LLM-Assisted Spanish Dialect Corpus Construction

Jessica C. Ramirez Vidal, Hiroki Ouchi, Sakriani Sakti

Nara Institute of Science and Technology
Ikoma, Nara 630-0192, Japan
jessica.claribel_ramirez.jg1@naist.ac.jp, {hiroki.ouchi, ssakti}@is.naist.jp

Abstract

This study presents a multi-dialect, pragmatically annotated Spanish corpus designed to address persistent gaps in the representation of regional varieties and communicative functions in existing linguistic and NLP resources. The corpus focuses exclusively on Spanish dialects spoken in the Americas, selecting one representative dialect per country and incorporating a single neutral Castilian variety for comparative purposes. Dialects are organized into five regional groups: Mexican, Central American, Caribbean, South American, and Rioplatense Spanish. Corpus development follows a multi-stage workflow in which a seed lexicon composed of openly licensed material from sources such as Wikipedia, Project Gutenberg, and curated random and synthetic data is used to initiate the LLM-based text generation. Each base sentence is expanded into dialect-specific variants and annotated with pragmatic and domain labels, producing a fully parallel dataset that supports cross dialect comparison. A multi-stage correction pipeline combining automated scripts, controlled LLM-based editing, and manual review ensures syntactic well-formedness and dialectal authenticity while eliminating language-switching and hallucination errors. The final version of the corpus covers 20 dialects and contains, 40,000 annotated sentences, released in both JSON and plain-text formats for use in a wide range of NLP tasks.

Keywords: Spanish Dialect, Pragmatic, Domain

1. Introduction

Large Language Models (LLMs) have transformed the methodological landscape of corpus construction, offering new possibilities for scaling, diversifying, and annotating linguistic datasets. Traditional corpus-building practices, often constrained by limited resources, time-intensive manual annotation, and uneven representation across dialects and modalities are increasingly complemented by LLM-driven techniques that enable rapid data generation, systematic variation, and multi-layer linguistic tagging. These models facilitate the creation of corpora for low-resource languages and dialects. As a result, LLMs are becoming central tools in contemporary corpus linguistics by accelerating data creation and enabling the integration of multimodal data sources.

Spanish is one of the most widely spoken languages globally and is formally established as an official language in a total of 21 countries spanning Europe, the Americas and Africa. Spanish is the second most spoken language in the United States, and is widely used in countries like Andorra, Gibraltar, Belize, Aruba, the Virgin Islands, among others (Molina Martos, 2024).

2. Related Work

Research on Spanish variation has long emphasized the language's extensive dialectal diversity across Latin America and the Iberian Peninsula. Quesada Pacheco (2021) examines phonological changes in Central American Spanish, while Molina Martos (2024) proposes a dialectal division focused on Peninsular Spanish, and Schlumpf and Carreira (2024) present a corpus of contemporary Equatoguinean Spanish, underscoring the need for resources that include less documented varieties.

The rapid expansion of LLM-based corpus construction has reshaped methodological possibilities, prompting a wave of studies exploring how human expertise and model-driven automation can be combined. Hybrid workflows have been examined in depth: Weissweiler et al. (2025) show that collaborative human–LLM approaches can effectively support the creation and evaluation of corpora targeting specific syntactic constructions, while Morin and Larsson (2025) introduce a large-scale unsupervised pipeline for automatic annotation. McCallum and Mizumoto (2025) provide a broader assessment of these developments, emphasizing that although LLMs can accelerate data creation, they also introduce risks of inconsistency and hallucination that require careful methodological control. Similar concerns appear in domain-specific applications, such as Sakai et al. (2024), who employ LLMs to construct a simultaneous-interpretation corpus, and Shen et al. (2024) together with Ma et al. (2025), who investigate how general-purpose LLMs can be repurposed for specialized annotation tasks, including pragmatic labeling.

Moreover, Kawasaki (2026) documents systematic digital linguistic bias in models' handling of Spanish lexical variation, showing a tendency to overproduce features associated with high-resource dialects such as Mexican Spanish. Pozhemetskiy (2025) evaluates LLM-generated Andalusian Spanish and finds that although models can approximate regional traits, their outputs remain unstable and prone to

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 153–159
11–12 May 2026.

overgeneralization. Work on low-resource dialect translation, including Yakhni and Chehab's (2025) study of Lebanese Arabic and shared tasks on dialect-to-MSA translation (Abdelaziz et al., 2024), highlights persistent challenges in capturing dialectal variation within multilingual systems. Surveys of machine translation in the LLM era (Ataman et al., 2025) further underscore the persistence of these issues across multilingual settings.

Despite the vast dialect landscape of Latin America, current research on Spanish dialect corpora does not include all Latin American countries; in general, studies focus on and compare only a small subset of dialectal varieties (Carcelén-Guerrero et al., 2025; Fuchs & González, 2022; Hernández Mena et al., 2017). Furthermore, large Spanish corpora that provide a global view of Latin American variation are not freely available for full-scale data extraction and independent use. To address this gap, the main goal of this study is to develop a multi-dialect Spanish corpus with pragmatic annotation and domain labeling, to be used as a resource for NLP tasks.

3. Scope

This study focuses exclusively on Spanish dialects spoken in the Americas, selecting one representative dialect per country to ensure broad yet manageable coverage of regional variation. In addition to these American varieties, the corpus includes a single neutral or standardized Castilian variety to serve as a comparative reference point. Importantly, the corpus does not incorporate the internal dialectal diversity of Spain nor Equatorial Guinean Spanish as these fall outside the intended geographic and linguistic scope of the study.

Selecting one representative dialect per country ensures geographic coverage while reducing internal complexity. Including a single neutral Castilian variety provides a useful reference point for contrastive analysis while keeping focus on American Spanish.

Dialect	Country
Mexican	Mexico
Caribbean	Dominican Republic, Cuba, Puerto Rico
Central American	Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica, Panama
South American	Ecuador, Venezuela, Chile, Colombia, Bolivia, Paraguay
Rioplatense	Argentina, Uruguay

Table 1: Countries per dialect

Although numerous classifications of Spanish dialects exist, this study adopts a five-group regional framework based on geographical location: Mexican, Central American, Caribbean, South American, and Rioplatense. The Rioplatense variety is treated as a separate category rather than being subsumed under the broader South American group due to its distinctive phonological and lexical characteristics (table 1).

4. Corpus construction

Figure 1 illustrates the overall architecture of the corpus-construction pipeline, outlining the sequential and parallel processes involved in generating, validating, and annotating the dataset. The workflow begins with a Seed Lexicon consisting of sentences in neutral Spanish. This lexicon feeds into the LLM, which produces the initial textual material and simultaneously supports two annotation pathways: pragmatic annotation and domain annotation. The primary output of the LLM is the Raw Corpus, a collection of unprocessed text samples generated according to the dialectal and pragmatic constraints defined at the outset. This raw material then undergoes a full linguistic integrity assessment, including a syntactic check and correction stage in which grammatical inconsistencies and unnatural constructions are identified and revised to ensure linguistic accuracy and dialectal authenticity. Through this multi-stage workflow we obtain the final version of the corpus.

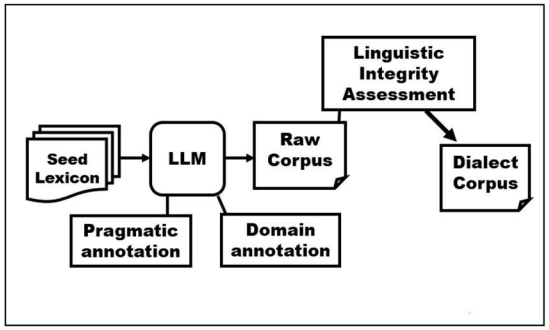

Figure 1: Dialect corpus architecture

4.1 Data collection

The seed lexicon was compiled from multiple data sources, which are summarized in Table 2.

Data Source	Sentences	%
Wikipedia	500	25
Random and Synthetic data	1,000	50
Project Gutenberg	500	25
Total	2,000	100

Table 2: Seed data

The lexicon contains 2,000 sentences extracted from Wikipedia (25%), Project Gutenberg (25%), and a set of random or synthetic texts (50%).

4.1.1 Wikipedia

Wikipedia is an online multilingual encyclopedia that provides openly licensed articles on a wide range of topics, making it a valuable and widely adopted resource in the NLP community. [1]

We downloaded the Spanish dump data from 10/01/2025. We pruned the dataset to include only articles containing a list or table referencing a Latin American country, subsequently extracting the Spanish titles and their cross-dialectal equivalents.

4.1.2 Project Gutenberg

Project Gutenberg served as an additional source for constructing the seed lexicon, offering access to a large collection of public-domain books that can be freely downloaded and redistributed. [2]

We downloaded Spanish-language and extract just sentences that contain dialogue markers (e.g., "–") as these tend to provide naturalistic conversational structures.

4.1.3 Random and Synthetic data

The use of synthetic data has emerged as a practical strategy in corpus construction, enabling the controlled creation of sentences that supplement traditional text sources.

To complement the dataset, we incorporated a substantial portion of random and synthetic data into the seed lexicon. The synthetic sentences were generated by prompting Large Language Models—specifically Gemini and ChatGPT-5 to produce the Spanish sentences that could later be expanded into dialect-specific variants.

4.2 LLM

Zero-shot prompting was applied specifically with ChatGPT-5, which was instructed to translate each Spanish sentence into the 19 target dialects.

> Translate: ¿Cómo estás? to Caribbean dialects: Dominican Spanish, Puerto Rican Spanish, Cuban Spanish, Mexican Dialect: Mexican Spanish, Rioplatense dialects: Argentinian Spanish, Uruguayan Spanish, Central_American dialects: Guatemalan Spanish, Honduran Spanish, Costa Rican Spanish, Panamanian Spanish, South_ American dialects: Venezuelan Spanish, Peruvian Spanish, Colombian Spanish, Ecuatorian Spanish, Bolivian Spanish, Chilean Spanish, Paraguayan Spanish.

Figure 2: Zero-shot prompt

Figure 2 presents the prompt template used to instruct the model to produce dialectal variants of the greeting ¿Cómo estás? (How are you?).

4.3 Annotation Schema

Pragmatic and domain annotation for the seed-lexicon sentences was carried out using ChatGPT-5. For the Wikipedia portion of the dataset, we relied on the platform's existing folksonomy as the primary source for assigning domain tags.

4.3.1 Pragmatic Annotation

Pragmatic annotation in this corpus focuses on labeling the communicative function and interactional role of each sentence, allowing the dataset to capture how speakers use language in context rather than only its grammatical form. Each sentence is assigned one or more pragmatic categories that reflect its intended communicative purpose.

For example, the greeting ¿Cómo estás? receives two pragmatic tags: informal and greeting. Other categories include functions such as request, offer, refusal, expression of thanks, and additional common interactional acts. These labels help distinguish sentences that may be structurally similar but serve different purposes in communication. The annotation scheme is designed to remain consistent across all dialects, ensuring that equivalent functions can be compared even when expressed with different lexical or syntactic choices.

4.3.2 Domain Annotation

In addition to pragmatic labels, each sentence in the corpus is assigned a domain tag that identifies the general thematic area or situational context to which the content belongs. Domain tags help distinguish sentences based on topic rather than communicative function, allowing the corpus to capture variation in vocabulary, register, and usage across different subject areas.

The domain categories were defined to be broad enough to accommodate cross-dialectal variation while remaining specific enough to differentiate common thematic areas. Each seed sentence receives one primary domain tag, selected based on its lexical content and contextual cues. The annotation scheme is applied uniformly across all dialects to ensure comparability and consistency throughout the corpus.

4.4 Raw Corpus

After completing all stages of data collection and processing, dialectal expansion through zero shot prompting, and LLM based pragmatic and domain annotation, we obtain a raw corpus, which constitutes the first version of the dataset. At this stage, the corpus contains all sentences in their 19 dialectal variants, each accompanied by

[1] https://es.wikipedia.org/

[2] https://www.gutenberg.org/

pragmatic and domain tags, but it has not yet undergone manual verification, normalization, or quality-control filtering. This initial version serves as the foundation for subsequent cleaning, consistency checks, and refinement steps that will lead to the finalized release of the corpus.

Figure 3 shows an example from the raw corpus where the automatic dialect generation produced an error in the Paraguayan variant. Instead of generating Paraguayan Spanish, the model translated the sentence into Guaraní.


```
{
  "Neutral": "¿Cómo estás?",
    "Pragmatic_tags": [
    "Informal",
    "Greetings"
  ],
   "Domain": "Social",
  "Caribbean_dialects": {
    "Dominican": "¿Cómo tú 'tá?",
    "Puerto_rican": "¿Cómo tú 'tas?",
    "Cuban": "¿Qué bolá?"
  },
  "Mexican_dialect": {
    "Mexican": "¿Qué onda?"
  },
  "Rioplatense_dialects": {
    "Argentinian": "Che, ¿cómo andás?",
    "Uruguayan": "Bo, ¿cómo andás?"
  },
  "Central_american_dialects": {
    "Guatemalan": "¿Qué tal vos?",
    "Honduran": "¿Qué pedo vos?",
    "Salvadorian": "¿Qué ondas?",
    "Nicaraguan": "¿Qué tuanis?",
    "Costa_rican": "¿Pura vida, mae?",
    "Panamanian": "¿Qué sopa?"
  },
  "south_american_dialects": {
    "Venezuelan": "¿Qué más pues?",
    "Peruvian": "¿Qué tal causa?",
    "Colombian": "¿Quiubo?",
    "Ecuadorian": "¿Qué ñaño?",
    "Bolivian": "¿Qué tal, hermano?",
    "Chilean": "¿Cómo estai?",
    "Paraguayan":"¿Mba' éichapa?"
  }
}
```


Figure 3: Raw corpus showing a mistranslation in the case of Paraguayan Spanish

Paraguay is the only Spanish-speaking country in Latin America in which both Spanish and Guaraní share full official status. Guaraní is spoken by the vast majority of the population, while Spanish dominates formal domains such as government, education, and the media. This coexistence produces widespread code-switching and the mixed variety jopará, and it also creates a distinctive challenge for LLMs, which may misinterpret requests for Paraguayan Spanish by defaulting to Guaraní or blending the two languages.

This example illustrates a typical issue found in the uncorrected corpus and highlights why subsequent cleaning and correction steps are necessary to ensure that each dialectal variant remains in Spanish and accurately reflects the intended dialect.

4.5 Linguistic Integrity Assessment

Establishing the linguistic integrity of the corpus was a central component of the construction process, given that the dataset integrates both real and automatically generated text across nineteen Spanish dialects. Linguistic integrity refers to the degree to which each sentence is structurally well formed, semantically coherent, pragmatically appropriate, and aligned with its assigned domain. Because the corpus is intended to support research on dialectal variation, pragmatic behavior, and domain-specific language use, it was essential to implement a systematic validation procedure capable of identifying and correcting irregularities introduced during generation, transformation, or correction. This assessment was designed to preserve the linguistic richness of each dialect while ensuring that the final dataset meets the standards required for reliable computational and linguistic analysis.

To achieve this, we developed a multi-layered validation workflow that examined the corpus at four complementary levels: syntactic structure, semantic coherence, pragmatic function, and domain relevance. Each level targets a distinct aspect of linguistic well-formedness, allowing for a comprehensive evaluation of the corpus beyond surface-level correctness. The assessment combined automated detection methods with targeted human inspection and controlled model-assisted correction, following a minimal-edit principle to maintain dialectal authenticity. By addressing linguistic integrity across these interconnected dimensions, the resulting corpus achieves a balance between structural accuracy, communicative plausibility, and functional alignment, providing a robust foundation for downstream tasks in natural language processing and corpus-based linguistic research.

To operationalize this framework, we organized the assessment into four complementary components, each targeting a distinct dimension

of linguistic well-formedness. The first component examines the structural properties of the text, verifying that sentences are syntactically complete and free of artifacts introduced during generation or transformation. The second focuses on semantic coherence, evaluating whether sentences express plausible and internally consistent meanings. The third addresses pragmatic integrity, assessing the appropriateness and functional role of discourse markers, politeness strategies, and other communicative cues across dialects. Finally, the fourth component evaluates domain integrity, confirming that each sentence meaningfully aligns with its assigned topical category. Together, these layers provide a comprehensive evaluation of the corpus and establish the foundation for the detailed analyses presented in the following subsections.

4.5.1 Syntactic Integrity Assessment

The syntactic integrity assessment focused on verifying that all sentences in the corpus were structurally well formed and free of artifacts introduced during automatic generation or dialectal transformation. Because the dataset integrates both real and synthetic text across nineteen Spanish dialects, syntactic irregularities can arise from incomplete generation, malformed constructions, or unintended distortions introduced during post-processing. This component of the validation pipeline aimed to identify and correct such issues while preserving the dialectal features and communicative intent of each sentence.

The assessment began with an automated detection phase designed to identify systematic structural anomalies. Custom scripts scanned the corpus for duplicated tokens, truncated sentences, irregular punctuation, spacing inconsistencies, and other artifacts commonly associated with large-scale text generation. These automated checks provided an efficient first pass that established a consistent baseline of structural well-formedness across the dataset. The automated corrections were intentionally conservative, avoiding interventions that might inadvertently alter dialect-specific constructions or pragmatic cues.

However, many syntactic phenomena relevant to Spanish dialectology require linguistic judgment to evaluate accurately. For this reason, sentences flagged as ambiguous, structurally complex, or dialectally marked were subjected to targeted human inspection. This review focused on constructions known to vary across dialects, such as clitic placement, subject-verb agreement patterns, voseo conjugations and regionally variable word order. The goal of this inspection was not to standardize dialectal variation but to correct only those issues that compromised syntactic coherence. A minimal-edit principle guided this process: adjustments were made only when necessary to restore structural integrity without altering the dialectal identity of the texestablished a consistent baseline of structural well-formedness across the dataset. The automated corrections were intentionally conservative, avoiding interventions that might inadvertently alter dialect-specific constructions or pragmatic cues.

In cases where a sentence required clarification or restructuring beyond minor edits, a controlled model-assisted correction procedure was applied. Prompts were designed to produce minimally edited versions that preserved the original dialectal features These model-generated corrections were subsequently manually reviewed and corrected, in some cases,to ensure that they did not introduce cross-dialect contamination, stylistic normalization, or unintended shifts in meaning.

While syntactic validation establishes the structural well-formedness of the corpus, structural correctness alone does not guarantee that sentences convey coherent or plausible meanings. For this reason, the next stage of the assessment focuses on semantic integrity, examining whether each sentence expresses a consistent and contextually meaningful proposition.

4.5.2 Semantic Integrity Assessment

The semantic integrity assessment focused on ensuring that each sentence conveyed a coherent, contextually plausible meaning and maintained internal logical consistency. Because the corpus includes both real and automatically generated text, semantic irregularities can arise from incomplete generation, unintended shifts in meaning, or inconsistencies introduced during transformation or correction. To address these issues, we conducted a systematic review of semantic coherence across all dialects.

Sentences were examined for logical completeness, ensuring that propositions were fully expressed and did not contain abrupt interruptions or unresolved references. We also evaluated internal consistency, verifying that each sentence maintained a stable perspective, temporal frame, and thematic focus. Particular attention was given to cases where automatic generation produced contradictory statements, improbable scenarios, or semantically incompatible combinations of predicates and arguments. These issues were corrected through minimal edits designed to restore coherence while preserving the original dialectal features and communicative intent.

Semantic plausibility was another key dimension of this assessment. We inspected sentences for realistic event structures and naturalistic combinations of actions, participants, and

settings. When semantic anomalies were identified sentences modified without altering their dialectal character. In cases requiring clarification or slight restructuring, controlled model-assisted corrections were applied using prompts designed to maintain meaning while preserving dialectal features. All outputs were subsequently inspected to ensure that no unintended semantic shifts were introduced.

Semantic coherence provides a foundation for meaningful interpretation, but naturalistic language use also depends on the appropriate deployment of pragmatic cues. Building on the semantic evaluation, the next component examines the pragmatic integrity of the corpus.

4.5.3 Pragmatic Integrity Assessment

The pragmatic integrity assessment examined whether the communicative functions encoded in the corpus were expressed appropriately and consistently across dialects. Pragmatic features such as discourse markers, politeness strategies, hedging devices, intensifiers, and stance expressions play a central role in shaping interpersonal meaning and vary substantially across Spanish varieties. Ensuring their correct use was therefore essential for maintaining the authenticity and functional richness of the corpus.

We inspected each sentence for the appropriate deployment of pragmatic elements relative to its communicative intent. This included evaluating whether discourse markers fulfilled their intended roles in structuring information, signaling transitions, or managing interpersonal alignment. The use of dialect-specific markers was examined for both functional relevance and dialectal appropriateness. The use of the pronoun 'you' that can be used as tú, usted, and vos depending of the dialect and politeness level was reviewed to ensure that it aligned with the expected register and social stance encoded in the text.

When pragmatic inconsistencies were identified sentences were minimally adjusted to restore pragmatic coherence while preserving dialectal authenticity. Controlled model-assisted corrections were applied only when necessary, followed by human inspection to ensure that pragmatic intent and dialectal features were maintained.

Although pragmatic features contribute to the communicative richness of each sentence, the corpus also relies on domain labels that organize texts according to topical content. The final component of the assessment therefore evaluates domain integrity.

4.5.4 Domain Integrity Assessment

We reviewed synthetic sentences to see whether they described situations, actions, or concepts that made sense for the domain. When a sentence was too vague, too general, or clearly off-topic, we made small edits to bring it back into alignment. If a sentence needed more substantial adjustment, we used controlled model-assisted correction to produce a clearer and more domain-appropriate version, and then checked it manually to ensure that the correction did not alter the dialect or pragmatic features of the text.

The Wikipedia portion of the corpus benefited from the platform's existing folksonomy, which provides a rich and reliable set of community generated categories. These categories offered a natural way to identify the topical domain of each article without requiring additional annotation or manual classification. Because Wikipedia's tagging system is collaboratively maintained and continuously refined, it provided a stable and well structured source of domain information. As a result, the sentences extracted from Wikipedia showed strong alignment with their assigned domains, and only minimal adjustments were needed during the domain integrity assessment. This made the Wikipedia data a particularly robust and straightforward component of the corpus.

4.6 Corpus

The corpus represents the fully and cleaned version of the dataset produced after all stages of data collection, dialectal generation, pragmatic and domain annotation, linguistic correction, and quality-control procedures. It contains parallel sentences for 19 Spanish dialects, including one neutral Peninsular baseline and one dialectal variant for each American country. Each sentence is accompanied by pragmatic labels, domain tags. The The final corpus comprises 40,000 sentences distributed evenly across 20 dialectal varieties, yielding 2,000 unique sentences per dialect.

5. Conclusion

This paper presented the construction of a multi dialect, pragmatically annotated Spanish corpus that integrates data from Wikipedia, Project Gutenberg, and controlled synthetic generation. By combining dialectal expansion, LLM-based pragmatic and domain annotation, and a hybrid correction workflow involving automated scripts, manual review, and controlled LLM prompting, we produced a high quality dataset covering 19 Spanish varieties. The final corpus preserves dialectal authenticity while providing consistent pragmatic and domain labels, syntactic well formedness, and standardized metadata. This resource contributes a scalable and reproducible methodology for building richly annotated dialect corpora and offers a foundation for research in dialectology, pragmatics, corpus linguistics, and NLP. An important avenue for future work is the construction of a corresponding speech corpus,

which would provide complementary evidence on dialectal pronunciation patterns.

6. Bibliographical References

Ataman, D., Birch, A., Habash, N., Federico, M., Koehn, P., & Cho, K. (2025). Machine translation in the era of large language models: a survey of historical and emerging problems. Information, 16(9), 723.

Abdelaziz, A. A. A., Elneima, A. H., & Darwish, K. (2024, May). LLM-based MT data creation: Dialectal to MSA translation shared task. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024* (pp. 112-116).

Abe, K., Matsubayashi, Y., Okazaki, N., & Inui, K. (2018). Multi-dialect neural machine translation and dialectometry. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Carcelén-Guerrero, A., Posio, P., Kachel, S., & Uclés-Ramada, G. (2025). CoLaGe: Corpus for the Study of Language and Gender in two varieties of spoken Spanish. Corpora, 20(2), 269-285.

Fuchs, M., & González, P. (2022). Perfect-Perfective variation across Spanish dialects: a parallel-corpus study. Languages, 7(3), 166.

Hernández-Mena, C. D., Meza-Ruiz, I. V., & Herrera-Camacho, J. A. (2017). Automatic speech recognizers for Mexican Spanish and its open resources. *Journal of applied research and technology*, *15*(3), 259-270.

Kawasaki, Y. (2026). Digital Linguistic Bias in Spanish: Evidence from Lexical Variation in LLMs. *arXiv preprint arXiv:2602.09346*.

Lu, H., Cheng, G., Luo, L., Zhang, L., Qian, Y., & Zhang, P. (2025, April). Slide: Integrating speech language model with llm for spontaneous spoken dialogue generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

Ma, B., Li, Y., Zhou, W., Gong, Z., Liu, Y. J., Jasinskaja, K., ... & Plank, B. (2025, July). Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8679-8696).

McCallum, L., & Mizumoto, A. (2025). Using Llms for Corpus Linguistics Research: Promise, Disappointment and Progression. *Disappointment and Progression*. Available at SSRN: https://ssrn.com/abstract=5224441 or http://dx.doi.org/10.2139/ssrn.5224441

Molina Martos, Isabel. (2024). Spanish Dialect Classifications. Dialectología. 12. 309-342. 10.1344/Dialectologia2024.2024.10.

Morin, C., & Larsson, M. M. (2025). A large-scale, unsupervised pipeline for automatic corpus annotation using LLMs: variation and change in the English consider construction. *arXiv preprint arXiv:2510.12306*.

Pozhemetskiy, R. (2025). *Translation into a low-resource language system: the ability of ChatGPT to create texts in Andalusian dialect of Spanish language* (Master's thesis, Itä-Suomen yliopisto).

Quesada Pacheco, M. Á. (2021). Dialectología histórica del español de América Central. Nivel fonético-fonológico. *Revista de Historia de la Lengua Española*, (16), 67-100.

Sakai, Y., Makinae, M., Kamigaito, H., & Watanabe, T. (2024, November). Simultaneous interpretation corpus construction by large language models in distant language pair. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 22375-22398).

Schlumpf, S., & Carreira, S. (2024). Presentación de un corpus para el estudio del español actual.

Shen, J., Tenenholtz, N., Hall, J. B., Alvarez-Melis, D., & Fusi, N. (2024). Tag-LLM: Repurposing general-purpose LLMs for specialized domains. *arXiv preprint arXiv:2402.05140*.

Weissweiler, L., Köksal, A., & Schütze, H. (2025). Hybrid Human-LLM Corpus Construction and LLM Evaluation for the Caused-Motion Construction. *Northern European Journal of Language Technology*, *11*(1), 27-57.

Yakhni, S., & Chehab, A. (2025). Fine-tuning llms for low-resource dialect translation: The case of lebanese. *arXiv preprint arXiv:2505.00114*.