The Case of Spanish as a Pluricentric Language: Challenging the Monolingual Bias in NLP to Improve Cultural Adequacy of LLMs

Anonymous authors
Paper under double-blind review

Abstract

This position paper argues that the Natural Language Processing (NLP) community's oversight of Spanish's pluricentric nature undermines the development of culturally adequate models. Achieving truly effective NLP requires acknowledging the inherent cultural nuances embedded in language, yet a prevalent misconception persists that a singular "standard Spanish" originates primarily from Spain. Drawing on interdisciplinary insights, we believe that the distinction between "correct" and "exemplary" linguistic Spanish forms is key to effectively addressing the challenges posed by Spanish pluricentricity. This distinction allows the recognition of each Spanish-speaking nation as a distinct standardization center, where "exemplary" language is inherently community-defined. Maldonado Cárdenas (2012) applied this distinction to differentiate Spanish varieties, but with limited coverage. Motivated by these limitations, we propose a community-focused annotation framework to generate data for improving cultural adequacy in Large Language Models (LLMs), emphasizing broader engagement and contribution recognition. We then critically examine current multicultural datasets, highlighting shortcomings (e.g., limited representation, missing variation metadata), underscoring the urgent need for a more inclusive and culturally aware approach.

1 Introduction

From a functional perspective, language is a tool that enables the communication of thoughts, ideas, and concepts. Indeed, while linguistic competence is key to using language effectively, it is usually not enough to know the literal ("dictionary") meaning of words and compositional rules of the corresponding language (Ovchinnikova, 2012). Equally crucial is integrating the social and cultural knowledge shared by a language community. This topic has been a matter of interest in social and behavioural sciences, such as ethnography, sociolinguistics, and dialectology. However, the interest of the NLP community is recent, largely due to the limitations of current LLMs in their ability to adequately represent cultural variation (Jin et al., 2024). A clear example of this is the Spanish language.

According to the 2024 Yearbook of the Cervantes Institute (Fernández & Mella, 2024), there are 600 million Spanish speakers[1]. Spanish is the predominant language in 21 countries, which implies the existence of a great number of geographic varieties that influence the performance of language models ((Bogantes et al., 2016; Castillo-lópez et al., 2023), as cited in Grandury, 2024). Yet, this fact is often overlooked in favor of a perceived standard variety. For instance, the LDC catalog[2] and the ELRA Map[3] that Joshi et al. (2020) consulted to classify Spanish as a *quintessential rich-resource language* typically list all resources under generic labels like "Spanish; Castilian" or simply "Spanish". Although resources targeting particular variants do exist, identifying them often relies solely on keywords present in their titles, rather than proper metadata.

[1] Combining natives, limited competence, and foreign language learners.
[2] https://catalog.ldc.upenn.edu
[3] https://catalog.elra.info/en-us

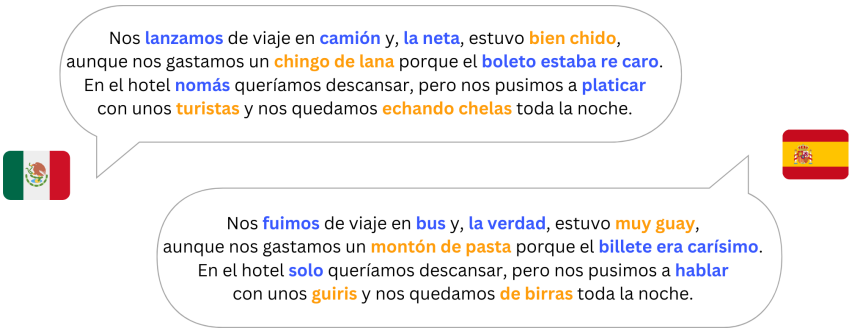

Figure 1: A Glimpse into Dialectal Variation. Here we illustrate some variations in vocabulary across two Spanish-speaking countries. Blue highlights common terms, while orange marks more casual expressions.

As shown in Figure 1, failing to account for different cultural contexts can lead to misunderstandings, including lexical choices, grammatical forms and stylistic conventions. Given its critical role in communication, how can we best define and measure the cultural adequacy of LLMs? How can insights from social sciences be applied to the evaluation of LLM cultural adequacy? These questions outline the central topic of this paper. We take the position that it is crucial to acknowledge Spanish's pluricentric nature from the earliest stages of data annotation. To that end, the subsequent section delves into the pluricentric status of Spanish, synthesizing key sociolinguistic insights that support our argument for adopting a community-focused approach to fostering cultural adequacy in NLP.

2 The Pluricentric Status of Spanish

The pluricentric nature of a language is directly related to its *standard variety*, as pluricentric languages have multiple centers from which standards emerge (Maldonado Cárdenas, 2012). Pöll (2021) observes that Spanish presents a unique case, as the Royal Spanish Academy historically sought to impose a single standard. However, the independence movements in Latin America triggered ongoing debates about what constitutes "correct" or "proper" Spanish. In this regard, Coseriu (1990) proposes the following distinction: a form can be CORRECT if it is accepted only within the context of a given situation or EXEMPLARY if it is perceived as the standard. This implies that a CORRECT form (e.g., Mexico: *pior*) would be rejected in situations where the EXEMPLARY form (e.g., Mexico: *peor*) is valid. Of the different approaches to Spanish's pluricentric nature, we align with Bierbach (2000) perspective, as it recognises **each Spanish-speaking nation as a distinct center of linguistic standardization**. In other words, each of those nations has their own EXEMPLARY forms. Building upon this proposition, Maldonado Cárdenas (2012) conducted a study using data collected through surveys and interviews with different speakers. As a result, the following groups were identified:

1. Pan-Hispanic forms (*exemplary* in Spain and America[4])
2. Pan-American forms (*exemplary* only in Spanish-speaking America)
3. Forms with widespread prestige in America
4. National forms (*exemplary* in a single country)

The participants were five natives of nine Spanish-speaking countries (i.e., 45 participants in total). Key inclusion criteria included a university degree and the place of residence, which had to coincide with the country of origin. Maldonado Cárdenas (2012) study confirms the presence of forms with national, regional and supraregional validity.

[4]In Spanish-language literature, "America" is used to denote the continent as a whole (i.e., North, Central, and South America).

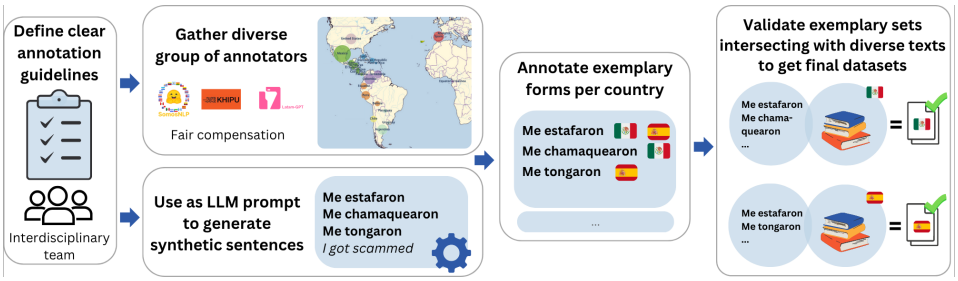

Figure 2: Overview of the main steps in our proposed framework.

2.1 Culture and Adequacy

Adequacy has to be seen in relation to action; something is *adequate* in relation to the purpose of what is done. Adequacy is thus a relation between means and purpose, and is thereby process-oriented (Reiss, 1983). In the case of language, adequacy can be understood as a relation between the communicative purpose and the chosen linguistic means. This highlights the process-oriented nature of language production, where speakers, driven by culturally shaped communicative goals, constantly make decisions about *lexical choices*, *grammatical forms* (e.g., tenses, moods), and other linguistic features to effectively convey their intended meaning.

Drawing from the definitions outlined in Sec. 2 and extending the concept of *adequacy*, we argue that the degree to which something is *culturally adequate* is related to the exemplarity of the lexical and grammatical forms.

3 Proposed Annotation Framework

The creation of a diverse data source is a crucial step towards our long-term objective of ensuring the inclusion of all Spanish variations in NLP research and applications. Drawing upon the insights presented in Section 2, we argue that the key to a diverse data collection and annotation process relies on the active involvement of experts and native speakers from different Spanish-speaking countries. A logical first step is replicating Maldonado Cárdenas (2012)'s study on a larger scale, increasing participant diversity and geographical coverage. The proposed framework, illustrated in Figure 2, consists of four steps:

Crowdsourcing for data collection. A foundational step in launching an inclusive NLP campaign is actively involving the target community in the data collection process. While crowdsourcing offers access to a diverse pool of annotators, ensuring truly representative data requires strategic collaboration with established communities within the relevant linguistic domains. To achieve this, we propose to engage with research groups and well-established Spanish-speaking NLP communities. One such community is SomosNLP[5], with thousands of social media followers and an active Discord server with 2,000 members who frequently participate in open-source multilingual initiatives. Other relevant, active groups include KHIPU[6], a biannual AI conference held in Latin America, and the National Center of AI in Chile (CENIA). To incentivize participation and foster collaboration, we propose simple yet significant public recognition. For example, offering them visibility through logo inclusion in our acknowledgements and actively seeking their input on key project decisions. Broad community participation is crucial for ensuring comprehensive representation across the Spanish-speaking world. Consequently, meticulous documentation of annotator origin (country/region) is crucial for data collection. Unlike Maldonado Cárdenas (2012) participant criteria, we believe that it is necessary to consider individuals with varying levels of formal education. Thus, we propose to expand outreach efforts to include schools,

[5] https://somosnlp.org
[6] https://khipu.ai

Dataset	Spanish Subset Size	Spanish-speaking Countries
BLEnD (Myung et al., 2024)	40k QA pairs	ES, MX
INCLUDE (Romanou et al., 2024)	550 QA pairs	PE, ES
Global MMLU (Singh et al., 2024a)	14k QA pairs	BO, HN, MX, PE, +
CVQA (Romero et al., 2024)	10k QA pairs	AR, CO, CL, EC, ES, MX, UY
Kaleidoscope (Salazar et al., 2025)	1.5k QA pairs	AR, CO, ES
Aya Collection (Singh et al., 2024b)	4.5M pairs	Unknown
#Somos600M (Grandury, 2024)	2.3M pairs	AR, ES, CL, CO, CR, MX, PE, PY, VE
FineWeb2 (Penedo et al., 2025)	54k annotations	Unknown

Table 1: Size and Spanish-speaking countries represented in each dataset, combining the information from the corresponding papers and datasets available on the Hugging Face Hub.

companies, and public institutions. Fair compensation for annotators is essential. Options may include co-authorship and/or financial incentives where possible.

Synthetic data generation. In Maldonado Cárdenas (2012)'s study, participants were shown different linguistic forms and asked to select the one they considered exemplary. We propose the implementation of a similar approach by leveraging the generative capabilities of LLMs to synthesize a comparable dataset. Alternatively, everyday scenario-based dialogue generation can be used to create a culturally diverse dataset. This proposal better aligns with the fact that cultural adequacy is reflected in everyday activities and is closely tied to the ability to clearly understand or communicate a specific purpose (Reiss, 1983). This technique has prior applications in NLP, particularly in chat dialogue (Li et al., 2022) and commonsense reasoning (Ostermann et al., 2018; 2019).

Annotation plan. Since the study of cultural adequacy addresses a sociolinguistic phenomenon, it is essential to consider an interdisciplinary team of experts. Researchers from NLP, linguistics, and social sciences should join efforts in prompt engineering, structuring the annotation guidelines and planning the corresponding annotator training.

Exemplary data validation. Evaluation is one of the most critical and challenging aspects of this framework. Determining whether a given linguistic form is "exemplary" does not have an absolute ground truth. While geographic dialect dictionaries and glossaries can serve as references, ultimately, what is considered exemplary is a community-driven decision. Our proposal considers implementing Maldonado Cárdenas (2012)'s validation approach, which involves analyzing the frequency of words and phrases deemed exemplary within corpora of texts from the corresponding countries such as Española (2024); de Henares: Universidad de Alcalá (2024).

4 Dataset Overview

Over the past year, we have seen a growing interest in the creation of multicultural datasets, which are usually the product of international and cross-institutional collaborations. In Table 1 we summarise the most notable ones. **BLEnD** is a multiple-choice question answering (MCQA) benchmark focused on everyday knowledge representing 16 regions and 13 languages, **INCLUDE** is a benchmark composed of a collection of MCQA exams with a focus on culture covering 41 languages, **Global MMLU** is a machine-translated version of MMLU to 42 languages and annotated with cultural information, **CVQA** is a multimodal MCQA benchmark showcasing everyday situations and covering 39 country-language pairs, **Kaleidoscope** is a multimodal MCQA benchmark derived from exam questions covering 18 languages and 14 subjects, **Aya** is an instructions dataset covering 65 languages (71 including dialects and scripts), **#Somos600M** is an instructions dataset in Spanish covering 9 countries created by the teams of a hackathon, and **FineWeb2** is a multilingual collection of texts annotated by educational level. Here, we compare these data collection efforts across several dimensions, evaluating their approach to representing Spanish linguistic variation and their utility to promote culturally adequate NLP.

Community engagement. Most campaigns prioritize native or fluent speakers, but the extent of community involvement varies. The team behind INCLUDE launched an open call for contributions while actively engaging with linguistic groups and regional associations. Similarly, Aya, FineWeb2, Kaleidoscope, and #Somos600M embrace a fully community-driven approach, encouraging open contributions, particularly from native speakers within the CohereForAI open-science[7], Hugging Face[8], and SomosNLP communities. Global MMLU, on the other hand, includes both professional and community translators, but the level of participation varies across languages, limiting the potential for consistent cultural representation. BLEnD collects contributions from professional native speakers, and CVQA was created through a network of personally invited international research groups.

Demographic reporting. Among the analyzed campaigns, CVQA, BLEnD and Aya provide the annotators' age and gender. Additionally, CVQA includes the annotators' country of residence and country of origin. BLEnD also reports their residence duration and educational level. Aya does not disclose more of the demographic information they collected. #Somos600M publishes the gender and country of origin of the contributors, disclosed by each team. However, GlobalMMLU, Kaleidoscope, INCLUDE, and FineWeb2 lack demographic reporting, making it difficult to assess the representativeness of their data.

Spanish variations. A significant limitation across these datasets is the underrepresentation of Spanish linguistic diversity, often limited to a few countries or lacking clear origin information for contributions. Aya acknowledges regional diversity in its language contributions but does not provide a precise breakdown of the countries involved. Similarly, Global MMLU hired professional translators for Spanish without reporting which dialectal variations were covered. FineWeb2 incorporates Spanish, regional dialects, and indigenous languages, but lacks information on the origin of Spanish-speaking contributors. INCLUDE, CVQA, Kaleidoscope, and #Somos600M provide the country of origin per sample, annotated by the community data contributors.

Compensation and incentives. We find two compensation models: academic recognition and direct monetary incentives. CVQA, INCLUDE, and Kaleidoscope recognize contributors through paper co-authorship, and Aya only acknowledges contributors. BLEnD provides monetary compensation, though details are vague beyond a minimum payment threshold. Global MMLU compensates professional translators but does not offer remuneration to community collaborators, potentially creating disparities in data quality across languages. #Somos600M provides access to computing, acknowledgements and support to publish a paper, in addition to prizes for the best projects. FineWeb2 does not provide information about its compensation strategy.

Data quality and evaluation. In BLEnD, there is at least one author who is a native speaker of the language and originally from the country/region represented. For INCLUDE and Kaleidoscope, each community volunteer who processed an exam was held accountable for its quality. FineWeb2 named "Language Leads" to support the community volunteers and ensure the quality of the annotations. The projects presented to #Somos600M were evaluated by a jury. The other dataset papers do not disclose information about their data validation process.

5 Conclusion

While Spanish is not typically considered a low-resource language in NLP, existing resources often exhibit significant gaps in their coverage of the diverse linguistic variations found across the Spanish-speaking regions. As LLMs are increasingly used to create synthetic evaluation datasets (Zhao et al., 2024), it is essential to ensure their cultural adequacy. This paper argues that NLP research must explicitly acknowledge the pluricentric nature of Spanish and actively incorporate strategies to account for this linguistic diversity. We hope that the framework and dataset overview presented help guide future efforts towards developing more inclusive and representative NLP resources for Spanish.

[7] https://cohere.com/research/open-science

[8] https://huggingface.co

Limitations

Multilinguality in LATAM. The framework proposed in this paper focuses on Spanish, a pluricentric language with well-documented geographic variation. However, our long-term aim is to represent the diversity of the whole Latin American community. In this region, Spanish is the predominant language but it coexists with other pluricentric Romance languages (i.e., Portuguese and French) and indigenous languages (e.g., Quechua, Guaraní, Maya, Aimara, and Náhuatl). Extending this methodology to other pluricentric languages might need additional adaptation. In particular, tailoring to each sociolinguistic context the annotation guidelines, data collection strategies, reach to local communities, and evaluation criteria to align with language-specific cultural norms.

Existing linguistic resources. In this work, we mention some existing resources created by experts in linguistics and sociology, such as dictionaries and glossaries. However, we did not provide an exhaustive list, nor delve deeply into their potential applications within NLP. Future work should explore how these resources can be systematically integrated into NLP methodologies to enhance cultural adequacy assessments.

Community representation. Although we emphasize the need for broad community participation, there are inherent challenges in ensuring truly representative data collection. We plan to implement this methodology and refine it with our experience, sharing key takeaways and recommendations with the broader community to improve data collection campaigns.

Subjectivity in annotation. Assessing cultural adequacy and linguistic exemplarity involves subjective judgments influenced by annotators' personal experiences, educational background, and exposure to different dialects. Our current proposal does not discuss how to take into account these factors in the annotation process. An important point to keep in mind is the disagreement between annotators, which should be expected and treated as a valuable signal rather than an error.

Ethics Statement

Value human contribution to AI. NLP models and applications rely heavily on human labor, including data annotation and linguistic expertise. Our framework explicitly values and incorporates contributions from native speakers and domain experts, ensuring that human input remains central. Our proposal recognizes that fairly compensating human contributors, which is essential in preventing the devaluation of their labor under the guise of AI-driven automation.

AI cannot solve all societal problems. AI is often framed as a solution to complex societal issues, but it cannot, by itself, resolve systemic problems such as linguistic discrimination or cultural erasure. While our framework contributes to a better representation of Spanish varieties, it does not eliminate the broader socio-political challenges that impact language use and recognition.

Synthetic data generation. The success of the framework relies on the availability of high-quality, diverse textual data. Utilizing LLMs for synthetic data generation allows for the creation of the intended extensive dataset. However, it presents its own risks, including reinforcement of existing biases in the models and overrepresentation of standardized linguistic forms at the expense of less-documented variations.

References

Mechtild Bierbach. " spanisch-eine plurizentrische sprache?" zum problem von" norma culta" und varietät in der hispanophonen welt. *Vox romanica*, 59:143, 2000.

Diana Bogantes, Eric Rodríguez, Alejandro Arauco, Alejandro Rodríguez, and Agata Savary. Towards lexical encoding of multi-word expressions in Spanish dialects. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis

(eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2255–2261, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1358/.

Galo Castillo-lópez, Arij Riabi, and Djamé Seddah. Analyzing zero-shot transfer scenarios across Spanish variants for hate speech detection. In Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri (eds.), *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pp. 1–13, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.vardial-1.1. URL https://aclanthology.org/2023.vardial-1.1/.

Eugenio Coseriu. El español de américa y la unidad del idioma. In *Actas del I Simposio de Filología Iberoamericana*, pp. 43–75. Pórtico Zaragoza, 1990.

Alcalá de Henares: Universidad de Alcalá. Corpus del proyecto para el estudio sociolingüístico del español de españa y de américa. https://preseea.uah.es/corpus-preseea, 2024. Accessed: 2025-02-6.

Real Academia Española. Banco de datos (corpes xxi) [en línea]. corpus del español del siglo xxi (corpes). https://www.rae.es/corpes/, 2024. Accessed: 2025-02-6.

Francisco Moreno Fernández and Héctor Álvarez Mella. Demografía del español en el mundo 2024. In *El español en el mundo. Anuario del Instituto Cervantes 2024*, pp. 30–97. Instituto Cervantes, 2024.

María Grandury. The #somos600m project: Generating nlp resources that represent the diversity of the languages from latam, the caribbean, and spain. In *North American Chapter of the Association for Computational Linguistics Conference: LatinX in AI (LXAI) Research Workshop*, 2024.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez Adauto, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. Multilingual trolley problems for language models. In *Pluralistic Alignment Workshop at NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=vrHErHkCNo.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560/.

Dawei Li, Yanran Li, Jiayi Zhang, Ke Li, Chen Wei, Jianwei Cui, and Bin Wang. C^3KG: A Chinese commonsense conversation knowledge graph. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1369–1383, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.107. URL https://aclanthology.org/2022.findings-acl.107/.

Mireya Maldonado Cárdenas. Español como lengua pluricéntrica: Algunas formas ejemplares del español peninsular y del español en américa. *Español,¿ desde las variedades a la lengua pluricéntrica?.-(Lengua y sociedad en el mundo hispánico; v. 30)*, pp. 95–122, 2012.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages, 2024. URL https://arxiv.org/abs/2406.09948.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1564/.

Simon Ostermann, Michael Roth, and Manfred Pinkal. MCScript2.0: A machine comprehension corpus focused on script events and participants. In Rada Mihalcea, Ekaterina Shutova, Lun-Wei Ku, Kilian Evang, and Soujanya Poria (eds.), *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pp. 103–117, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-1012. URL https://aclanthology.org/S19-1012/.

E. Ovchinnikova. *Integration of World Knowledge for Natural Language Understanding*. Atlantis Thinking Machines. Atlantis Press, 2012. ISBN 9789491216534. URL https://books.google.es/books?id=jfJUHOncFzkC.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*, 2025.

Bernhard Pöll. *Spanish Today: Pluricentricity and Codification*, pp. 163–183. Cambridge University Press, 2021.

Katharina Reiss. Adequacy and equivalence in translation. *The Bible Translator*, 34(3):301–308, 1983.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*, 2024.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark, 2024. URL https://arxiv.org/abs/2406.05967.

Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh, Fabian Farestam, Angelika Romanou, Danylo Boiko, Dipika Khullar, Mike Zhang, Dominik Krzemiński, Jekaterina Novikova, Luísa Shimabucoro, Joseph Marvin Imperial, Rishabh Maheshwary, Sharad Duwal, Alfonso Amayuelas, Swati Rajwal, Jebish Purbey, Ahmed Ruby, Nicholas Popovič, Marek Suppa, Azmine Toushik Wasi, Ram

Mohan Rao Kadiyala, Olga Tsymboi, Maksim Kostritsya, Bardia Soltani Moakhar, Gabriel da Costa Merlin, Otávio Ferracioli Coletti, Maral Jabbari Shiviari, MohammadAmin farahani fard, Silvia Fernandez, María Grandury, Dmitry Abulkhanov, Drishti Sharma, Andre Guarnier De Mitri, Leticia Bossatto Marchezi, Setayesh Heydari, Johan Obando-Ceron, Nazar Kohut, Beyza Ermis, Desmond Elliott, Enzo Ferrante, Sara Hooker, and Marzieh Fadaee. Kaleidoscope: In-language exams for massively multilingual vision evaluation, 2025. URL https://arxiv.org/abs/2504.07072.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*, 2024a.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.620. URL https://aclanthology.org/2024.acl-long.620/.

Raoyuan Zhao, Abdullatif Köksal, Yihong Liu, Leonie Weissweiler, Anna Korhonen, and Hinrich Schuetze. SynthEval: Hybrid behavioral testing of NLP models with synthetic CheckLists. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7017–7034, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.412. URL https://aclanthology.org/2024.findings-emnlp.412/.