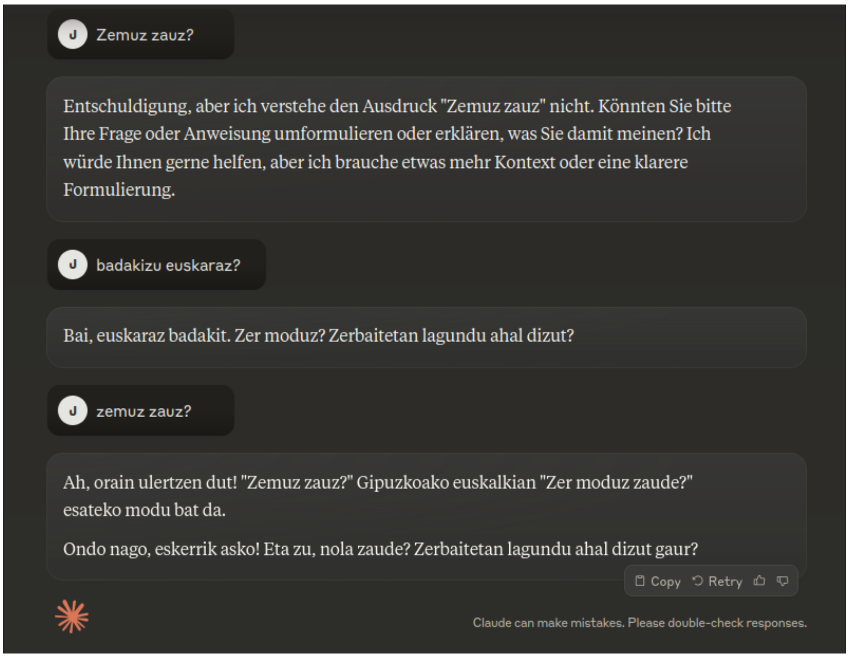
This language variability creates difficulties in several Natural Language Processing (NLP) tasks, which for example involve Question-Answering (QA), Natural Language Inference (NLI) or dialogue system tasks. For instance, Figure 1 illustrates the difficulties of Claude to understand a simple question

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.

jaione.bengoetxea@ehu.eus (J. Bengoetxea)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-11-17

Figure 1: Claude responds in German to a question written in dialectal Basque. When asked "How are you?" in a central dialect, Claude responds in German, saying that it cannot understand the question. After giving some context by asking if they speak Basque, they are able to correctly identify and respond (caption from July 2024).



Zemuz zauz?

Entschuldigung, aber ich verstehe den Ausdruck "Zemuz zauz" nicht. Könnten Sie bitte Ihre Frage oder Anweisung umformulieren oder erklären, was Sie damit meinen? Ich würde Ihnen gerne helfen, aber ich brauche etwas mehr Kontext oder eine klarere Formulierung.

badakizu euskaraz?

Bai, euskaraz badakit. Zer moduz? Zerbaitetan lagundu ahal dizut?

zemuz zauz?

Ah, orain ulertzen dut! "Zemuz zauz?" Gipuzkoako euskalkian "Zer moduz zaude?" esateko modu bat da.

Ondo nago, eskerrik asko! Eta zu, nola zaude? Zerbaitetan lagundu ahal dizut gaur?

in a Basque dialect out of context. Consequently, interest in the automatic processing of language variability has gained considerable interest in recent years, as demonstrated by the success of workshops such as VarDial, which in 2025 celebrated its 12th edition.

Nevertheless, the majority of current research has focused on a limited list of languages (e.g., Arabic, German, English), thus leaving a grand majority of languages unexplored. We consider expanding this field of research to a more diverse set of languages to be of utmost importance, as providing variability-aware NLP tools is an essential step towards building more equitable NLP systems [2]. This will ultimately provide more accessible resources for every user, despite their social, educational or communication background.

Nowadays, NLP relies on Large Language Models (LLMs), which are usually trained on standard varieties of language and need large amounts of data to obtain an acceptable performance. Consequently, due to the lack of language variability data in the training process of LLMs, they considerably struggle to analyze non-standard texts and their performance in tasks such as NLI and QA drops when language variation is present [2].

Therefore, the aim of this thesis is to contribute to this field of growing interest by exploring language variability in Basque. We consider Basque an interesting and challenging language to work on due to its low-resource nature, as well as its high linguistic variation between its dialects. The thesis will be carried out in the HiTZ research group, under the Language Analysis and Processing (UPV/EHU) doctoral program and directed by Rodrigo Agerri and Itziar Gonzalez-Dios.

2. Background and Related Work

In recent years, there has been an increasing interest in language variability in NLP. This section introduces the works that have been presented in the fields of dialect and register processing.

2.1. Geographic Variation

Regarding dialects, research has been conducted in several tasks such as dialect identification [3], sentiment analysis [4], Machine Translation (MT) [5] or dialogue systems [6]. In fact, Aepli and

Sennrich [7] explored cross-lingual transfer between closely related varieties by adding character-level noise to high-resource data to improve generalization. Moreover, Ramponi and Casula [3] pretrained LLMs for geographic variation of Italian tweets. Finally, Demszky et al. [8] 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 [9] propose a bilingual lexicon induction method for German dialects using LLMs, while Hassan et al. [10] 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 being 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. For more information on dialectal research in NLP, Joshi et al. [2] provides a comprehensive survey of the latest works, and Faisal et al. [11] establishes an extensive variability benchmark for several languages.

2.2. Registers

Regarding language variation in terms of register, two main tasks have been researched: style transfer [12] and register classification [13]. The most relevant for this project is style transfer, which involves converting a sentence from one register to another. Rao and Tetreault [12] studied the transformation from informal to formal English, finding that Neural Machine Translation (NMT) achieved the highest formality, while their rule-based approach best preserved meaning. Briakou et al. [14] introduced XFORMAL, a multilingual dataset with formal sentences derived from informal ones in Brazilian Portuguese, French and Italian, highlighting that there is still potential for improvement in multilingual style transfer. However, the majority of the works have been carried in English.

2.3. Language Variation in Basque

In Basque dialectology, Zuazu [15] established an extensive and comprehensive descriptive representation of features of modern Basque dialects. In NLP, Estarrona et al. [16] worked on a morpho-syntactically annotated corpus of Basque historical texts as an aid in the normalization process. Moreover, Uria and Etxepare [17] introduced a corpus of syntactic variation in northern Basque dialects. Additionally, some dialectal benchmark works have included Basque in their experimentation, where they presented benchmarks for MT with northern Basque dialects [18, 11]. However, no work has yet dealt with southern Basque dialects in NLP. Following Zuazu [15]'s work, southern Basque dialects would be Western (traditionally linked to the province of Biscay), Central (traditionally linked to the province of Gipuzkoa) and Navarrese.

Regarding Basque registers, some linguistic theory research on registers involves studies on academic Basque [19, 20], as well as an informal form of Basque called 'hika' [21]. The closest work to NLP is Alonso-Ramos and Zabala [22], who extracted academic vocabulary lists to create a writing aid tool. Thus, to the best of our knowledge, no previous work has been done involving register processing and understanding in the field of NLP.

3. Description of the Proposed Research, Including the Main Hypotheses for Research

This research will explore the effect of language variation on the performance of LLMs in Basque, as well as examine methods to improve their behavior when linguistic variability is added. Based on a linguistic study, language variability performance will be evaluated in NLP tasks such as Natural Language Inference (NLI) and Question Answering (QA), exploring data collection methods and implementation to improve the robustness of language variability in these tasks.

Our main hypothesis is that LLMs will struggle to perform certain tasks when using linguistically diverse data as input, especially given the inherently high variability of Basque. Therefore, our first objective will be to perform a thorough evaluation of current NLP resources. This is a novel and ambitious line of research for several reasons: (i) currently, there is no available dataset that covers language variability in Basque regarding southern dialects or different registers (ii) although some aspects of language variability can be generalized, there are many other aspects that are language specific, thus data collection methods as well as system building approaches will need to be adapted to Basque, which represents a considerable scientific challenge (iii) there is no NLP system that extensively supports linguistic variability in Basque.

Achieving the main objective of the thesis will have two main benefits. First of all, the results of this thesis will contribute to the understanding of the underlying linguistic mechanisms inherent to dialectal and register variation in Basque. Secondly, the development of variability-aware NLP systems will bring benefits to several other fields such as QA, NLI, discourse and dialogue systems, summarization or MT, consequently bringing the field of Language Technology closer to more accessible and equitable NLP tools. In order to achieve this main objective, the following intermediate tasks have been outlined:

Task 1: Analysis of variation. A linguistic variation analysis, both in terms of dialects and registers, is essential. More specifically, the adaptation of Zuazu [15]'s work on dialectal Basque features to language technology tools is imperative for the automatic processing of geographical variation. Additionally, establishing well-distinguished register boundaries for Basque is necessary for the automatic processing of different formality sentences.

Task 2: Data collection. Conducting a data collection process for low-resource environments specific to the linguistic variation of Basque will be essential, thus obtaining the first linguistic variability dataset in Basque. First, a search and collection of publicly available data will be carried out. This process will be complemented with some experimentation based on paraphrasing and MT approaches, such as rule-based permutations, lexical normalization or style transfer methods. Additionally, manual adaptation of datasets into linguistically diverse text will also be explored to obtain gold label quality data.

Task 3: Data augmentation with generative language models. An investigation of different techniques to take advantage of language models for text generation such as monolingual (Latxa [23]) and multilingual (Bloom [24]; GPT-4 [25]) LLMs will be conducted, thus facilitating the generation of synthetic data in language variability tasks in Basque. This will serve both as data transformation as well as a data augmentation step, which will be significantly relevant for low-resource environments, as current Deep Learning and Neural Network approaches often demand considerable amounts of data.

Task 4: Assessment. The performance of monolingual and multilingual state-of-the-art LLMs will be evaluated in different NLP tasks when language variability is present. With this assessment, the shortcomings and limitations of LLMs will be identified and analyzed.

Task 5: Development and evaluation of variability-aware LLMs. The assessed LLMs will be adapted, thus improving their performance based on the analysis carried out in Task 4 and the data collected and generated in Tasks 2 and 3. The tasks of QA and NLI will be evaluated in terms of linguistic variability as studied in Task 1. Due to data scarcity, the foreseen methods are zero- and few-shot techniques, which rely on none or few training data points for experimentation.

In summary, our main hypothesis is that LLMs encounter difficulties when dealing with linguistic variation across specific tasks, especially in Basque, a high variability language. Consequently, our objective will be to provide novel resources (such as linguistically diverse datasets, either manually created, collected or automatically generated), as well as to evaluate the performance of current NLP tools. Finally, an experimentation step will be carried out to improve the understanding and performance ability of current NLP resources when dealing with linguistically diverse data and tasks.

4. Methodology and the Proposed Experiments

Variability processing in NLP is currently marked by deep learning neural systems, often supported by methods based on linguistic features. However, the scarcity of training data, particularly for low-resource languages like Basque, poses a significant challenge in language variability processing.

Thus, this project will require a thorough dialect- and register-aware data collection process, which will cover a wide range of text types, providing us with a large scope of linguistic variability. In other words, dialects, registers and text types are interconnected: text types are influenced by the appropriate register for each context, while certain dialects are more suitable for specific registers. Thus, obtaining different text types will inherently provide us with the linguistic variation that this thesis aims to study, as well as create robust tools that could deal with different types of texts.

In this context, our methodological proposal will consist of the following novelties: (i) adapting data collection methods for Basque language variability, (ii) establishing Basque-specific evaluation criteria for variability detection (iii) exploring few-shot and zero-shot approaches to reduce the need for costly manual data collection and annotation.

The following experiments have been proposed, which have been organized yearly:

Year 1: linguistic analysis of variation and data collection We will start working on Task 1 by conducting an extensive study of language variation in Basque. This will imply the analysis of dialectal features [15], as well as the creation of a general formality typology for Basque, based on previous domain-specific analyses [22].

Large amounts of real and natural linguistically variable data can be found in local news articles, oral transcriptions, or subtitles, while registers that differ from the neutral form of language could be extracted from academic and scientific texts [22], legal texts or political speech [26, 27].

In terms of data collection, a gold standard dataset will be created by manually adapting the evaluation partition of a NLI dataset (XNLI-eu [28]) into different variations of Basque. This will allow us to perform some baseline experiments to measure the level of linguistic variability in understanding these tasks currently in the NLP field.

Additionally, some (semi-)automatic methods will be explored, such as rule-based settings, lexical normalization, or LLM prompting in order to obtain additional silver parallel data.

The experiments and objectives planned for this year are the following:

1. Theoretical framework for the analysis and processing of variability in Basque, which will establish concise criteria to determine if a sentence has an adequate dialectal and/or formal form.
2. Development of data collection methods for low-resource environments specific for the linguistic variation of Basque, exploring rule-based settings as well as lexical normalization approaches and LLM prompting.
3. Provide the first manually-adapted, publicly available dataset of Basque language variation of southern dialects.
4. Baseline evaluation of the effect of language variability in the task of NLI. Zero-shot experiments will be conducted to analyze the effect of variability data in the fine-tuning step.
5. Publication of data collection approaches for language variability in low-resource environments.

Year 2: generation of language variability data for low-resource environments The feature-typologies and variability data obtained in the previous year will be used to work on a second iteration of Task 2, this time focusing on the generation of text containing variability. For this purpose, Task 3 will focus on the study of different methods of language generation, experimenting with the creation of different types of linguistic variability, and evaluating the generated variation in terms of the theoretical framework previously established.

In this respect, we will work on expanding the adaptation of XNLI-eu to larger evaluation data as well as training data. Additionally, we will work on other tasks such as QA by adapting available datasets (e.g., BertaQA [29]) into variability by using the data adaptation methods previously explored.

The experiments and objectives planned for this year are the following:

1. Development of techniques for the automatic generation of synthetic data with variability through experiments with generative language models.
2. Variability datasets for QA and NLI tasks. The training data will be obtained through the automatic generation of linguistic variability. The gold standard dataset (manually permuted by native Basque linguists, experts in the corresponding variation) obtained in the previous year will be used for evaluation.
3. Evaluation of the impact of variability in the performance of NLI and QA tasks, now with expanded data. Results will be compared against previously established zero-shot baselines.
4. Publication of synthetic data generation results for low-resource languages, as well as evaluation of NLI and QA tasks.

Year 3: test and improve Basque LLM performance with linguistic variability Work on Task 4 will continue by focusing on the expansion of the number of LLMs evaluated on the previous baseline for QA and NLI tasks with variation, as well as improving the capacity of those LLMs to process linguistic variability. In order to do so, Task 5's focal point will be to experiment with the state-of-the-art monolingual as well as multilingual LLMs available at the time of experimentation, with the aim of improving their ability to process linguistic variability in the tasks of QA and NLI.

It is also foreseen to go on a PhD stay to a foreign institution, where methods to deal with variation in other languages will be explored to see if they are also applicable to Basque or our analysis is applicable to other languages. This will contribute to the understanding of language-specificity in the linguistic variability field of NLP, while exploring a multilingual or language-agnostic approach.

The experiments and objectives planned for this year are the following:

1. Evaluation of monolingual and multilingual LLMs in the tasks of variability-aware QA and NLI.
2. Development of variability-aware LLMs, improving their performance for Basque language variability.
3. During the PhD stay, expansion of the language variation methods to other languages, thus analyzing the level of knowledge transfer ability when it comes to language variability.
4. Publication about the development of Basque variability-aware LLMs.
5. Publication on cross-lingual knowledge transfer across dialects of different languages.

Year 4: final experiments and thesis write-up A final iteration of Task 5 will be conducted, taking the most interesting conclusions obtained from previous years and rounding off new experiments in the first months of the year. Then the thesis will be written, and its defense preparation will be done.

The experiments and objectives planned for this year are the following:

1. Finish Task 5, thus finishing tasks and experiments of previous years.
2. Publication of the results of the language-variability aware LLMs.
3. Write-up of the PhD thesis and defense.

5. Specific Issues of Research to be Discussed

This thesis will work on the processing of language variation in Basque, both when it comes to dialects as well as registers. In doing so, the following challenges will need to be addressed:

- **Defining variation boundaries in Basque NLP:** One of the central difficulties is determining dialectal and register-based boundaries. How can we reliably annotate or detect linguistic variation when boundaries are often fluid and context-dependent? This raises both linguistic and methodological questions, especially for low-resource languages.

- **Data scarcity:** In NLP, no work has been done in terms of southern Basque dialects or registers. Therefore, I would like to discuss some strategies to overcome this constraint, such as data collection methods in low-resource environments. All collected data and developed software would be made publicly available under free licenses to support reproducibility and scientific advancement.
- **Synthetic data generation:** To alleviate data scarcity, synthetic data generation via LLMs presents a promising approach, as we can create sentence pairs by prompting models. I have currently tested some zero-shot prompting methods, and I would additionally like to discuss methods to produce authentic variability while avoiding overfitting or bias toward overly standardized forms.
- **Evaluation of generated data:** Evaluating the linguistic quality and task-relevance of generated data is a major challenge. I will start by manually evaluating a small sample of generated text in order to assess its quality. However, I would like some feedback on some form of scalability, so that I can expand this evaluation to larger amounts of highly variable data.

Acknowledgments

This thesis is funded by the Basque Government pre-doctoral grant (PRE_2024_1_0028).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] E. Coseriu, La geografía lingüística, volume 11, Universidad de la República, Facultad de Humanidades y Ciencias, 1956.

[2] A. Joshi, R. Dabre, D. Kanojia, Z. Li, H. Zhan, G. Haffari, D. Dippold, Natural language processing for dialects of a language: A survey, ACM Comput. Surv. 57 (2025). URL: https://doi.org/10.1145/3712060. doi:10.1145/3712060.

[3] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Y. Scherrer, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann, M. Zampieri (Eds.), Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19/. doi:10.18653/v1/2023.vardial-1.19.

[4] A. Ball-Burack, M. S. A. Lee, J. Cobbe, J. Singh, Differential tweetment: Mitigating racial dialect bias in harmful tweet detection, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 116–128. URL: https://doi.org/10.1145/3442188.3445875. doi:10.1145/3442188.3445875.

[5] O. Kuparinen, A. Miletić, Y. Scherrer, Dialect-to-standard normalization: A large-scale multilingual evaluation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 13814–13828. URL: https://aclanthology.org/2023.findings-emnlp.923/. doi:10.18653/v1/2023.findings-emnlp.923.

[6] T. Alshareef, M. A. Siddiqui, A seq2seq neural network based conversational agent for gulf arabic dialect, in: 2020 21st International Arab Conference on Information Technology (ACIT), 2020, pp. 1–7. doi:10.1109/ACIT50332.2020.9300059.

[7] N. Aepli, R. Sennrich, Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics,

Dublin, Ireland, 2022, pp. 4074–4083. URL: https://aclanthology.org/2022.findings-acl.321/. doi:10.18653/v1/2022.findings-acl.321.

[8] D. Demszky, D. Sharma, J. Clark, V. Prabhakaran, J. Eisenstein, Learning to recognize dialect features, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2315–2338. URL: https://aclanthology.org/2021.naacl-main.184/. doi:10.18653/v1/2021.naacl-main.184.

[9] E. Artemova, B. Plank, Low-resource bilingual dialect lexicon induction with large language models, in: T. Alumäe, M. Fishel (Eds.), Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), University of Tartu Library, Tórshavn, Faroe Islands, 2023, pp. 371–385. URL: https://aclanthology.org/2023.nodalida-1.39/.

[10] H. Hassan, M. Elaraby, A. Y. Tawfik, Synthetic data for neural machine translation of spoken-dialects, in: S. Sakti, M. Utiyama (Eds.), Proceedings of the 14th International Conference on Spoken Language Translation, International Workshop on Spoken Language Translation, Tokyo, Japan, 2017, pp. 82–89. URL: https://aclanthology.org/2017.iwslt-1.12/.

[11] F. Faisal, O. Ahia, A. Srivastava, K. Ahuja, D. Chiang, Y. Tsvetkov, A. Anastasopoulos, Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages, ArXiv abs/2403.11009 (2024). URL: https://api.semanticscholar.org/CorpusID:268513057.

[12] S. Rao, J. Tetreault, Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 129–140. URL: https://aclanthology.org/N18-1012/. doi:10.18653/v1/N18-1012.

[13] E. Eder, U. Krieg-Holz, M. Wiegand, A question of style: A dataset for analyzing formality on different levels, in: A. Vlachos, I. Augenstein (Eds.), Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 580–593. URL: https://aclanthology.org/2023.findings-eacl.42/. doi:10.18653/v1/2023.findings-eacl.42.

[14] E. Briakou, D. Lu, K. Zhang, J. Tetreault, Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3199–3216. URL: https://aclanthology.org/2021.naacl-main.256/. doi:10.18653/v1/2021.naacl-main.256.

[15] K. Zuazu, Euskalkiak. Euskararen dialektoak, Elkar, 2008.

[16] A. Estarrona, I. Etxeberria, R. Etxepare, M. Padilla-Moyano, A. Soraluze, Dealing with dialectal variation in the construction of the Basque historical corpus, in: M. Zampieri, P. Nakov, N. Ljubešić, J. Tiedemann, Y. Scherrer (Eds.), Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online), 2020, pp. 79–89. URL: https://aclanthology.org/2020.vardial-1.8/.

[17] L. Uria, R. Etxepare, 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 (2012) 117–135.

[18] M. M. I. Alam, S. Ahmadi, A. Anastasopoulos, CODET: A benchmark for contrastive dialectal evaluation of machine translation, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1790–1859. URL: https://aclanthology.org/2024.findings-eacl.125/.

[19] I. Z. Unzalu, The elaboration of basque in academic and professional domains, in: L. Grenoble, P. Lane, U. Royneland, N. O. Murchadha (Eds.), Linguistic Minorities in Europe Online, De Gruyter Mouton, Berlin, Boston, 2019. URL: https://doi.org/10.1515/lme.9612443, 2020.

[20] G. Bereziartua Etxeberria, M. M. Boillos Pereira, Euskara, hizkuntza akademikoa: laburpenen sistematizazioa helburu, Euskera Ikerketa Aldizkaria (2022) 33–62. URL: https://euskera-ikerketa.euskaltzaindia.eus/index.php/euskera/article/view/6. doi:10.59866/eia.vi67.6.

[21] B. M. Aseguinolaza, G. B. Etxeberria, Hitanoa euskal hiztunen komunitate garaikidean: molde zaharretatik ertz berrietara, Bat: Soziolinguistika aldizkaria (2022) 135–164.

[22] M. Alonso-Ramos, I. Zabala, Hartaes-vas: Lexical combinations for an academic writing aid tool in spanish and basque, in: CEUR Workshop Proceedings, volume 3224, CEUR-WS. org, 2022, pp. 22–25.

[23] J. Etxaniz, O. Sainz, N. Miguel, I. Aldabe, G. Rigau, E. Agirre, A. Ormazabal, M. Artetxe, A. Soroa, Latxa: An open language model and evaluation suite for Basque, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 14952–14972. URL: https://aclanthology.org/2024.acl-long.799/. doi:10.18653/v1/2024.acl-long.799.

[24] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. G. et al., Bloom: A 176b-parameter open-access multilingual language model, 2023. URL: https://arxiv.org/abs/2211.05100. arXiv:2211.05100.

[25] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. A. et al., Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[26] J. Alkorta, M. I. Quintian, Adding the Basque parliament corpus to ParlaMint project, in: D. Fišer, M. Eskevich, J. Lenardič, F. de Jong (Eds.), Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 107–110. URL: https://aclanthology.org/2022.parlaclarin-1.15/.

[27] N. Escribano, J. A. Gonzalez, J. Orbegozo-Terradillos, A. Larrondo-Ureta, S. Peña-Fernández, O. Perez-de Viñaspre, R. Agerri, BasqueParl: A bilingual corpus of Basque parliamentary transcriptions, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3382–3390. URL: https://aclanthology.org/2022.lrec-1.361/.

[28] M. Heredia, J. Etxaniz, M. Zulaika, X. Saralegi, J. Barnes, A. Soroa, XNLIeu: a dataset for cross-lingual NLI in Basque, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 4177–4188. URL: https://aclanthology.org/2024.naacl-long.234/. doi:10.18653/v1/2024.naacl-long.234.

[29] J. Etxaniz, G. Azkune, A. Soroa, O. L. de Lacalle, M. Artetxe, Bertaqa: How much do language models know about local culture?, 2024. URL: https://arxiv.org/abs/2406.07302. arXiv:2406.07302.