OpenCor: Latin American and Iberian Languages Open Corpora Forum

Livy Real[1,2], **Valeria de Paiva**[3]
[1] Instituto Kunumi, Belo Horizonte, Brazil
[2] Instituto de Computação – Universidade Federal do Amazonas, Manaus, Brazil
[3] Topos Institute, Berkeley, USA
livy@kunumi.com, valeria@topos.institute

Abstract

The availability of open resources and corpora is a fundamental requirement for research in Natural Language Processing (NLP) and Computational Linguistics; however, languages spoken in Latin America and the Iberian Peninsula, particularly indigenous, minority, and regional varieties, remain structurally under-resourced and under-represented. This paper presents a historical account of OpenCor (Latin American and Iberian Languages Open Corpora Forum), a community-driven initiative created to promote, document, and discuss open corpora and lexical resources for these languages. Conceived as a collaborative forum rather than a competitive evaluation venue, OpenCor focuses on data creation, licensing practices, sustainability, and community building. Between 2018 and 2024, OpenCor was organized as a recurring workshop co-located with major conferences, fostering dialogue across countries, institutions, and linguistic traditions. By documenting the initiative's motivations, organizational trajectory, submission trends, and the diversity of resources presented, this paper aims to preserve institutional memory, highlight the often-invisible labor of corpus development, and provide a reference for future initiatives dedicated to openness and linguistic diversity.

Keywords: Language Resources, Corpora, Latin American Languages, Iberian Languages

1. Introduction

The availability of open corpora and lexical resources is a central requirement for the development of Natural Language Processing (NLP) and Computational Linguistics research. While major languages benefit from a growing ecosystem of openly licensed datasets, tools, and benchmarks, languages spoken in Latin America and the Iberian Peninsula have historically faced structural challenges related to visibility, funding, infrastructure, and coordination. These challenges are particularly acute for under-resourced languages, regional varieties, and languages with limited institutional support.

In this context, **OpenCor** (Latin American and Iberian Languages Open Corpora Forum) emerged as a community-driven initiative aimed at mapping, discussing, and promoting open corpora and lexical resources for these languages. Rather than focusing on system comparison or shared tasks, OpenCor was conceived as a forum for resource presentation, discussion of licensing practices, sustainability, and community building. Between 2018 and 2024, OpenCor was organized as a recurring workshop co-located with major conferences in the field in both Latin America and the Iberian Peninsula, fostering dialogue across countries, institutions, and linguistic traditions.

This paper presents a historical report of the OpenCor initiative, covering its motivations, organizational trajectory, meetings, and the evolution of the resources discussed within the forum. By documenting this history, we aim to preserve institutional memory, highlight community efforts that are often invisible in traditional evaluation venues, and provide a reference for future initiatives focused on openness and linguistic diversity. The project maintains a public web presence at https://opencor.gitlab.io/.

This paper makes four principal contributions. First, it presents a documented historical timeline of OpenCor from 2018 to 2024. Second, it provides an analysis of submission patterns and participation trends across editions. Third, it offers a curated overview of the open corpora showcased through the initiative, highlighting their linguistic, methodological, and regional diversity. Finally, it reflects on the structural challenges involved in sustaining open linguistic infrastructures, situating OpenCor within broader debates about research recognition, funding asymmetries, and long-term maintenance.

2. OpenCor Motivation

The primary motivation behind OpenCor is the absence of a consolidated, openly accessible list of corpora and lexical resources for Latin American and Iberian languages. Although many resources exist, information about them is fragmented across personal webpages, institutional repositories, or conference proceedings. This fragmentation makes it difficult for researchers, students, and practitioners to discover existing datasets, assess their licensing conditions, and reuse them responsibly. The initiative is inspired by

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 21–28

Linguateca[1], established in 1998 as a distributed resource center for the computational processing of Portuguese, with a mission centered on facilitating access to existing resources, collaboratively developing new ones, and organizing shared evaluation efforts (Santos, 2003).

The name *OpenCor* was chosen with a deliberate double meaning. It refers to *Open Corpora*, the central focus of the initiative, but also to *cor*, the Latin word for *heart*. This second meaning serves as a metaphor for openness, generosity, and inclusion. A common saying across the Iberian Peninsula and Latin America describes something as 'as big as a mother's heart, where there is always room for one more'. Historically, many workshops and shared tasks in NLP prioritize system performance and competition, which can inadvertently marginalize discussions about data creation, annotation labor, licensing, and long-term maintenance. OpenCor was designed to be: first, a list of open resources; second, a complementary space, a forum explicitly dedicated to resources rather than models, and to collaboration rather than competition.

Another important motivation was the production of a historical timeline of open resources and community efforts. By tracking meetings, submissions, and presented corpora over time, OpenCor functions as a living record of how openness has evolved in the regions. The OpenCor website was therefore not only an organizational hub but also a public archive documenting editions, programs, calls for papers, and links to resources.

The initial discussions that led to OpenCor took place during a Linguistic Data Consortium (LDC)[2] meeting in Mexico City, sponsored by the University of Pennsylvania, in 2018. At that moment, participants reflected on the structural asymmetry in the production and consolidation of linguistic infrastructures: although many corpora and initiatives concerned Latin American and Iberian languages, centralized documentation and resource aggregation were often located outside the regions where these languages are spoken. This observation reinforced the need for a regionally grounded forum dedicated to mapping, documenting, and sustaining open resources from within the community itself. OpenCor is also motivated by the recognition that many researchers in Latin America and the Iberian Peninsula face significant funding constraints. The workshop consistently aimed to lower barriers to participation by allowing "workshop-only" registration, using free submission platforms, and inviting speakers who were already attending the host conference or able to participate without additional financial support. Finally, OpenCor meetings always allowed for 'non-archival' submissions, aiming to gather the community that had already published their works in another venue but were interested in discussing and sharing their resources.

OpenCor focuses on corpora that are freely accessible and reusable, thereby supporting transparency and cumulative research. It prioritizes resources relevant to the languages spoken in Latin America and the Iberian Peninsula, including indigenous and minority languages that are often underrepresented in mainstream NLP research. At the same time, it seeks to engage a broad scholarly community, encompassing researchers in NLP, linguistics, education, and related fields.

The initiative serves two complementary purposes:

1. As a **curated, evolving list of open corpora**, improving discoverability and reuse of existing resources;
2. As a **forum and meeting point** for corpus creators and users, providing space for discussion of methodologies, annotation practices, and long-term maintenance issues.

3. OpenCor Development

OpenCor has been active since the late 2010s and has been organized as a recurring forum, often co-located with established conferences and workshops in Computational Linguistics and NLP. By 2024, the project had reached its fifth edition, reflecting sustained community interest and ongoing relevance.

OpenCor has evolved from a small, informal effort into a recognized venue for presenting and discussing corpus-related work that often does not fit comfortably within traditional publication formats.

Across all editions, OpenCor was organized as a community-driven workshop with a strong commitment to openness and accessibility. EasyChair was consistently used as the submission and review platform due to its free availability for small meetings. Invited speakers were selected based on their contributions to open resources and their ability to participate without dedicated travel funding.

The first OpenCor workshop[3] was held in 2018 in Canela, Brazil, co-located with PROPOR, the International Conference on Computational Processing of Portuguese. The workshop was organized following an invitation from Prof. Aline Villavicencio and established the core format for subsequent editions. Of the eight accepted papers, three of them (almost half) were not presented due to a lack of funding, a challenge that would recur in later editions. The invited speaker for this first edition was

[1] https://linguateca.pt/

[2] https://www.ldc.upenn.edu/

[3] https://opencor.gitlab.io/opencor-2018/

Year	Event	Location	Format	Submitted	Accepted	Keynote
2018	PROPOR	Canela (BR)	In-person	10	8	A. Medina Urrea
2019	PLAGAA	Guanajuato (MX)	In-person	9	7	Fernanda López
2020	PROPOR	Évora (PT)	Hybrid	9	7	Marcos Garcia
2021	STIL	Online	Online (full day)	10	10	Valeria de Paiva
2024	PROPOR	Santiago (ES)	In-person	11	7	–

Table 1: Conference statistics by year

Alfonso Medina Urrea (El Colegio de México, Red Temática de Tecnologías del Lenguaje, Mexico), who presented on Mexican corpora and their applications (Medina Urrea, 2018). Importantly, funding from PROPOR itself enabled the participation of the invited speaker, making this the only edition of OpenCor with direct financial support for this purpose.

In 2019, OpenCor[4] was held in Guanajuato, Mexico, as part of the 4th PLAGAA, *Taller Mexicano de Detección de Plagio y Análisis de Autoría*[5]. In this edition Adrian Pastor López Monroy (Centro de Investigación en Matemáticas, Mexico) contributed as the local chair. Among the submissions, two focused on Portuguese and five on Spanish or other languages. The invited speaker was Fernanda López (Universidad Nacional Autónoma de México), who presented CLOE (*Corpus de Lengua Oral del Español*) (López-Escobedo and Solorzano-Soto, 2016).

The 2020 edition[6] took place in Évora, Portugal, co-located with PROPOR. The invited speaker was Marcos Garcia (Universidade de Santiago de Compostela, Galicia), who presented an overview of open language resources for Galician (García and Crespo-Otero, 2022). This edition marked an important moment in connecting resource-building efforts across the Iberian Peninsula with those in Latin America, particularly among closely related Ibero-Romance languages such as Portuguese and Galician.

Originally planned to take place in Fortaleza, Brazil, OpenCor 2021[7] was held fully online due to the COVID-19 pandemic and co-located with STIL, the *Symposium in Information and Human Language Technology*. For the first time, OpenCor was held as a full-day workshop. The invited speaker was Valeria de Paiva (Topos Institute), who spoke about open linguistic resources for Brazilian Portuguese (Rademaker et al., 2017; de Paiva et al., 2016). The online format enabled broader participation across countries, partially mitigating travel-related funding barriers, while also introducing new organizational challenges typical of virtual events.

After a hiatus, OpenCor returned in 2024 (OpenCor at PROPOR 2024). Due to the density of the main conference program, no invited speaker was scheduled. The continued interest, reflected in the number of submissions, demonstrated the sustained relevance of the forum.

3.1. OpenCor Trends

Across its editions from 2018 to 2024, OpenCor received a total of 49 paper submissions, of which 39 were accepted, resulting in an overall acceptance rate of approximately 80%. Acceptance rates varied across years, ranging from 64% in 2024 to 100% in 2021, the latter reflecting the inclusive character of the fully online edition held during the COVID-19 pandemic.

Rather than reflecting low selectivity, these acceptance rates are indicative of the workshop's curatorial and community-building role. Submissions were evaluated primarily on their contribution to the documentation, creation, or dissemination of open corpora and lexical resources, with an emphasis on clarity, reusability, and licensing transparency. In several editions, a non-negligible number of accepted papers could not be presented due to travel and funding constraints, particularly in in-person events, underscoring structural barriers faced by researchers working on language resources in Latin America and the Iberian Peninsula.

The relatively stable number of submissions across editions further suggests sustained community interest, despite the absence of dedicated funding, travel grants, or shared-task incentives. These trends support the interpretation of OpenCor as a long-term community infrastructure effort, rather than a competitive evaluation venue.

Despite the geographic scale and academic diversity of both focused regions, the absolute number of submissions to OpenCor has remained relatively modest across editions. This should not be interpreted as a lack of interest or research activity, but rather as a reflection of structural participation constraints. Limited access to travel funding continues to be a major barrier for many researchers

[4] https://opencor.gitlab.io/program-2019/

[5] https://sites.google.com/view/plagaa2019

[6] https://opencor.gitlab.io/opencor-2020/

[7] https://comissoes.sbc.org.br/ce-pln/stil2021/programSTIL2021.pdf

in the region, particularly for international conferences. Although the 2021 edition was held fully online and achieved full acceptance, the exceptional conditions of the pandemic make it difficult to treat this edition as representative evidence that online venues alone are sufficient to address participation asymmetries. Additionally, submission patterns show a higher concentration of contributions from Brazil and Mexico, which are the countries of affiliation of the workshop organizers, suggesting that local academic networks and proximity to organizers play a significant role in community engagement when broader institutional support is lacking.

4. Open Corpora List

A central outcome of OpenCor has been the accumulation of a curated list[8] of open corpora and lexical resources for Latin American and Iberian languages. Building on the tradition of initiatives such as Linguateca, OpenCor similarly emphasizes documentation, accessibility, and community-driven reuse. The resources presented across editions are heterogeneous in scope, modality, language coverage, and level of maturity.

Rather than enforcing a fixed taxonomy, OpenCor emphasized descriptive presentations, allowing authors to discuss data collection methodologies, annotation schemes, licensing decisions, and known limitations. Over time, the OpenCor website accumulated links to these resources, gradually forming a distributed list of open datasets.

It is important to note that OpenCor does not collect linguistic data. The corpora described here correspond exclusively to the resources voluntarily submitted by participants of the workshops and related events. As a consequence, the set of languages represented in OpenCor should not be interpreted as a deliberate or comprehensive coverage of the languages of Latin America and the Iberian Peninsula. Many relevant languages are not present, including, for example, Basque, Tikuna, and Guarani Kaiowá, simply because no resources for these languages were submitted to the initiative.

OpenCor's list, despite being vastly incomplete, brings together a wide variety of open corpora and lexical resources for Latin American and Iberian languages, including both newly developed datasets and resources created before the initiative. Inclusion in the list has not been restricted to works presented at OpenCor meetings: authors may also request the addition of their resources independently, through a public submission form, available from the website of the initiative.

This author-initiated list reflects OpenCor's emphasis on community inclusion, visibility, and reuse. As of 2025, the list contains 40 resources. The contributions showcased at OpenCor cover a wide range of corpus types, beginning with traditional written and spoken corpora for Spanish (Sánchez Fernández and Medina Urrea, 2020) and Portuguese (Santos et al., 2018; Okano et al., 2020). These include both general-purpose and domain-specific collections, such as product reviews (Real et al., 2019; dos Santos Silva et al., 2024), oil and gas domain (Freitas et al., 2023) or standardized reference corpora (Crespo et al., 2023; Mendes, 2024).

Beyond language-specific corpora, the program highlights regional and national resources from Latin America and the Iberian Peninsula, addressing linguistic variation and language diversity (Pichel Campos et al., 2019). These include resources for Catalan (Rodriguez-Penagos et al., 2021) and multilingual survey questionnaires (Zavala-Rojas et al., 2021), reflecting the initiative's commitment to capturing regional diversity. Complementing these text-level datasets are structured lexical resources, such as high-coverage morphological lexicons (e.g. MorphoBR (Alencar et al., 2018)) and terminological databases (e.g. PetrolNer (Coelho and de Freitas, 2020)), as well as oral language collections (Junior et al., 2024; Othero and Ayres, 2014) that provide specialized annotation for phenomena such as emotional children's speech (Pérez-Espinosa et al., 2020), thereby broadening the functional range of available linguistic data.

A core concern of OpenCor is representation of underrepresented varieties and minority languages. Examples include corpora documenting indigenous languages, for instance, the Wixarika-Spanish Parallel Corpus (Mager et al., 2018), which provides aligned text material for an indigenous Mexican language, and ongoing efforts to build annotated parallel corpora for Nheengatu (de Alencar, 2023), the lingua franca in the Amazon basin, alongside resources for Sri Lanka Portuguese (Silva and Trigo, 2022) and Southern Quechua (Cardenas et al., 2018). Finally, the initiative also embraces non-spoken and multimodal language resources, such as multimedia corpora developed for applications like acoustic interaction research (Rascon et al., 2018) and sign language translation (Núñez-Reyes, 2016), underscoring OpenCor's engagement with multiple modalities of linguistic data.

This inclusive approach has brought visibility to corpora often marginalized in mainstream NLP venues, particularly work on Indigenous languages of Latin America and regional languages of the Iberian Peninsula. It highlights not only the existence of such data, but also the labor, linguistic diversity, and communities involved in their creation and maintenance. Notably, several of these

[8] https://opencor.gitlab.io/corpora/

corpora were first presented at OpenCor meetings during their early stages of development and were officially released at a later time.

5. Difficulties and Challenges

Despite its successes, OpenCor faces persistent structural challenges. These include limited and unstable funding, particularly for minority and indigenous language resources, where financial precarity directly affects both scope and continuity. The initiative also contends with the coordination costs inherent in geographically dispersed and institutionally diverse communities, which require sustained collaboration across national and disciplinary boundaries. Beyond initial creation, the long-term sustainability of corpora remains a significant concern, encompassing issues of hosting, licensing clarity, documentation, and ongoing updates. Finally, the labor involved in corpus development continues to suffer from insufficient academic recognition, as data creation and curation are often undervalued in hiring, promotion, and research assessment frameworks.

These difficulties are not unique to OpenCor, but reflect broader systemic issues in open scientific infrastructure. Funding has been the most persistent challenge faced by OpenCor. Apart from the initial support in 2018, the workshop never had dedicated funding for travel grants, invited speakers, or infrastructure. A funding request submitted to NAACL (NAACL Emerging Regions fund) in 2020 to support the 2021 edition was denied. As a result, participation often depends on authors' ability to self-fund or already be present at the host conference. While options such as "workshop-only" registration help reduce costs, several accepted papers across editions were not presented due to financial constraints.

These challenges highlight the broader structural issue that initiatives focused on data, openness, and linguistic diversity often struggle to secure funding compared to model-centric or commercially oriented research agendas. These challenges are not unique to OpenCor but reflect structural dynamics within the Latin American NLP ecosystem. To better situate OpenCor within this landscape, the following section examines related regional initiatives and their organizational models.

6. More Latin American Initiatives

OpenCor is not an isolated effort, but part of a landscape of community-driven initiatives dedicated to NLP and open linguistic resources in Latin America. Over the last decade, the region has witnessed the emergence of multiple forums, collectives, and research networks addressing structural gaps in visibility, infrastructure, and representation within global NLP research. Rather than forming a coordinated ecosystem, these initiatives typically operate independently, often without sustained interaction or mutual awareness. This fragmentation reflects structural constraints and contributes to the difficulty of sustaining long-term initiatives, many of which emerge, transform, or disappear over time.

Within this broader and fragmented landscape, we highlight a small number of initiatives of which we are aware, selected for their relevance to the development, visibility, and organization of language-related research in Latin America. This selection is not intended to be exhaustive, but rather illustrative of different models of community organization and engagement in the region. Although these initiatives differ in scope, structure, and objectives from OpenCor, we include them to situate our work within this wider regional context and to acknowledge that it forms part of an ongoing collective effort rather than a standalone endeavor.

Within the NLP community, one visible initiative is AmericasNLP[9], a workshop series focused on Indigenous languages of the Americas and often co-located with major international conferences such as ACL and NAACL. The venue promotes the development of corpora and shared datasets in low-resource settings and brings together researchers from universities in Latin America and Europe alongside participants affiliated with large technology companies. This configuration, also present in smaller venues, such as the Workshop on NLP for Indigenous Languages of Lusophone Countries[10], reflects a broader strategic interest in expanding multilingual AI systems to historically underrepresented languages. At the same time, the growing involvement of major industry actors raises questions about governance, ownership, and long-term control of linguistic resources, particularly where corporate infrastructures intersect with community-based language initiatives.

In Brazil, the Brazilian Symposium on Information and Human Language Technology[11] (STIL) constitutes probably the longest-running and most consolidated NLP venue in Latin America. Organized under the Brazilian Computing Society since 2003, STIL has historically served as a primary outlet for the presentation of open corpora, treebanks, and lexical resources for Brazilian Portuguese, while also welcoming work on Spanish and other languages. The 4th OpenCor edition was co-located with STIL, reflecting the overlap in community and goals, while maintaining distinct scopes.

[9] https://github.com/AmericasNLP

[10] https://sites.google.com/view/illc-nlp-2024/home

[11] https://sites.google.com/view/ce-pln/eventos/stil

Beyond formal academic venues, grassroots and independent collectives have played a decisive role in advancing open linguistic infrastructures. In Mexico, *Comunidad Elotl*[12] exemplifies a bottom-up model centered on the development of free and open-source tools and corpora for Indigenous languages such as Nahuatl, Otomí, and Mixtec. Similarly, in the Andean region, initiatives such as *Siminchik*[13] and *Siminchikkunarayku*[14] have focused on building open speech corpora for Quechua (Cardenas et al., 2018) and other native languages of Peru (Zevallos et al., 2022), often relying on community participation and crowdsourcing to overcome institutional limitations. Interestingly, many Latin American researchers focused on native language preservation are nowadays working with underrepresented languages of Spain or even with open data, but in other European countries, which may reflect structural constraints whereby researchers committed to endangered languages and open data struggle to find institutional spaces in Latin America that enable sustained, long-term, and continuous research efforts.

At a broader regional level, initiatives such as Khipu[15], a Latin American meeting on Artificial Intelligence from 2019, and institutions like CENIA[16] (*Centro Nacional de Inteligencia Artificial*) in Chile contribute to cross-national collaboration and infrastructure building. While their scope extends beyond language technologies, these initiatives reinforce a regional culture of openness, collaboration, and data sharing that directly benefits NLP research. Likewise, open data forums such as *Abrelatam Con Datos*[17] have shaped the policy and civic discourse around data as a public good, creating favorable conditions for open linguistic resources development since 2013.

Open data and data reuse are also central concerns for communities operating outside academic NLP itself, particularly those engaged with public information, transparency, and political accountability in Latin America. One such initiative is Coda.Br[18] (Conferência Brasileira de Jornalismo de Dados e Métodos Digitais), which occupies a space distinct from Computational Linguistics research while engaging directly with practices and debates around openness and data access. Organized since 2015, Coda.Br brings together journalists, civic technologists, and professionals interested in politics, transparency, and the use of open data. A distinctive feature of the conference is its commitment to regional decentralization, with recent editions held annually in the Amazon region, fostering participation beyond traditional academic and technological centers. Although not focused on language resources or NLP research, Coda.Br contributes to a broader culture of open data and critical engagement with public information that intersects with the conditions under which data, including linguistic data, are created, accessed, and sustained in Latin America.

Within this landscape, OpenCor occupies a complementary and specialized role. Rather than serving as a general NLP venue, it provides a dedicated forum for the documentation, discussion, and dissemination of open corpora and lexical resources for Latin American and Iberian languages. By focusing explicitly on data, licensing, and sustainability, OpenCor contributes to the regional ecosystem by addressing a critical but often under-recognized layer of NLP research infrastructure.

Taken together, these initiatives reveal a regional ecosystem characterized by decentralization, fragmentation, and limited structural funding. At the same time, they demonstrate that Latin America is actively shaping the development of NLP and lexical resources, rather than merely consuming technologies produced elsewhere. Researchers, collectives, and institutions across the region have built open corpora, organized conferences, and sustained collaborative networks, often under significant structural constraints. Within this context, OpenCor occupies a specific infrastructural niche: a forum dedicated not to model performance, but to the documentation and sustainability of open corpora. Its persistence across editions, despite the absence of stable funding, illustrates both the vitality and the precarity of data-centered initiatives in Latin America.

7. Conclusion

This paper has presented a historical account of OpenCor, a community-driven forum dedicated to open corpora and lexical resources for the languages of Latin America and the Iberian Peninsula. Between 2018 and 2024, OpenCor functioned as a space for documenting resources, discussing licensing and sustainability practices, and fostering exchange among researchers working on under-resourced, regional, and minority languages.

By consolidating information about meetings, submissions, and the resources presented across editions, this paper contributes to the preservation of institutional memory around an initiative whose outcomes extend beyond traditional publication metrics. In particular, it makes visible forms of labor, such as data collection, annotation, curation,

[12] https://elotl.mx/
[13] https://watuchi.org/
[14] https://siminchikkunarayku.pe/
[15] https://khipu.ai/
[16] https://cenia.cl/
[17] 'Open Latin America with Data'; https://abrelatam.org
[18] https://escoladedados.org/coda/

and maintenance, that are often underrepresented in mainstream NLP venues.

Rather than positioning OpenCor as a model to be replicated or sustained indefinitely, this historical record highlights both the possibilities and the structural constraints faced by the community-led efforts centered on openness and linguistic diversity. As such, the paper aims to serve as a reference point for researchers, organizers, and institutions interested in understanding how open language resource initiatives have emerged, operated, and evolved within the specific conditions of Latin America and the Iberian Peninsula.

7.1. Limitations

This paper provides a brief historical account of the OpenCor initiative and does not aim to be an exhaustive survey of all open corpora efforts in Latin America and the Iberian Peninsula. While we sought to reference relevant NLP groups, workshops, and initiatives, the landscape of language resource development in the region is broad, heterogeneous, and continuously evolving, with many efforts occurring outside established academic venues. The initiatives discussed also reflect, in part, the academic networks and geographic contexts in which OpenCor was organized, which may result in the under-representation of some communities or languages.

8. Bibliographical References

Leonel Figueiredo de Alencar, Bruno Cuconato, and Alexandre Rademaker. 2018. MorphoBR: An Open Source Large-Coverage Full-Form Lexicon for Morphological Analysis of Portuguese. *Texto Livre: Linguagem e Tecnologia*, 11(3):1–25.

Ronald A. Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. In *Proceedings of the Workshop "Improving Social Inclusion Using NLP: Tools, Methods and Resources" co-located with the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 21–28, Miyazaki, Japan. European Language Resources Association (ELRA).

Leonardo Gularte Coelho and Larissa Astrogildo de Freitas. 2020. Construção do corpora calendário brasileiro de saúde. In *Anais da Semana Integrada de Inovação, Ensino, Pesquisa e Extensão (SIIEPE)*, Pelotas, Brazil. Paper presented at SIIEPE 2020, Universidade Federal de Pelotas.

Maria Clara Ramos Morales Crespo, Maria Lina de Souza Jeannine Rocha, Mariana Lourenço Sturzeneker, Felipe Ribas Serras, Guilherme Lamartine de Mello, Aline Silva Costa, Mayara Feliciano Palma, Renata Morais Mesquita, Raquel de Paula Guets, Mariana Marques da Silva, Marcelo Finger, Maria Clara Paixão de Sousa, Cristiane Namiuti, and Vanessa Martins do Monte. 2023. Carolina: a general corpus of contemporary brazilian portuguese with provenance, typology and versioning information.

Leonel Figueiredo de Alencar. 2023. Yauti: A tool for morphosyntactic analysis of nheengatu within the universal dependencies framework. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 135–145, Porto Alegre, RS, Brazil. Sociedade Brasileira de Computação (SBC).

Valeria de Paiva, Livy Real, Hugo Gonçalo Oliveira, Claudia Freitas, Alexandre Rademaker, and Alberto Simões. 2016. An overview of portuguese wordnets. In *Proceedings of the 8th Global WordNet Conference*, Bucharest, Romania.

Lucas Nildaimon dos Santos Silva, Ana Cláudia Zandavalle, Carolina Francisco Gadelha Rodrigues, Tatiana da Silva Gama, Fernando Guedes Souza, Phillipe Derwich Silva Zaidan, Alice Florencio Severino da Silva, Karina Soares, and Livy Real. 2024. Repro: A benchmark dataset for opinion mining in brazilian portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese – Vol. 1*, pages 432–440, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Cláudia Freitas, Elvis de Souza, Maria Clara Castro, Tatiana Cavalcanti, Patrícia Ferreira da Silva, and Fábio Corrêa Cordeiro. 2023. Recursos linguísticos para o pln específico de domínio: o petrolês. *Linguamática*, 15(2):51–68.

Marcos García and Alfredo Crespo-Otero. 2022. A targeted assessment of the syntactic abilities of transformer models for galician-portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, volume 13208 of *Lecture Notes in Computer Science*, pages 46–56. Springer.

Isaac Junior, Gabriela Wick-Pedro, Cláudia Barros, and Oto Vale. 2024. Roda viva boundaries: an overview of an audio-transcription corpus. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*.

Fernando López-Escobedo and Jorge Solorzano-Soto. 2016. Propuesta de clasificación de un banco de voces con fines de identificación forense. *Linguamática*, 8(1):33–41.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Alfonso Medina Urrea. 2018. Pluricentric languages: New perspectives in theory and description. *Nueva Revista de Filología Hispánica*, 66(1):211–214.

Amália Mendes. 2024. The reference corpus of contemporary portuguese: Corpus design and case study on discourse markers. In Miguel Calderón Campos and Gael Vaamonde, editors, *Linguistic Corpora and Big Data in Spanish and Portuguese*, pages 145–178. Walter de Gruyter GmbH.

Alba Núñez-Reyes. 2016. Agrupamiento de textos cortos en dominios cruzados. Technical report, Repositorio Institucional, Instituto de Lingüística, Universidad Autónoma Metropolitana.

Emerson Yoshiaki Okano, Zebin Liu, Donghong Ji, and Evandro Eduardo Seron Ruiz. 2020. Fake news detection on fake.br using hierarchical attention networks. In *Computational Processing of the Portuguese Language*, pages 143–152, Cham. Springer International Publishing.

Gabriel de Ávila Othero and Mônica Rigo Ayres. 2014. Anotação morfológica automática de corpus de língua falada: desafios ao Aelius. *Texto Livre: Linguagem e Tecnologia*, 7(2):44–60.

José Ramón Pichel Campos, Pablo Gamallo Otero, and Iñaki Alegria Loinaz. 2019. Measuring diachronic language distance using perplexity: Application to English, Portuguese, and Spanish. *Natural Language Engineering*, 26(4):433–454.

Humberto Pérez-Espinosa, Juan Martínez-Miranda, Ismael Espinosa-Curiel, Josefina Rodríguez-Jacobo, Luis Villaseñor-Pineda, and Himer Avila-George. 2020. Iesc-child: An interactive emotional children's speech corpus. *Computer Speech & Language*, 59:55–74.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Caleb Rascon, Ivan V. Meza, Aldo Millan-Gonzalez, Ivette Velez, Gibran Fuentes, Dennis Mendoza, and Oscar Ruiz-Espitia. 2018. Acoustic interactions for robot audition: A corpus of real auditory scenes. *The Journal of the Acoustical Society of America*, 144(5):EL399–EL403.

Livy Real, Marcio Oshiro, and Alexandre Mafra. 2019. B2w-reviews01: an open product reviews corpus. In *Proceedings of the Symposium in Information and Human Language Technology (STIL 2019)*, Brazil. Dataset and paper presented at STIL 2019.

Carlos Rodriguez-Penagos, Carme Armentano-Oller, Marta Villegas, Maite Melero, Aitor Gonzalez, Ona de Gibert Bonet, and Casimiro Carrino Pio. 2021. The catalan language club.

Manuel Alejandro Sánchez Fernández and Alfonso Medina Urrea. 2020. Hacia el etiquetado de estados informativos en el corpus periodístico del noroeste de méxico (copenor). *Signos Lingüísticos*, 16(31):164–181.

Diana Santos. 2003. Relatório linguateca: Relatório relativo ao período 2000–2003. Technical report, Linguateca.

Diana Santos, Cláudia Freitas, and Eckhard Bick. 2018. Obras: a fully annotated and partially human-revised corpus of brazilian literary works in public domain. *In CorLex*.

C. Silva and L. Trigo. 2022. *PtLanka*. CLUP – Centro de Linguística da Universidade do Porto, Porto.

Diana Zavala-Rojas, Danielly Sorato, Lidun Hareide, and Knut Hofland. 2021. [MCSQ] Multilingual Corpus of Survey Questionnaires. *Meta: Journal des traducteurs*.

Rodolfo Zevallos, Luis Camacho, and Nelsi Melgarejo. 2022. Huqariq: A multilingual speech corpus of native languages of Peru for Speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5029–5034, Marseille, France. European Language Resources Association.