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



María Grandury • • , Diana Galvan-Sosa • •

Universidad Politécnica de Madrid | PEPFL | University of Cambridge | SomosNLP

Overview

- Main problem: The "standard Spanish" misconception hinders truly effective NLP.
- Our position: NLP must account for cultural diversity by recognizing the pluricentricity of Spanish.
- Our proposal: Community-driven annotation framework that validates each Spanish-speaking nation as a center of linguistic standardization.

The Pluricentric Status of Spanish

Nos lanzamos de viaje en camión y, la neta, estuvo bien chido, aunque nos gastamos un chingo de lana porque el boleto estaba re caro. En el hotel nomás queríamos descansar, pero nos pusimos a platicar con unos turistas y nos quedamos echando chelas toda la noche.



Nos fuimos de viaje en bus y, la verdad, estuvo muy guay, aunque nos gastamos un montón de pasta porque el billete era carísimo. En el hotel solo queríamos descansar, pero nos pusimos a hablar con unos guiris y nos quedamos de birras toda la noche.

- Pluricentric languages have multiple centers from which standards emerge.
- Coseriu (1990) distinction: Linguistic forms are either CORRECT (situational) or **EXEMPLARY** (standard).
- Our central claim: **Cultural** adequacy hinges on the exemplarity of linguistic and gramatical forms.

Proposed Annotation Framework

Define clear

annotation

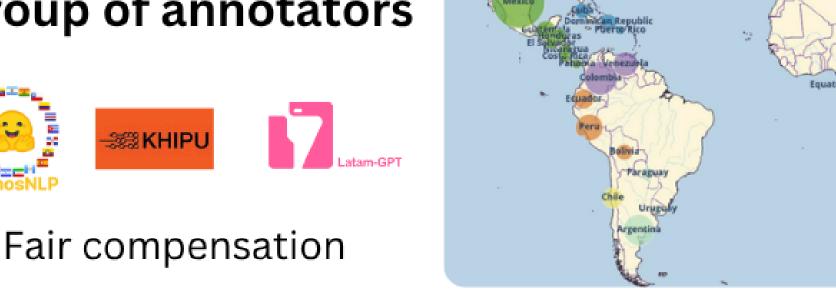
guidelines

Crowdsourcing for data collection

Gather diverse group of annotators







Use as LLM prompt to generate synthetic sentences

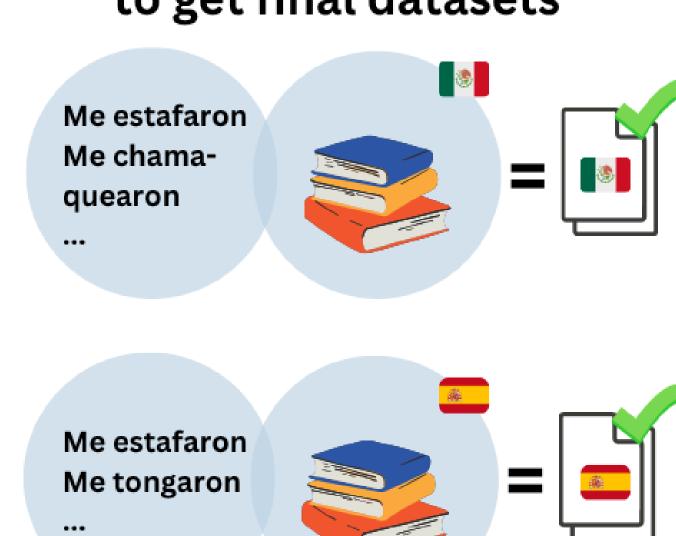
Me estafaron Me chamaquearon Me tongaron I got scammed

Annotate exemplary forms per country



Validation

Validate exemplary sets intersecting with diverse texts to get final datasets



Annotation plan

Interdisciplinary

team

Synthetic data generation

Spanish Representation in Multicultural Datasets

Dataset	Spanish Subset Size	Spanish-speaking Countries
BLEnD (Myung et al., 2024)	40k QA pairs	ES, MX
INCLUDE (Romanou et al., 2024)	550k QA pairs	PE, ES
Global MMLU (Singh et al., 2024)	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., 2024)	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

- Community engagement
- Demographic reporting
- Spanish variations
- Compensation
- **Data quality and eval**

- Most prioritize native or fluent speakers. Variable community involvement.
- Inconsistent: half of them omit it, others detail country of origin and residence.
- Significant underrepresentation of Spanish linguistic diversity.
- Two models: academic recognition and direct monetary incentives.
- Inconsistent: half of them omit it, others rely on "language leads".

Ongoing work

- Survey existing resources (dictionaries, glossaries) created by linguistic and sociological experts.
- Review multilingual and Spanish-only shared tasks.

Join our cultural data collection!

We aim to create a **cultural benchmark** and an alignment dataset for Ibero-America. Join us to improve the representation of your country!

