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, 🟢 University of Cambridge, 🔶 SomosNLP

MELT Workshop @COLM | October 10, 2025

INDEX

MAIN PROBLEM

The "standard Spanish" misconception hinders truly effective NLP.

OUR POSITION

NLP must account for cultural diversity by recognizing the pluricentricity of ES.

OUR PROPOSAL

Community-driven annotation framework, each nation as center of standardization.

600M

Spanish speakers

SomosNLP

SomosNLP

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.

THE PLURICENTRIC STATUS OF SPANISH

- Pluricentric languages have multiple centers from which standards emerge.
- Definition by Coseriu (1990): Linguistic forms are either CORRECT (situational, “pior”) or **EXEMPLARY (standard, “peor”)**.
- Each Spanish-speaking nation as a distinct center of linguistic standardization.
- Our central claim: **Cultural adequacy** hinges on the exemplarity of linguistic and grammatical forms.

“It is crucial to acknowledge Spanish’s pluricentric nature from the earliest stages of data annotation.”

PROPOSED FRAMEWORK

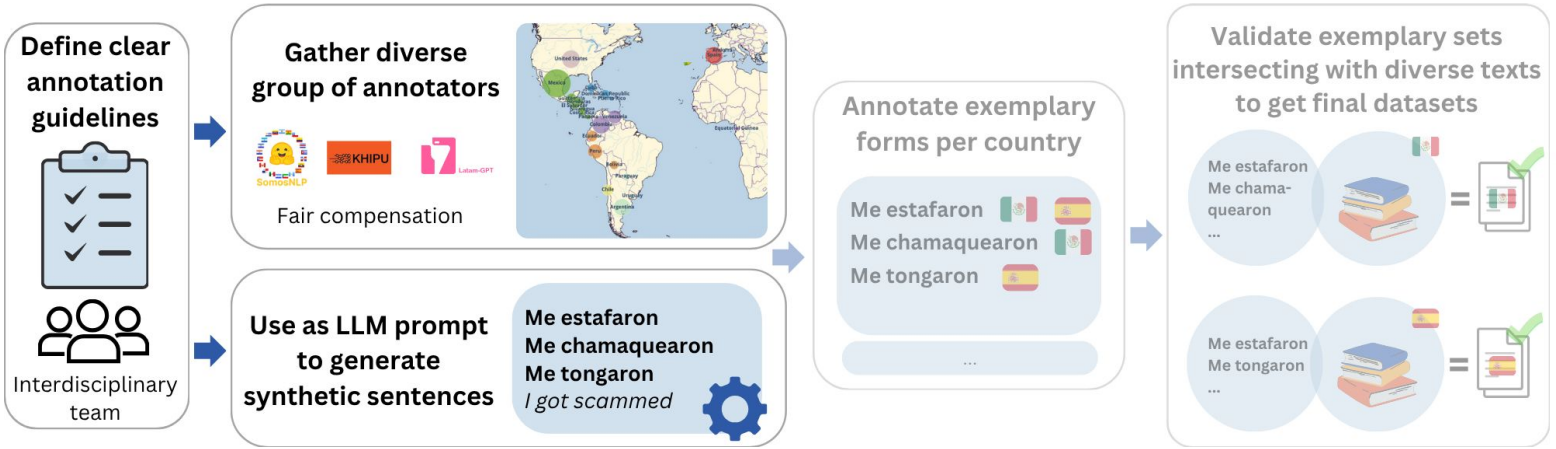
#1: ANNOTATION PLAN

PROPOSED FRAMEWORK

#2: CROWDSOURCING FOR DATA COLLECTION

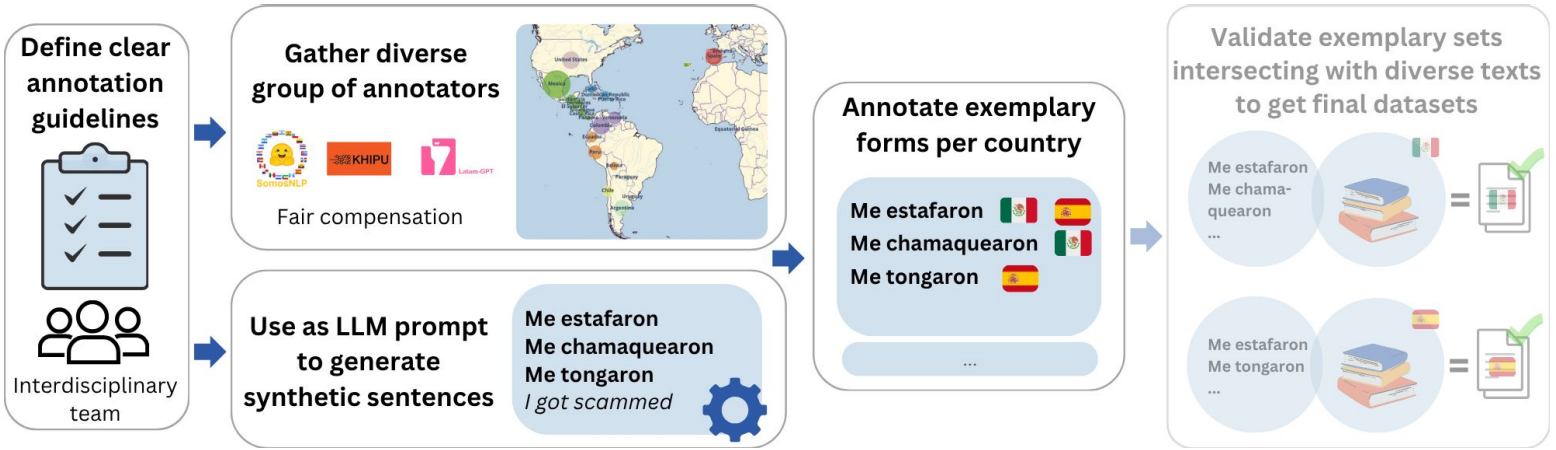
PROPOSED FRAMEWORK

#3: SYNTHETIC DATA GENERATION



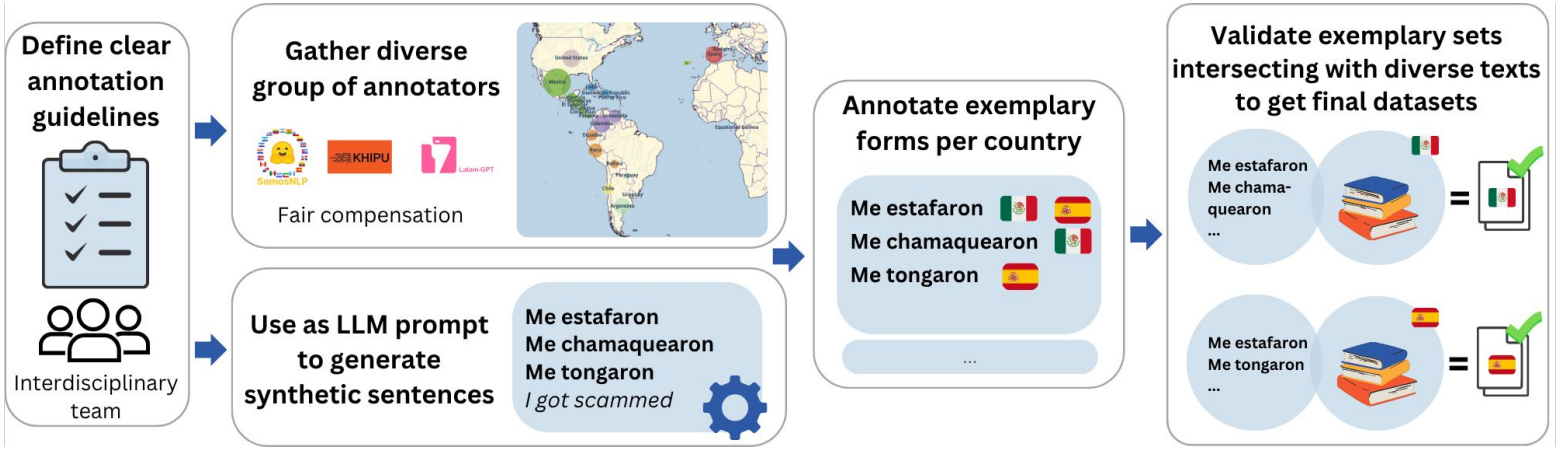
PROPOSED FRAMEWORK

#4: EXEMPLARY DATA ANNOTATION



PROPOSED FRAMEWORK

#5: EXEMPLARY DATA VALIDATION



CULTURAL DATA COLLECTION CAMPAIGNS

- Most **are open to the community**
- Poor **demographic reporting**
- Significant underrepresentation of **Spanish linguistic diversity**
- **Compensation** ranging from acknowledgments to monetary
- **Data quality** review: “language leads”, peer-review, or no info

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

PILOT CAMPAIGN

SOMOSNLP CULTURAL HACKATHON

- Goal: Cultural preference dataset with LLM Arena
- Data quality: Illustrated guidelines, quiz, and peer-review

ONGOING AND FUTURE WORK

- Survey existing **resources** (dictionaries, glossaries) created by **linguistic and sociological** experts.
- Review multilingual and Spanish-only **shared tasks**.
- Extend the pilot collection campaign.

THANK YOU VERY MUCH!

María Grandury and Diana Galvan-Sosa

mariagrandury@somosnlp.org

NLP RESOURCES IN SPANISH

LANGUAGE MODELS & PRE-TRAINING DATA

- LMs: local initiatives, multilingual open-source & open-weights
- Corpora: library (BNE), web (CEREAL, Red-Pajama, FineWeb2)

Source: CEREAL, which comes from the **Spanish portion of OSCAR**

NLP RESOURCES IN SPANISH

FORUMS, SHARED TASKS & BENCHMARKS

- IBERLEF, CLEF, SemEval, ACL
- **60% reviewed datasets do not include country labels, 30% come from Spain**
- LATAM Spanish: MX, UY
- Co-existing languages: ILENIA, AmericasNLP

NLP RESOURCES IN SPANISH

LEADERBOARDS

- Multilingual leaderboards: Usually benchmarks translated into Spanish
- Original Spanish: LM Arena, SCALE, **ODESIA**
- Original co-existing languages: CLUB (Catalan)

ODESIA

Odesia Core Tasks

Spanish | English

System	Arithmetic mean	EXIST 2022: Sexism detection	EXIST 2022: Sexism categorisation	DIPROMATS 2023: Propaganda identification	DIPROMATS 2023: Coarse propaganda characterization	DIPROMATS 2023: Fine-grained propaganda characterization	DIAI 202 Disab detec
Xlm roberta large	**0.5873**	0.7663	0.5593	0.8186	0.5343	0.4527	0.785
Qwen2.5-7B	**0.5858**	0.7816	0.5526	0.8282	0.5572	0.4528	0.645

Bilingual, 14 original EN/ES tasks, difficulty calibration, zero contamination

PROPOSALS TO INCREASE DIVERSITY

#6: IBERO-AMERICAN RESOURCES & EVAL WORKSHOP

Target **dialectal and cultural diversity**, fair multicultural evals, highlight initiatives & foster collaboration

Recurrent workshop, (non)-archival tracks, 1-day: posters, keynote talks, panel & open discussion

Built upon hackathons (2000 registrations), **ACL** BoF + social (60 participants), **COLM** social (today)

CONCLUSIONS AND NEXT STEPS

Contributions & current situation	Planned steps	Long-term vision
• La Leaderboard • First Spanish varieties benchmark • Extended INCLUDE-ES from 2 to 17 countries • *There is still limited data labeled with country information*	• Organize workshop and shared task • Extend INCLUDE to 21 countries • Include LATAM benchmarks in La Leaderboard and provide per-variety results	• **Multicultural evaluation as default** • Diverse, responsible, open, community-led evaluation that closes the gap between reported and real-world performance

LA LEADERBOARD

PROPOSALS TO INCREASE DIVERSITY

DATA COLLECTION FRAMEWORK

- Objective