

Bridging Ibero and Latin American NLP communities

Eugenio Martínez, Luis Chiruzzo, Pablo Gamallo, Rafael Muñoz, German Rigau

CEATIC (UJA), UdelaR, CiTIUS (USC), CENID (UA), HiTZ (UPV/EHU)
emcamara@ujaen.es, luischir@fing.edu.uy, pablo.gamallo@usc.gal,
rafael@dlsi.ua.es, gernan.rigau@ehu.eus

Abstract

LANLP focuses on community-driven resource development and evaluation for Iberian languages, and diverse Latin American languages (including indigenous and minority languages). We aim to bridge regional communities to share initiatives, corpora and tools. LANLP fills this gap, fostering new contacts between Iberian and Latin American NLP research groups. The goals are to (1) highlight challenges in processing these languages, (2) share novel datasets and models, and (3) catalyze future collaborations and shared tasks. We emphasize both academic rigor and community inclusivity, encouraging contributions from established researchers and grassroots language advocates alike.

Keywords: Networking, NLP communities, digital divide

1. Introduction

We organize a Networking Symposium on Latin American NLP (LANLP), focusing on natural language processing for the diverse languages of the Iberian Peninsula and Latin America. This region includes major world languages (e.g., Spanish (~558M speakers), Portuguese (~267M)), as well as regional and indigenous languages. For example, Latin America alone hosts tens of millions of speakers of Quechua (~10M), Guaraní (>6M), Nahuatl (~2M), and Aymara (~2M), among many others. Such languages are highly under-resourced: over 88% of the world's languages remain largely unsupported by language technologies.

This networking event addresses that gap by promoting collaboration on ethically and culturally sensitive resource creation, evaluation, and novel methods for low-resource multilingual NLP in Iberian and Latin American languages and varieties. Our goal is to bring together communities (SEPLN, PROPOR, AmericasNLP, and CLARIAH-ES) to share cutting-edge research, language resources, and best practices.

LANLP focuses on community-driven resource development and evaluation for Iberian languages, and diverse Latin American languages (including indigenous and minority languages). We aim to bridge regional communities: for instance, past forums such as OpenCor¹ note that “Latin American and Iberian communities... did not have an established event” to share initiatives, corpora, and tools. LANLP tries to fill this gap, fostering new contacts between Iberian and Latin American NLP research groups.

The goals are to:

1. Highlight challenges in processing these languages;
2. Share novel datasets and models;
3. Catalyze future collaborations and shared tasks.

We emphasize both academic rigor and community inclusivity, encouraging contributions from established researchers and grassroots language advocates alike.

2. LANLP Communities

2.1. SEPLN & IberLEF

The Spanish Society for Natural Language Processing (SEPLN), founded in 1983, has played a central role in structuring the scientific and technological community for Natural Language Processing (NLP) in Spain. Its contribution goes beyond the organization of an annual conference: SEPLN has progressively consolidated a stable ecosystem for research, evaluation, dissemination, technology transfer, and collaboration among universities, research centers, public institutions, and companies. This long-term community-building role has been essential for strengthening both scientific excellence and the practical impact of language technologies in the Iberian context.

Within this ecosystem, the journal *Procesamiento del Lenguaje Natural* (PLN Journal) is a core instrument for scientific dissemination. The journal, published biannually, has become a reference venue for research in NLP in Spanish and related contexts. Its relevance is not only historical, but also institutional, as reflected in recognized quality indicators. In particular, the journal has obtained the FECYT quality seal and is indexed in ESCI (Emerging Sources Citation Index, Web of Science).

¹<https://opencor.gitlab.io/>

2025:Q2 Linguistics, Q4 Computer Science. Artificial Intelligence). In addition, SEPLN highlights its visibility through bibliometric indicators and its role as a trusted publication channel for the community. These quality markers reinforce the journal's strategic value for the consolidation and international visibility of NLP research.

SEPLN has also strengthened the technology transfer and industry engagement dimension of the field through initiatives such as TecnoLing. Integrated into the SEPLN conference framework, TecnoLing is designed as a technology showcase and networking space that brings together companies, startups, R&D centers, and public-sector actors. Its format typically includes demonstrations, exhibitor stands, presentations, and networking sessions, explicitly aimed at facilitating knowledge transfer, identifying collaboration opportunities, and promoting the adoption of language technologies in real-world settings. TecnoLing is therefore especially important because it helps transform the conference environment into a broader innovation ecosystem, where academic advances can more easily connect with industrial needs and societal applications.

In the area of shared evaluation, the trajectory that has led to IberLEF (Iberian Languages Evaluation Forum) represents one of the most significant contributions of the SEPLN community. IberLEF emerged from earlier evaluation campaigns such as TASS and IberEval, and has evolved into a consolidated forum for the comparative evaluation of NLP systems developed by research groups and companies. Its value lies in promoting reproducible evaluation, shared datasets and benchmarks, and coordinated progress on challenges relevant to Iberian languages.

IberLEF is better understood through its scale and thematic diversity. Recent editions have included a large number of shared tasks and broad international participation, with contributions spanning a wide range of topics such as sentiment analysis, harmful content detection, inclusive language, simplification and easy-to-read text, multimodal analysis (e.g., memes), historical text processing, question answering, clickbait detection, and Portuguese-focused tasks, among others. This diversity reflects the maturity of the forum and its capacity to support both foundational and applied research across multiple languages and domains.

Taken together, SEPLN, the PLN Journal, IberLEF, and TecnoLing constitute a highly valuable infrastructure for NLP in Spain and the broader Iberian research space. They combine scientific publication, editorial quality standards, shared evaluation practices, and technology transfer mechanisms in a coherent ecosystem. This integrated structure is particularly relevant for fostering sustainable collaboration, improving research visibility,

and ensuring that advances in NLP and AI can generate tangible scientific, economic, and social impact.

2.2. PROPOR

PROPOR refers both to a long-standing research community dedicated to the computational processing of Portuguese and related varieties, and to the biennial International Conference on the Computational Processing of Portuguese, which has served as its main scientific meeting since 1993. Over more than three decades, the PROPOR community and conference have provided a central forum for research on written and spoken Portuguese, fostering methodological advances, language resource development, and collaboration between academia and industry. The conference has alternated between Brazil and Portugal, reflecting the transatlantic nature of the Portuguese-speaking world, and has recently expanded its scope to more explicitly include Galician, as evidenced by the 2024 edition hosted in Santiago de Compostela, marking a significant step toward a recognition of Galician as a variety of the Lusophone diasystem. Both the community and the conference are managed by a Steering Committee that is renewed every two years and includes Portuguese, Brazilian, and Galician researchers.²

While PROPOR has successfully consolidated a transatlantic research community around Portuguese (and more recently Galician) technology, it has historically evolved in parallel with other Iberian and Latin American NLP communities, particularly those focused on Spanish and indigenous languages. By alternating its conference venues between Brazil, Portugal and Galicia, PROPOR has encouraged sustained collaboration between research groups on all continents, promoting the sharing of methodologies, language resources, evaluation frameworks, and technological developments. The conference and its associated community provide a stable organizational structure and a long-standing forum for scientific exchange, supporting both academic research and industrial applications. This transatlantic dimension positions PROPOR as a key hub for multilingual research, enabling studies on many language varieties of Portuguese across several continents, including not only America and Europe, but also Africa, Asia and Oceania. This contributes to the development of inclusive and interoperable language technologies for the Portuguese linguistic space.

²<https://propor.org>

2.3. AmericasNLP

AmericasNLP³ is a workshop that has been running since 2021 with the objective of bringing together researchers of NLP and computational linguistics applied to indigenous languages throughout the Americas. The aim is to encourage and increase the visibility of work on indigenous languages of the Americas, by encouraging research on NLP, computational linguistics, corpus linguistics and speech for indigenous languages, to connect researchers and professionals from underrepresented communities and native speakers of endangered languages, and more generally to promote machine learning approaches suitable for low-resource languages.

So far there have been five editions of the workshop, with an upcoming sixth edition on the way, co-located either with one of the *CL conferences or with NeurIPS. In total, more than 100 papers have been published in the proceedings of the workshop, including system papers of the participants of the eight shared tasks organized during this time, which include: Machine Translation for low-resource indigenous languages, creation of educational materials, Automatic Speech Recognition of indigenous languages, and MT metrics for indigenous languages.

2.4. CLARIAH-ES

In the current context of digital transformation, study, research, and development in the humanities, arts and social sciences require scientific and technological infrastructures that allow for the computational processing of textual, visual, numerical and/or audio data. The CLARIAH-ES⁴ (F. et al., 2024) infrastructure supports and contributes to the management and coordination in Spain of the European Research Infrastructure Consortia (ERIC) CLARIN⁵ (focused on digital data and processes related to Language) and DARIAH⁶ (focused on digital data and processes related to arts and humanities scholars). Although both CLARIN and DARIAH are independent ERIC infrastructures, some European countries have formed joint CLARIAH consortia. Both infrastructures promote multilingualism, digital methods, interoperability, maintenance and reuse of resources, open science, visibility and scientific cooperation in Europe, thus overcoming the fragmentation of research communities and increasing the impact of their research.

CLARIAH-ES contributes to the advancement of research in the humanities and social sciences, as well as to its strategic positioning in interna-

tional projects and programs. With this in mind, CLARIAH-ES seeks to bring together research groups and initiatives that have a stake in these research infrastructures and that wish to reduce the digital divide, promoting new multidisciplinary lines of research in the humanities, arts and social sciences (and beyond) by facilitating their digital transformation with the help of language technologies.

The ambitious goals of CLARIAH-ES can only be achieved by bringing together the necessary resources in terms of data, computing facilities, and knowledge that are not available to any one research group in Spain. CLARIAH-ES is formed by a multidisciplinary group of twelve leading research centers in Language Technologies (TL), Artificial Intelligence (AI), High Performance Computing (HPC), linguistic experts in the official languages in Spain (Spanish, Catalan, Basque and Galician), and experts in digital transition in the areas of humanities, arts and social sciences.

In summary, the CLARIAH-ES research infrastructure seeks a tangible impact on society. Through greater exchange of knowledge, data, technologies, infrastructures, skills, and best practices, we aim to amplify the potential of our current research capabilities. This collaborative synergy will not only ensure the sustainability of tools and services, but also foster collaborative environments. Our aspirations transcend borders, encompassing Europe, Ibero-America, and the global stage as we endeavor to increase funding opportunities for vital infrastructures. The interdisciplinarity of the participating research groups, which includes areas of research as diverse as computer science, philology, social sciences, history, etc., ensures a broad contribution from different perspectives. By properly weaving together these perspectives, we can develop research results that are useful for promoting high-impact digital tools and artificial intelligence applications in different social science and digital humanities scenarios, including research and cultural infrastructures such as libraries and museums.

3. Related initiatives

3.1. ILENIA and ALIA

To understand the current trajectory of Ibero-American NLP, it is essential to examine large-scale institutional projects that provide the backbone for digital sovereignty. Two of the most significant initiatives in the Iberian context are **ILENIA** and **ALIA**, which focus on the structural development of language resources and foundational AI models for Spanish and the co-official languages of Spain.

The **ILENIA** project⁷ is a cornerstone initiative

³<https://americasnlp.org/>

⁴<http://www.chariah.es>

⁵<http://www.clarin.eu>

⁶<http://www.dariah.eu>

⁷<https://proyectoilenia.es/>

designed to promote the digital presence and technological parity of Spain’s linguistic diversity. By coordinating the efforts of leading research centers—including the BSC-CNS, HITZ Center (UPV/EHU), CiTIUS and ILG (USC), and CENID (UA)—ILENIA focuses on creating high-quality, interoperable Language Technology infrastructure. The main objectives are the following:

- **Multilingual Infrastructure:** The project facilitates the creation of advanced corpora, lexicons, and tools for Spanish, Catalan, Basque, and Galician.
- **Scientific Cooperation:** It aligns with the goals of **CLARIAH-ES** to reduce the digital divide and promote open science across the humanities and social sciences.
- **Technological Integration:** ILENIA ensures that resources follow common standards, enabling their reuse in a wide variety of AI applications, from machine translation to speech processing.

The **ALIA** project⁸ (*Alianza por la Inteligencia Artificial en Español*) represents a strategic push to develop a large-scale foundation model that is culturally and linguistically grounded in the Hispanic world. The main goals are the following:

- **Digital Sovereignty:** It aims to provide a “technological public good,” offering an open-source alternative that allows institutions and startups to build domain-specific applications without dependency on closed, proprietary APIs.
- **Mitigating Bias:** ALIA seeks to mitigate “algorithmic colonialism” by training models on regional legal, medical, and academic datasets rather than relying solely on translated English benchmarks.
- **Community Synergy:** ALIA benefits from the collaborative ecosystem, utilizing community-driven datasets and shared evaluation suites such as “IberoBench” (Baucells et al., 2025).

In the context of the LANLP Networking Symposium, these projects serve as successful blueprints for how institutional funding and academic expertise can be synthesized to achieve technological autonomy. The symposium aims to catalyze similar large-scale, transatlantic collaborations that ensure Ibero-American languages are central to the next generation of AI.

⁸<http://alia.gob.es>

3.2. SomosNLP

SomosNLP⁹ is a community of academic researchers, industry practitioners, and open-source contributors dedicated to creating and sharing resources that enable and accelerate the development of NLP in Ibero-American languages. The aim is to address linguistic inequity by connecting individuals from the region, collaboratively exploring unique challenges, and building the necessary open-access research infrastructure to enable rapid and relevant technological progress across the Spanish and Portuguese-speaking world. The community has been active since 2021 and was registered as a non-profit organization in 2025.

The SomosNLP initiative is built upon three core pillars: to promote open-source resource creation, to educate by providing access to high-quality content, and to connect for multidisciplinary collaboration.

The annual SomosNLP hackathon is the foundational mechanism for engagement, typically attracting over 300 registrations. These events facilitate resource creation collaboratively. The hackathons are framed as open innovation spaces where participants collectively identify, build, and contribute to open-source resources. The hackathons are open to all NLP practitioners—from industry experts to independent developers and academics—welcoming anyone regardless of their background or expertise. This collaborative model has generated significant and verifiable technical output on Hugging Face. The hackathon is supported by established companies: Hugging Face has been the gold sponsor since the first edition in 2022, and for the 2025 event, the hackathon secured sponsorship from Mistral and was recognised as a Cohere community grantee.

The education strategy focuses on making high-quality resources universally accessible. SomosNLP hosts a curated library of 70+ recorded talks and workshops on the public YouTube channel. The knowledge shared in the talks and workshops explicitly encourages open source code development, data sharing, and open methods/protocols—the essential building blocks for ethical, reproducible, and responsible NLP research.

To connect and bridge the gap between researchers across the Ibero-American community and the global NLP community, SomosNLP enables spaces for targeted networking. The SomosNLP Discord hosts over 2000 members who share resources, projects, events, relevant news, and support each other. Moreover, in 2025 SomosNLP started hosting Birds-of-a-Feather (BoF) events at major international conferences: ACL, COLM, and NeurIPS.

⁹<https://somosnlp.org>

A successful example of international cross-institutional open-source collaboration led by SomosNLP is “La Leaderboard” (Grandury et al., 2025), a LLM leaderboard for Spanish varieties and languages of Spain and Latin America. This leaderboard hosts 66 benchmarks in Spanish, Catalan, Basque, and Galician, donated by 13 research groups. The paper presenting La Leaderboard was accepted at ACL Main 2025, and the web interface receives around 3.000 monthly visits.¹⁰

3.3. LATAM-GPT: A Foundation for Regional Digital Sovereignty

The emergence of **LATAM-GPT** represents a pivotal shift in the Ibero-American AI landscape, moving the region from passive consumption of Northern-centric models to the active development of localized foundation models. Led by Chile’s National Center for Artificial Intelligence (CENIA) and supported by a coalition of over 60 institutions across 15 countries, LATAM-GPT is designed to address the “linguistic and cultural representation gap” inherent in global Large Language Models (LLMs).

LATAM-GPT is built upon the `Llama 3.1` architecture and operates at a scale of approximately 50 billion parameters, positioning it within the same performance tier as GPT-3.5. Unlike proprietary models where training data remains opaque, LATAM-GPT’s development prioritizes regional specificity through a curated corpus:

- **Data Volume:** A massive 18 – 20.5 TB dataset comprising over 1 billion documents.
- **Linguistic Diversity:** The model integrates high-resource languages (Spanish and Portuguese) alongside 50+ indigenous languages such as Quechua, Mapuche, and Guaraní.
- **Infrastructure:** Initial training was conducted using high-performance computing clusters, including nodes equipped with NVIDIA H200 GPUs, facilitated by partnerships with AWS and the Data Observatory.

The primary objective of the LATAM-GPT project is to mitigate *algorithmic colonialism*—the tendency of global models to project Anglo-Saxon cultural norms and linguistic structures onto non-English speakers. By training on regional legal, medical, and academic datasets, LATAM-GPT provides:

1. **Semantic Precision:** Nuanced understanding of regional slang, idioms, and local administrative terminology.

¹⁰<https://huggingface.co/spaces/la-leaderboard/la-leaderboard>

2. **Inclusion:** Reduced tokenization fragmentation for Spanish and Portuguese, which lowers computational costs and improves response quality for regional users.
3. **Technological Autonomy:** As a “technological public good,” it offers an open-source alternative for local governments and startups to build domain-specific applications without dependency on closed APIs.

3.4. LatinX in Natural Language Processing (LXNLP)

The **LatinX in Natural Language Processing (LXNLP)** workshop,¹¹ operating under the institutional umbrella of LatinX in AI (LXAI), serves as a primary affinity group dedicated to enhancing the visibility and professional advancement of researchers from Ibero-American and Latin American origins. This initiative addresses the systemic under-representation of LatinX scholars in core NLP venues, where historical participation from the Global South has faced significant structural and financial barriers.

3.5. Overview of the Accepted Contributions

The accepted papers offer a coherent snapshot of current research priorities at the intersection of Ibero- and Latin American NLP, while also reflecting the methodological and sociotechnical diversity of the communities involved. Taken together, they show that work in the region is not limited to adapting mainstream NLP pipelines to new languages, but is increasingly concerned with building language technologies that are community-aware, resource-conscious, multilingual, and deployable in socially or institutionally sensitive settings.

A first thematic cluster focuses on low-resource, Indigenous, and underrepresented languages, highlighting the need for approaches that move beyond simple technological transfer from high-resource settings. In this line, the oral-first interactive system for Guaraní speakers foregrounds conversational agency and culturally aligned interaction design in primarily oral language contexts, while AI-TraLow addresses machine translation for low-resource languages through a combination of curated data, linguistic resources, and efficient modeling. OpenCor complements these contributions by documenting the collective and infrastructural dimension of language technology, emphasizing the importance of open corpora, lexical resources, and the sustainability challenges behind community-driven resource creation.

¹¹<https://www.latinxinai.org/naacl-2024>

A second cluster centers on trustworthy and domain-sensitive language technologies, especially in regulated or high-stakes environments. SAFEWORDS proposes a reproducible framework for GDPR-aligned anonymization and evaluation across Spain's co-official languages, placing privacy, replicability, and governance at the core of LLM assessment. In parallel, mCS-LM and MedicaLLM explore applied multilingual and multimodal systems for customer service, incident management, and healthcare, respectively. Both contributions stress the importance of grounded generation, structured outputs, validation layers, and interoperability when deploying language technologies in real-world organizational contexts.

The program also includes work on foundational language modeling and linguistic representation, as illustrated by the study on sentence representations in Spanish BERT-like models. By analyzing how different layers encode syntactic and semantic information, this contribution provides a more fine-grained understanding of representational behavior in Spanish-language models, offering insights that are relevant not only for benchmarking, but also for downstream system design and evaluation.

Overall, the accepted papers reveal three broader tendencies. First, they confirm that resource creation, evaluation, and deployment must be treated as interconnected challenges, especially for languages and varieties that remain underrepresented in mainstream NLP. Second, they highlight a growing emphasis on responsible and context-aware AI, including privacy, accountability, interpretability, and cultural alignment. Third, they show that the field is advancing simultaneously at multiple levels: from infrastructures and corpora, to model analysis, to translation, speech, and domain-specific applications. In this sense, the program reflects a research agenda in which linguistic diversity, technical robustness, and societal relevance are increasingly inseparable.

4. Expected Outcomes

LANLP Networking Symposium at LREC 2026 is designed to be a catalyst for long-term community engagement. We anticipate several key outcomes categorized into academic, technical, and socio-economic domains.

4.1. Scientific and Technological Autonomy and Policy Influence

The symposium aims to provide the scientific backing needed for digital sovereignty.

We aim at generating a policy brief for regional science and technology funding agencies regarding the strategic importance of developing sovereign

AI infrastructure to avoid over-reliance on external proprietary APIs.

4.2. Formalization of a Regional Research Roadmap

A primary expected outcome is the publication of a *LANLP 2026 White Paper*. This document will synthesize the findings from the panel discussions and breakout sessions to establish a "Community Priority List." This roadmap can identify:

- **Linguistic Priorities:** Identifying specific dialects and indigenous languages in urgent need of digitization and resource creation.
- **Benchmarking Standards:** Establishing standardized evaluation protocols for Ibero-American NLP tasks, ensuring that regional models are measured against culturally relevant metrics rather than just translated English benchmarks.
- **Sustainability:** Establishing a permanent digital directory of Ibero-American NLP labs to facilitate easier cross-border internship and PhD exchange programs.

4.3. Consolidation of Open-Source Language Resources

By leveraging the co-location with LREC, we expect to accelerate the release and documentation of several key datasets.

- **Instruction-Tuning Corpora:** We expect to release the next iteration of open-source, instruction-tuned datasets specifically for regional Spanish and Portuguese variations.
- **Interoperability:** Ensuring that new resources from the region are integrated into global repositories such as the ELRA catalogs and the Hugging Face Hub, increasing their visibility to the global scientific community.

5. Towards a common Research Agenda

The LANLP community is unified by a commitment to advancing the state of Natural Language Processing through the lens of Ibero-American linguistic and cultural specificity. Our research agenda is organized around four primary pillars that address the unique challenges of our community.

5.1. Linguistic Resource Infrastructure and Evaluation

At the core of the LANLP mission is the expansion of the digital footprint for regional languages. This includes:

- **Primary Resource Creation:** Development of high-fidelity corpora, lexicons, and multimodal annotations for the Iberian and Latin American languages, with a focus on capturing both text and speech nuances.
- **Morphosyntactic Analysis:** Specialized research into the analysis and tagging of morphologically rich and under-documented languages (e.g., Basque, Mapudungun, Bribri), utilizing frameworks such as Universal Dependencies.
- **Tailored Benchmarking:** Designing evaluation metrics and benchmarks that move beyond translated English datasets to reflect the authentic linguistic realities of our languages and cultures.

5.2. New Developments in the Era of Large Language Models

The community explores the intersection of localized needs and global modeling trends, specifically:

- **Model Efficiency and Equity:** Investigating Small Language Models (SLMs), synthetic data generation, and strategies to mitigate "language domination" and the digital scarcity affecting Ibero-American varieties.
- **Transfer Learning and Multilinguality:** Advancing cross-lingual representations and embedding methods that bridge the gap between high-resource Spanish and Portuguese and the diverse minority languages of the continent.
- **Translation and Generation:** Optimizing Machine Translation (MT) and generation for regional variations and low-resource indigenous languages like Quechua, Aymara, and Nahuatl.

5.3. Dialectology, Contact, and Domain-Specific NLP

Ibero-America provides a unique laboratory for studying language in contact and social context:

- **Dialectal Variation and Code-Switching:** Developing robust methods for identifying and handling regional dialects and complex language contact, such as Spanish-Portuguese code-mixing and Spanish-Indigenous language intersection.

- **Speech and Audio Innovation:** Tailoring ASR and TTS systems to the prosody and phonology of Latin American Spanish, Brazilian Portuguese, and indigenous oral traditions.
- **Applied Socio-NLP:** Tackling domain-specific tasks—including sentiment analysis and hate-speech detection—within the specific socio-political contexts of Latin American digital media.

5.4. Ethics, Governance, and Participatory Methods

Recognizing the deep cultural stakes of our work, the LANLP community prioritizes:

- **Participatory Research:** Implementing community-driven methods, including crowdsourcing and citizen science, to ensure data collection is representative and inclusive.
- **Data Sovereignty and Rights:** Addressing the ethics of indigenous language rights, data governance, and the sustainability of resources within the framework of fair and transparent AI.
- **Digital Humanities:** Applying NLP to historical texts and cultural heritage to preserve and analyze the vast literary legacy of the Ibero-American space.

6. Conclusions

LANLP is positioned as a networking venue explicitly designed to bridge Iberian and Latin American NLP communities, with a strong emphasis on under-resourced and minoritized languages, ethical and culturally sensitive practices, and community-driven resource development and evaluation. The overview of existing ecosystems (e.g., SEPLN, CLARIAH-ES, PROPOR and Americas-NLP) makes clear that substantial expertise, resources, and infrastructures already exist on both sides of the Atlantic, but that they remain only partially connected in terms of shared benchmarks, interoperable resource pipelines, and sustained coordination mechanisms.

A key takeaway is that the most impactful contribution of LANLP is likely to be as a *connective layer* rather than a replacement for established venues: a place where communities align on common practices for resource documentation, evaluation design, and responsible collaboration, while still preserving the identity and strengths of each network. In this sense, evaluation-oriented communities and shared-task initiatives can provide reusable protocols for reproducibility and comparability, whereas

infrastructure-oriented initiatives can ensure long-term sustainability through standards, metadata, and interoperability.

At the same time, the current community map is necessarily incomplete. Several communities and initiatives are still emerging or are not yet fully represented, and the long tail of languages and varieties in Iberia and Latin America cannot be covered by any single event. This highlights an immediate agenda for LANLP: (i) broaden participation to include additional networks, grassroots language advocates, and regionally grounded initiatives; (ii) encourage shared evaluation suites and task designs that explicitly account for dialectal variation and code-switching; and (iii) promote governance models that support ethical data collection, long-term maintenance, and transparent reuse.

Overall, LANLP can catalyze durable cross-regional collaboration by prioritizing three outcomes: interoperable language resources, evaluation and benchmarking practices tailored to low-resource realities, and sustained community links that convert short-term networking into long-term joint work.

Acknowledgments

This work has been partially supported by the Strategic Networks CLARIAH-ES (RED2024-154077-E) and RutaMdL (RED2024-154067-E) and the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU – NextGenerationEU within the framework of the project "Desarrollo de Modelos ALIA". The authors also thank Aritz Farwell and Ainara Estarrona for helping with the organization of the symposium.

References

- Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. [IberoBench: A benchmark for LLM evaluation in Iberian languages](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.
- Carreras F., Estarrona A., Farwell A., Iruskieta M., Marco M., Melero M., Montejo A., Rigau G. Riño D., Romero D., Ros S., Sánchez E., and Sousa X. 2024. [Clariah-es: Strategic network](#)

[for the integration in the european research infrastructures in social sciences and humanities](#). In *SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, CEUR Workshop Proceedings Vol-3729*, La Coruña, Spain. SEPLN.

- María Grandury, Javier Aula-Blasco, Júlia Falcão, Clémentine Fourrier, Miguel González Saiz, Gonzalo Martínez, Gonzalo Santamaria Gomez, Rodrigo Agerri, Nuria Aldama García, Luis Chiruzzo, Javier Conde, Helena Gomez Adorno, Marta Guerrero Nieto, Guido Ivetta, Natàlia López Fuertes, Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, Helena Montoro Zamorano, Carmen Muñoz Sanz, Pedro Reviriego, Leire Rosado Plaza, Alejandro Vaca Serrano, Estrella Vallecillo-Rodríguez, Jorge Vallego, and Irune Zubiaga. 2025. [La leaderboard: A large language model leaderboard for Spanish varieties and languages of Spain and Latin America](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32482–32524, Vienna, Austria. Association for Computational Linguistics.