The PROPOR Ecosystem: Structure, Roles, and Evolution of Portuguese-Language NLP

Rafael O. Nunes[1], **Gustavo L. Tamiosso**[1], **Pedro L. C. de Andrade**[2], **Matheus S. de Aguiar**[1], **Rafael P. de Gouveia**[2], **Higor Moreira**[1], **Bruno Tavares**[1], **Laura P. de Gouveia**[2], **Felipe S. F. Paula**[1], **Andre Spritzer**[1], **Hidelberg O. Albuquerque**[3], **Nádia F. F. da Silva**[4], **Ellen P. R. S. Pereira**[3], **Dennis G. Balreira**[1] and **Joel L. Carbonera**[1]

[1]UFRGS, Brazil [2]USP, Brazil [3]UFRPE, Brazil [4]UFG, Brazil

{ronunes,gltamiosso,msaguiar,hmoreira,bruno.tsantos,spritzer,dgbalreira,jlcarbonera}@inf.ufrgs.br

felipesfpaula@gmail.com, {pedroandrade,rafael.p.gouveia,laura.gouveia}@usp.br

{hidelberg.albuquerque,ellen.ramos}@ufrpe.br, nadia.felix@ufg.br

Abstract

The PROPOR conference has been the primary venue for Portuguese-language Natural Language Processing (NLP) research for over two decades. This paper presents a longitudinal bibliometric analysis of PROPOR (2003–2024), examining thematic evolution, community structure, and scientific impact. The results reveal a shift from speech-oriented research toward text-based tasks, alongside the continued centrality of resources and linguistic theory. The community displays a stable yet diversified structure, combining institutional hubs with brokerage-driven leadership. Scientific impact is strongly concentrated, following a long-tail pattern that contrasts cumulative productivity with the rapid citation growth of recent editions. Overall, PROPOR emerges as a resilient regional linguistic ecosystem evolving in close alignment with broader NLP trends.

1 Introduction

The International Conference on Computational Processing of Portuguese (PROPOR) has stood out as one of the largest events focused on Natural Language Processing (NLP) for the Portuguese language, with a vast body of scientific production accumulated over two decades[1]. This bibliometric analysis examines the main research trends and the evolution of contributions published at PROPOR, addressing both core topics and the most explored tasks and linguistic variants. Furthermore, it investigates the role of the research community, collaboration networks, and the conference's position within the broader global NLP landscape.

Based on meta-science studies by Pramanick et al. (2025), Movva et al. (2024), and other key researchers, this work identifies patterns and dynamics in Portuguese-language NLP research. However, despite its two-decade history, to the best of our knowledge, no systematic, longitudinal bibliometric study has comprehensively mapped its community structure, thematic evolution, and scientific impact. This work aims to fill that gap.

This analysis is guided by the following research questions: **RQ1 (What?):** How has the thematic landscape of PROPOR evolved between 2003 and 2024, particularly regarding the distribution of research topics, NLP tasks, and the representation of Portuguese language variants? **RQ2 (Who?):** What are the prevailing patterns of author retention, renewal, and collaboration among researchers and institutions within the PROPOR community? **RQ3 (What Impact?):** How do citation dynamics relate to productivity and collaboration, and what does this reveal about the conference's scientific impact?

2 Related Work

Understanding the structure and evolution of the NLP research community has been addressed by several meta-scientific studies. Early work analyzed large-scale trends in major venues such as ACL and EMNLP using techniques like keyphrase extraction and topic modeling (Anderson et al., 2012). Later studies examined topic evolution and author behavior within the ACL Anthology (Gollapalli and Li, 2015), as well as the nature of contributions in top-tier NLP venues (Pramanick et al., 2025). More recently, large-scale analyses of arXiv data have tracked emerging trends and shifts in research focus, particularly with the rise of large language models (LLMs) (Movva et al., 2024).

While these studies provide valuable insights into the global NLP landscape, they largely focus on high-resource languages and flagship venues, offering limited visibility into the dynamics of regional or language-specific research communities. In this context, targeted analyses of specific languages and communities have been shown to reveal distinct trajectories and priorities (Schneider et al.,

[1]https://propor.org/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 342–350

2022; Souza et al., 2018).

To the best of our knowledge, the only prior work involving the PROPOR conference is the study by Leal et al. (2024), which investigates reproducibility practices across major Portuguese language venues. However, their focus is not on PROPOR as a research community or historical object, but rather on reproducibility issues at the paper level. In contrast, the present work offers the first comprehensive longitudinal analysis of PROPOR itself, examining its thematic evolution, community structure, and scientific impact over a period of more than two decades.

3 Corpus Annotation

This study analyzes PROPOR proceedings published between 2003 and 2024. We selected this period because full proceedings are available through Springer for 2003–2022[2] and through the ACL Anthology for 2024[3]. Earlier editions were excluded due to the lack of publicly available proceedings and incomplete archival records.

As Springer proceedings are not openly accessible, our analysis relies on paper titles, authors, affiliations, and abstracts. Since topic labels are not provided in the proceedings, the corpus was manually annotated by eight NLP researchers and one curator following the official PROPOR topic taxonomy[4]. Institutional affiliations and citation counts were automatically collected using the OpenAlex API[5], with missing metadata for 87 papers manually verified and completed.

Annotation followed established practices (Hovy and Prabhumoye, 2021; Santos et al., 2024; Pramanick et al., 2025). The classification categories were derived directly from the PROPOR Call for Papers and include:

- **Natural language processing tasks**
- **Natural language processing applications**
- **Natural language generation**
- **Information extraction and information retrieval**
- **Speech technologies**
- **Speech applications**
- **Resources, standardization and evaluation**
- **NLP-oriented linguistic description or theoretical analysis**
- **Distributional semantics and language modeling**
- **Language varieties and dialect processing**
- **Multilingual studies, methods, applications, and resources, Portuguese and/or Galician**

Although the official PROPOR call for papers has evolved over the years, the taxonomy used here was kept stable across all editions to allow longitudinal comparability. Categories that did not exist in earlier calls were applied retroactively based on the conceptual scope of the papers.

The 429 papers were distributed among four annotator pairs. Inter-annotator agreement was assessed using Cohen's Kappa (Cohen, 1960), with values ranging from 0.36 to 0.51. Subsequent adjudication discussions revealed recurring sources of disagreement, particularly between NLP tasks and NLP applications, as well as between speech technologies and speech applications. To address these issues, foundational methods were classified as tasks or technologies, while user-oriented systems were classified as applications. The category Language varieties and dialect processing was restricted to studies explicitly comparing two or more varieties, as all PROPOR papers inherently involve a variety of Portuguese.

Following guideline refinement, annotators conducted a consensus phase to reassess disputed cases and harmonize labels. This iterative process resulted in a consistent, consolidated topic annotation across the entire corpus, with full consensus reached on all disputed cases during the adjudication discussions. Detailed decision rules for borderline cases, such as the NLP Tasks vs. NLP Applications boundary and the scope of Language Varieties, are documented in the full annotation guidelines available in the project repository.[6]

4 Results

This section presents the results structured around the three research questions defined in Section

[2] https://link.springer.com/conference/propor

[3] https://aclanthology.org/venues/propor/

[4] http://www.wikicfp.com/cfp/servlet/event.showcfp?copyownerid=90704&eventid=190288

[5] https://docs.openalex.org/

[6] Full guideline and corpus: https://github.com/RafaelOleques/propor-ecosystem

1. Each subsection addresses a distinct analytical dimension: the conference's thematic landscape (RQ1), the structure and evolution of its research community (RQ2), and its scientific impact (RQ3).

4.1 RQ1: Thematic Landscape (What?)

Regarding research topics, we analyze their evolution from a historical perspective, combining three complementary views: the longitudinal development of each topic across editions (Figure 1), their overall distribution (Figure 2), and the ranking dynamics of topics that reached the top three positions in at least one year (Figure 3). To ensure comparability across time, we adopt a stable topic taxonomy aligned with the official calls for papers and program committee structure of PROPOR. A consistent color palette is maintained across all visualizations. In the longitudinal analysis, solid lines denote uninterrupted presence across editions, while dashed lines indicate periods of absence.

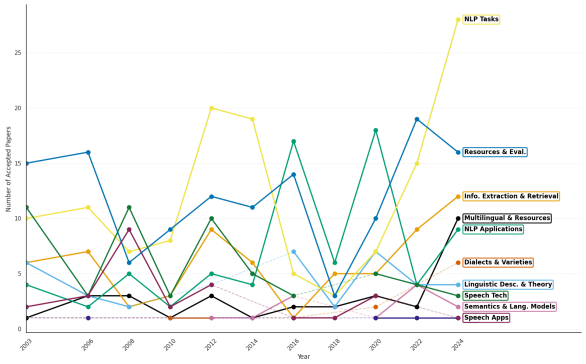

Figure 1: Longitudinal distribution of major research topics. Solid lines indicate years with publications, while dotted lines denote interpolated values for temporal gaps.

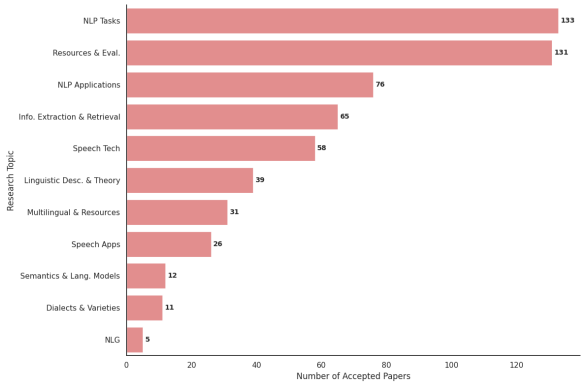

Figure 2: Overall distribution of research topics.

The overall topic distribution reveals two enduring pillars in the conference history: NLP Tasks

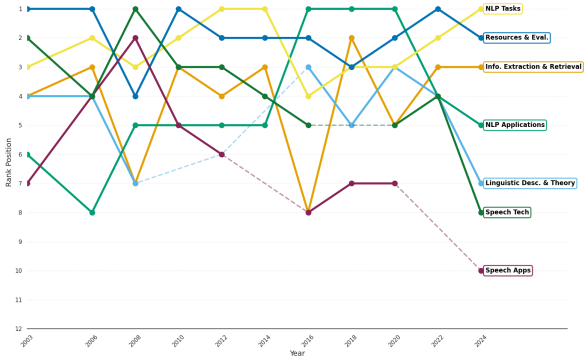

Figure 3: Rank evolution of research topics that reached the Top 3 in at least one conference edition.

($n = 133$) and Resources & Eval. ($n = 131$). Their sustained prominence reflects a dual commitment that has characterized PROPOR: the construction of linguistic infrastructure and the investigation of concrete language processing problems. This balance situates the conference at the intersection of methodological foundations and applied research, a defining feature of its identity.

Longitudinal trends (Figure 1) indicate that this balance has not remained static. NLP Tasks, which did not dominate the early editions, gradually consolidated its position and exhibits a pronounced increase in the 2024 edition. This growth temporally aligns with the broader emergence of LLMs (Movva et al., 2024) and the resulting expansion of task-oriented evaluation. While this trajectory mirrors global developments in NLP (Pramanick et al., 2025), its manifestation at PROPOR reflects a localized reappropriation, as task-driven research frequently engages with Portuguese data, linguistic variation, and region-specific evaluation challenges.

From Signal to Characters. Ranking dynamics (Figure 3) further illustrate a gradual reconfiguration of the conference scope. Text centered topics increasingly occupy leading positions, while speech related research, once central in the early editions, shows a sustained decline over time. Importantly, this decline is structural rather than episodic. Both Speech Tech and Speech Apps lose prominence without corresponding reemergence under alternative categories, suggesting that speech oriented research has progressively migrated to other venues. This shift marks a long term transition in PROPOR from signal processing oriented work toward textual semantics and application focused research.

Linguistic Theory as a Recurrent Anchor. Against this backdrop of methodological change, Linguistic Description & Theory ($n = 39$, see Figure 2) plays a distinctive historical role. As shown in Figure 1, this topic does not maintain continuous presence and experiences a near absence between 2010 and 2014. However, its subsequent resurgence in 2016 indicates a renewed engagement with linguistic theory. Rather than an anomaly, this pattern is characteristic of PROPOR, whose foundations explicitly bridge computational and descriptive linguistics. The recurrent return of theoretical work suggests an enduring institutional commitment to linguistic analysis, even as statistical and neural methods gain prominence.

Continuity and Reframing in Specialized Areas. Beyond the major axes, several specialized topics reveal important aspects of continuity and reframing within the conference. Multilingual & Resources exhibits a late but steady rise. While marginal until 2018, it grows consistently in subsequent editions, reaching ten papers in 2024. Rather than introducing an entirely new concern, this trend can be interpreted as a reactivation of longstanding interests in linguistic diversity, now reframed through contemporary cross-lingual methodologies. This development aligns closely with the historical mission of PROPOR as a venue attentive to non-English languages.

Information Extraction & Retrieval demonstrates remarkable stability across editions. It consistently occupies the third or fourth position in the rankings (Figure 3), functioning as a persistent utility layer that supports a wide range of research agendas. Its durability contrasts with more volatile topics and underscores its role as an enabling rather than trend driven area.

Natural Language Generation (NLG) appears as the least populated category ($n = 5$). This low count does not reflect a substantive absence of generative research, but rather a shift in disciplinary labeling. In the contemporary landscape, work involving text generation is rarely framed explicitly as NLG, a term historically associated with pipeline based architectures. Instead, such contributions are presented as model centered or task specific studies and are consequently absorbed into broader categories such as NLP Tasks or Semantics & Language Models. This terminological shift is itself indicative of changing conceptualizations within the field.

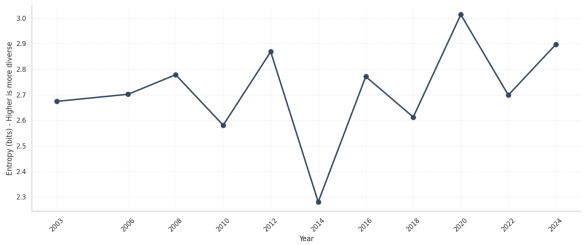

Figure 4: Evolution of Thematic Diversity (Shannon Entropy).

The thematic health of the conference is further assessed through the evolution of Shannon entropy (Figure 4). Rather than exhibiting a monotonic trajectory, entropy values reveal cycles of thematic concentration followed by diversification. The lowest entropy is observed in 2014 at approximately 2.28 bits, indicating a period of heightened focus. Subsequent peaks in 2020 and 2024, approaching 3.0 bits, reflect a more balanced distribution of research topics. These fluctuations suggest that PROPOR has repeatedly absorbed paradigm shifts without converging on a single dominant theme, maintaining a resilient and multifaceted research profile over time.

4.2 RQ2: The Community (Who?)

Shifting the analysis from research topics to the community structure, the data reveals a field currently undergoing a significant demographic transformation. By examining author retention, longevity, and collaboration habits, we identify three structural characteristics: a recent demographic expansion driven by external factors, a reliance on transient contributions, and a definitive shift toward larger collaborative teams.

4.2.1 How do Authors Interact with Conferences?

The *Hype* Expansion and Retention Challenges. The evolution of the author base (Figure 5) depicts a community in rapid expansion. The 2024 edition stands as a statistical outlier, recording an unprecedented influx of nearly 200 new authors, virtually doubling the figures of previous peaks. This surge strongly correlates with the global popularization of LLMs, which likely attracted researchers from adjacent fields and new students to the NLP domain, compounded by the higher total number of accepted papers in the 2024 edition compared to previous ones. However, this growth introduces a challenge regarding community stability: the reten-

tion rate (i.e., the percentage of returning authors) has remained relatively low, fluctuating between 15% and 25% in recent years. This suggests that while the conference successfully attracts new talent, it currently operates as a "revolving door," with a high volume of newcomers who may not necessarily establish a lasting presence in the community.

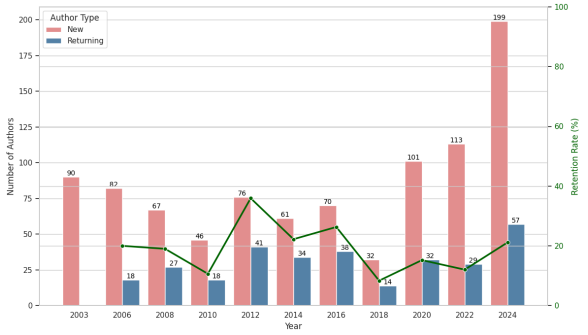

Figure 5: Evolution of authors, illustrating the absolute counts of **New** and **Returning** authors (bars, left axis) alongside the corresponding author **retention rate** (line, right axis) from the previous edition.

The *Hourglass* Demographics. This retention dynamic is further explained by the longevity analysis (Figure 6), which follows a strict power-law distribution. The vast majority of the community consists of transient contributors who appear in a single edition. This pattern is consistent with the academic lifecycle of Master's and PhD students, who drive the volume of submissions but leave the field upon graduation. Conversely, the community's continuity relies on a remarkably small fraction of authors who have participated in more than five editions. This structural duality suggests that the conference is sustained by two distinct forces: a massive, renewing base of students providing experimental volume, and a small, stable nucleus of senior researchers providing historical consistency and direction.

From Individual Efforts to Consortiums. Beyond demographics, the *modus operandi* of research has shifted. The analysis of team sizes (Figure 7) indicates the decline of small, advisor–student teams and the emergence of a *Big Science* paradigm in Portuguese NLP. In early editions, the median paper was authored by teams of two or three. By 2024, the distribution has shifted upward, with the upper quartile reaching five authors and outliers exceeding ten collaborators. This trend reflects the increasing complexity of modern

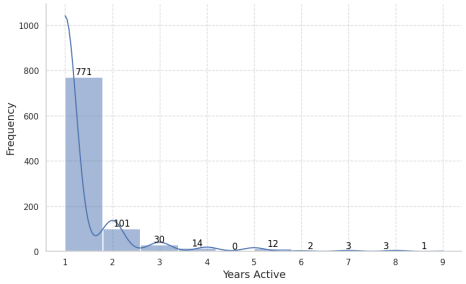

Figure 6: Distribution of author longevity. The histogram shows the absolute frequency of authors based on the total number of years they have published at the conference.

NLP: current state-of-the-art research requires a diverse set of skills, ranging from linguistic curation and engineering to heavy computational resource management, that exceeds the capacity of smaller teams.

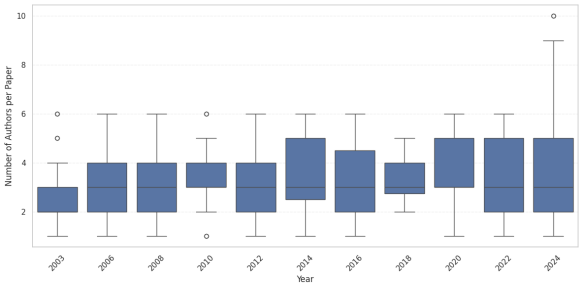

Figure 7: Distribution of team sizes by year.

4.2.2 Which Institutions are Present?

Complementing the demographic analysis, we examine the institutional backbone of PROPOR. Unlike the high turnover observed among individual authors, institutions serve as the stable nodes of the network, accumulating resources and reputation over decades. The data exposes a structure defined by high centralization around a transatlantic axis, recently challenged by a rapid peripheral expansion.

The hierarchy of production and impact (Figure 8 and Figure 9) reveals that the conference is anchored by a consolidated Luso-Brazilian oligopoly. At the forefront, the *University of São Paulo (USP)* establishes itself as the undisputed leader in both volume ($n = 72$ papers) and impact ($n = 560$ citations). Its dominance is not merely as a participant, but as the primary gravitational center for Portuguese NLP research in Latin America. On the European side, a duality emerges between volume and efficiency. While *ULisboa* ranks second in pro-

duction volume ($n = 62$), *INESC-ID* surpasses it in accumulated citations ($n = 401$ vs. $n = 356$), suggesting a highly efficient ratio of impact per paper. *USP*, *ULisboa*, and *INESC-ID* appear among the most frequent institutional contributors to the conference. Below these top-tier hubs, we observe a steep drop in volume, where regional leaders such as *Coimbra*, *UFSCar*, and *PUCRS* form a secondary tier. The significant gap between the top three and the rest of the field confirms a high degree of concentration in knowledge production.

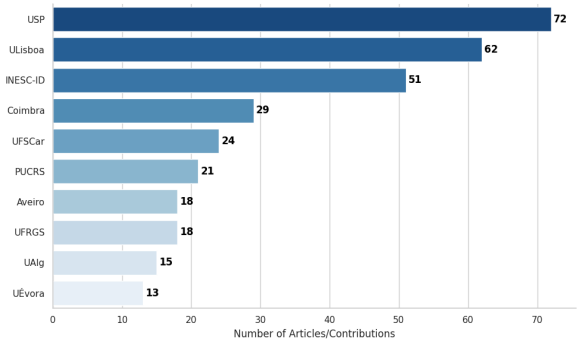

Figure 8: Top 10 most productive institutions by number of articles.

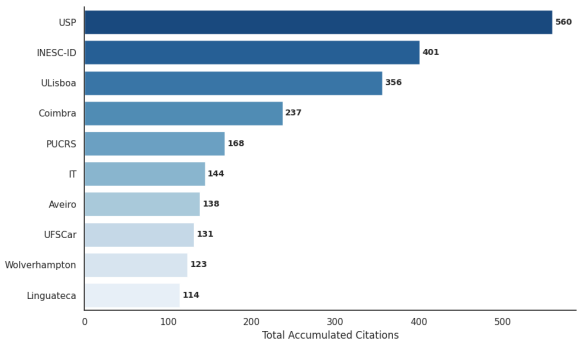

Figure 9: Top 10 institutions by total citations.

Network Topology and Cohesion. The visualization of the co-authorship network (Figure 10) and its density metrics (Figure 11) illustrates how these institutions interact to maintain community cohesion. The network graph reveals a "Hub-and-Spoke" topology, where the giant component is held together by central nodes (USP, ULisboa) acting as bridges connecting smaller, otherwise isolated institutions. This structure ensures that despite geographic dispersion, the community remains connected via these large research centers. Longitudinally, this expansion creates a *Density Paradox*: the network density has dropped significantly, from a peak of ≈ 0.12 in 2014 to a low of ≈ 0.04 in 2024. This decline should not be interpreted as a fragmentation of the community, but rather as a mathematical consequence of rapid expansion. As the conference evolves from a *small village* (where most nodes connected) to a *metropolis*, the number of possible connections grows quadratically, naturally diluting the density metric.

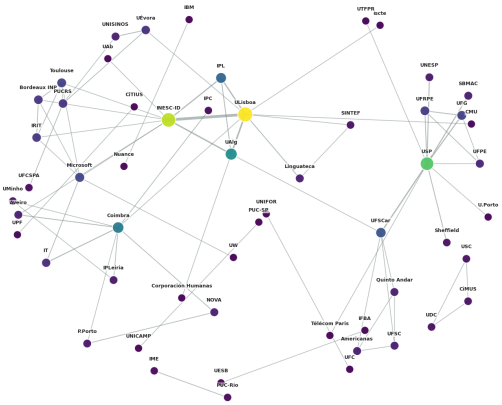

Figure 10: Co-authorship Network of institutions (collaborations ≥ 2 articles). Node size and color are proportional to productivity. Edge thickness represents the frequency of collaboration.

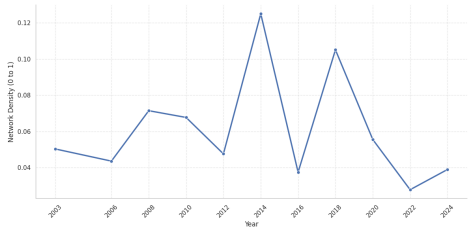

Figure 11: Evolution of the Density of the Institutional Collaboration Network.

Institutional Stability vs. Individual Volatility. Finally, the entry dynamics (Figure 12) highlight a crucial divergence between authors and institutions. The 2024 edition recorded a rare phenomenon where the number of new entering institutions ($n = 36$) surpassed the number of returning ones ($n = 34$). This suggests a massive influx of new actors, likely drawn by the LLM hype, entering at the periphery of the network, which further explains the drop in density. Despite this influx, institutional retention remains robust (fluctuating between 40% and 60%), significantly higher than author retention. This confirms that institutions function as the "memory" of the conference: while

students constitute the transient workforce, the laboratories and universities provide the necessary infrastructure and continuity for the field to evolve.

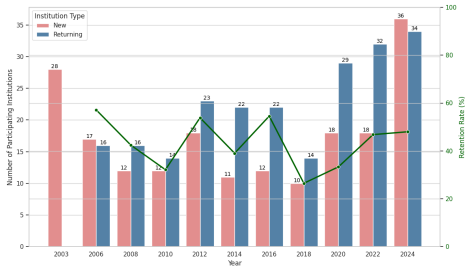

Figure 12: Distribution of new institutions entering PROPOR over time.

4.3 RQ3: Scientific Impact (What Impact?)

While aggregate citation counts capture the cumulative visibility of the conference, they offer limited insight into the mechanisms through which impact is generated. A combined analysis of citation distribution, index-based indicators, and citation velocity reveals a structurally concentrated impact profile, shaped by a small number of high-performing contributions and by changes in the temporal dynamics of scholarly attention.

Impact Overview. The citation distribution (Figure 13) indicates a pronounced concentration of impact. Most papers receive relatively few citations, while a very limited subset accounts for a disproportionate share of the total citation volume. This long tail configuration suggests that the conference impact is not the result of uniform incremental contributions, but rather of episodic strong influence outputs that shape the venue's external visibility.

This concentration effect is further substantiated by the persistent divergence between the h-index and g-index across five-year windows (Figure 14). The systematic dominance of the g-index over the h-index indicates that impact accrual is driven more by peak performance than by consistent medium-

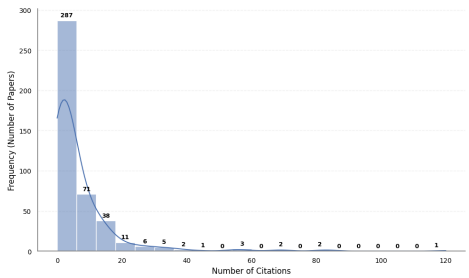

Figure 13: Distribution of citations.

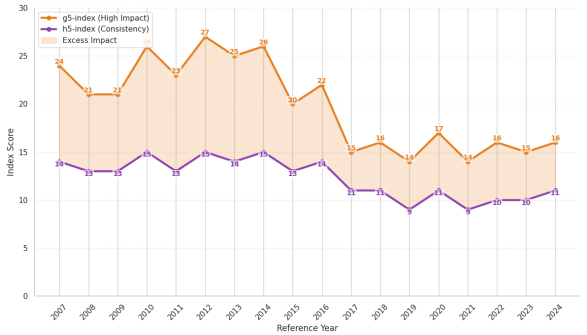

Figure 14: Evolution of scientific impact metrics. Calculated over a 5-year sliding window, the **h5-index** reflects the consistent production of impactful papers, while the **g5-index** accounts for the accumulated citations of highly cited works.

level citation activity. The magnitude of the gap between the two indices serves as a proxy for the structural reliance on outlier papers. This effect intensifies between 2010 and 2014, reaching a maximum in 2012, and weakens temporarily around 2018, in parallel with a contraction in publication volume, before re-stabilizing in the post-2020 period. These dynamics point to cycles of impact concentration rather than linear growth.

Beyond cumulative influence, citation velocity provides insight into how quickly research contributions are integrated into the scholarly discourse. The velocity analysis (Figure 15) reveals a clear temporal shift in impact formation. Earlier high-impact papers accumulated citations gradually over several years, whereas recent contributions achieve comparable or greater annual citation rates shortly after publication. The emergence of a 2024 paper with a citation rate of 18.5 citations per year exemplifies this acceleration and contrasts sharply with the slower uptake observed in earlier periods. This shift suggests a transformation in dissemination and consumption patterns, where immediacy increasingly complements longevity as a component of scientific impact. Additionally, the indexation of PROPOR 2024 proceedings in the ACL Anthology as open-access papers may have contributed to accelerating the pace at which recent contributions are discovered and cited, reducing the lag between publication and uptake that characterized earlier editions.

Thematic Impact. The analysis of scientific impact through research topics reveals a structural transition in the relevance of themes over the decades. By examining the trajectories of topics

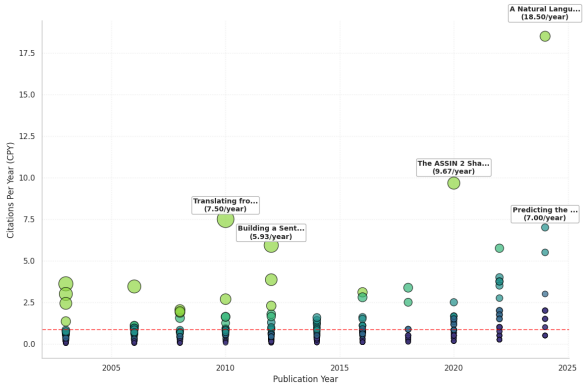

Figure 15: Scatter plot of research impact acceleration measured by Citations Per Year (CPY). Marker size is proportional to the **Total Citation Count**. The dashed red line denotes the dataset mean CPY.

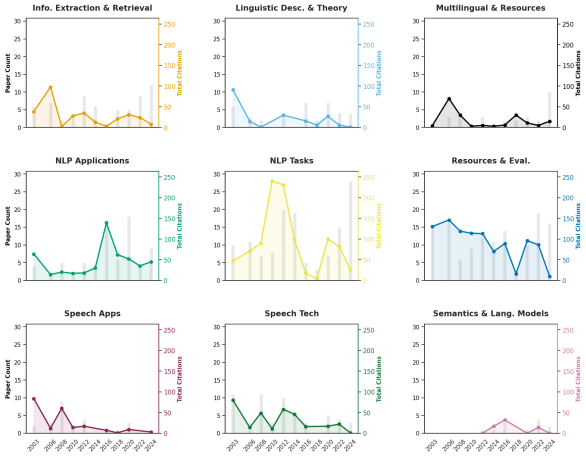

Figure 16: Facet grid comparing publication volume and citation impact over time for topics that ranked among the top three most frequent in at least one year.

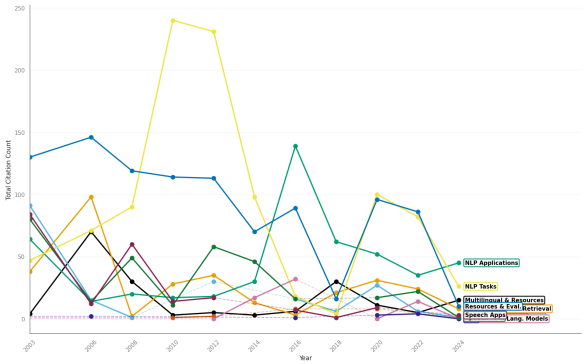

Figure 17: Temporal evolution of scientific impact by research category.

that reached the top tier of citations, we identify the primary drivers of the conference's intellectual influence and the shifting paradigms within the community, as seen in Section 4.1.

The longitudinal data (Figure 16 and 17) establishes *NLP Tasks* and *Resources & Evaluation* as the greatest impact of the topic for PROPOR. These categories consistently have the highest volumes of citations, with a historical peak occurring between 2010 and 2012. During this window, *NLP Tasks* recorded a massive surge, exceeding 240 citations in a single period—a level of influence that remains an outlier in the conference's history. This suggests that during the early 2010s, the community focused heavily on establishing core processing tasks that served as the basis for subsequent research.

A chronological assessment of these citations identifies a clear thematic rotation. In the initial editions, such as 2003, the impact was concentrated in speech-related areas and theoretical foundations, including *Speech Tech*, *Speech Apps*, and *Linguistic Description & Theory*, all of which registered more than 80 citations annually. However, these themes exhibit a persistent downward trend in the following editions. The *Linguistic Description & Theory* category, for instance, saw its impact diminish significantly after the initial peak, reaching levels near zero in the last decade. This decline highlights a shift in the community's interest away from traditional linguistic modeling toward more application-oriented and data-driven computational approaches.

In recent years, the diversification of impact has been evident through the emergence of new themes and cyclical interests. *NLP Applications* showed a notable spike in 2016, temporarily leading the citation volume for that edition, while *Semantics & Language Models* began to show significant bibliographic traction only after 2012, with peaks in 2016 and 2022. In contrast, niche areas like *Information Extraction & Retrieval* demonstrate a cyclical impact pattern, maintaining constant relevance through recurring peaks of interest. The lower absolute citation volumes observed across all categories in the 2024 edition should not be interpreted as a decline in research quality or relevance. Rather, this is an expected artifact of recency: papers published recently have had less time to accumulate citations, and their full impact will only become visible in future analyses.

Final Remarks. Taken together, these findings characterize the scientific impact of PROPOR as structurally concentrated but temporally dynamic. While a small number of papers continue to anchor the conference's long-term citation footprint, recent

editions exhibit faster cycles of recognition and engagement. This dual pattern aligns with broader changes in scholarly communication, while also reflecting the conference's ability to periodically produce contributions that resonate beyond its immediate community.

5 Conclusion

This paper presented the first longitudinal analysis of the PROPOR conference, examining over two decades of research output across thematic evolution, community structure, and scientific impact. The findings reveal a gradual shift from speech-oriented research toward text-based tasks, alongside the sustained relevance of linguistic resources and theory. PROPOR exhibits a stable community structure, supported by complementary institutional roles, while its scientific impact remains structurally concentrated but shows accelerated uptake in recent editions, reflecting its adaptation to contemporary NLP dynamics.

The study is limited by its reliance on bibliographic metadata and abstract-level analysis, particularly for earlier editions, as well as by the inherent abstraction of topic categorization. Future work could incorporate full-text and citation context analysis, and extend the comparison to other regional or language-specific NLP venues, thereby enabling a deeper understanding of how regional linguistic ecosystems, such as PROPOR, evolve within the global NLP landscape.

References

Ashton Anderson, Dan Jurafsky, and Daniel A. McFarland. 2012. Towards a computational history of the ACL: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21, Jeju Island, Korea. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Sujatha Das Gollapalli and Xiaoli Li. 2015. EMNLP versus ACL: Analyzing NLP research over time. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2002–2006, Lisbon, Portugal. Association for Computational Linguistics.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Daniel Leal, Anthony Luz, and Rafael Anchiêta. 2024. A reproducibility analysis of portuguese computational processing conferences: A case of study. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 605–609.

Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. 2024. Topics, authors, and institutions in large language model research: Trends from 17K arXiv papers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1223–1243, Mexico City, Mexico. Association for Computational Linguistics.

Aniket Pramanick, Yufang Hou, Saif M. Mohammad, and Iryna Gurevych. 2025. The nature of NLP: Analyzing contributions in NLP papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25169–25191, Vienna, Austria. Association for Computational Linguistics.

Joaquim Santos, Helena Freire Cameron, Fernanda Olival, Fátima Farrica, and Renata Vieira. 2024. Named entity recognition specialised for portuguese 18th-century history research. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 117–126.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.

Ellen Souza, Danilo Costa, Dayvid W Castro, Douglas Vitório, Ingryd Teles, Rafaela Almeida, Tiago Alves, Adriano LI Oliveira, and Cristine Gusmão. 2018. Characterising text mining: a systematic mapping review of the portuguese language. *Iet Software*, 12(2):49–75.