The faces of Latin American research on computational linguistics

Anonymous ACL submission

Abstract

Latin America (LatAm) is mainly represented by developing or under-developed countries, and given that their investment in research and development may be lacking, their presence in high-impact research communities such as computational linguistics (CL) could be marginal. This work aims to measure the presence of LatAm researchers in the CL community and improve the visibility of those underrepresented investigators. We extracted the metadata of all ACL Anthology publications with at least one LatAm researcher affiliated with an institution located in an LatAm country at the time of the publication. We found that only a small percentage (2.4 %) of CL publications have affiliations with institutions based in LatAm and that Alexander Gelbukh (Mexico) and Luciana Benotti (Argentina) are the most productive researchers. Our analysis also reveals that some countries in the region have not contributed to any CL publications. Despite these challenges, our results highlight the potential for growth and improvement. By shedding light on the underrepresentation of LatAm researchers in CL, this study aims to promote greater visibility and inclusivity within the community, ultimately fostering a more diverse and vibrant research landscape.

1 Introduction

Latin America (LA) is characterized by significant economic and educational disparities, with only a few countries, such as Chile, Argentina, Uruguay and Costa Rica, boasting very high Human Development Index (HDI) scores (United Nations, 2024). As a result, many LatAm countries struggle to invest in essential areas such as education, research, and development (Ortega et al., 2022), which can limit their participation in high-impact scientific communities like computational linguistics (CL). This limited investment can hinder the region's ability to develop a robust research infrastructure, train a skilled workforce, and produce innovative research outputs, ultimately affecting their global visibility and influence in fields like CL.

The term LatAm refers to a region formed by a set of countries on the American continent whose languages derive from Latin. Geographically, it includes most parts of the American continent, from the south at Tierra del Fuego in Chile to the Rio Grande at the border between Mexico and the United States.

The field of CL, encompassing natural language processing and human language technologies, has experienced a remarkable surge in popularity with the emergence of large language models (LLMs) capable of instruction-following (Touvron et al., 2023a,b). As LLMs become increasingly accessible to the general public, they fuel a growing demand for high-quality research in this area, which can help LatAm researchers leverage this momentum to gain more visibility.

Despite the growing importance of computational linguistics (CL), there is a notable lack of research on the presence and impact of LatAm researchers in this field. Currently, there is only a single study examining the situation in Brazil (Pardo et al., 2010) and a publication analyzing the global overview (Rungta et al., 2022), leaving a significant knowledge gap regarding the participation of LatAm researchers in CL. To address this gap and inform strategic decisions, it is essential to comprehensively measure the level of involvement of LatAm researchers in CL. This research can provide valuable insights into the current underrepresentation of LatAm in CL, facilitate the identification of factors contributing to this issue, and ultimately inform evidence-based interventions to promote greater representation, diversity, and inclusivity within the CL community.

This study aims to quantify the presence of LatAm researchers in the CL community, with the ultimate goal of enhancing their visibility and

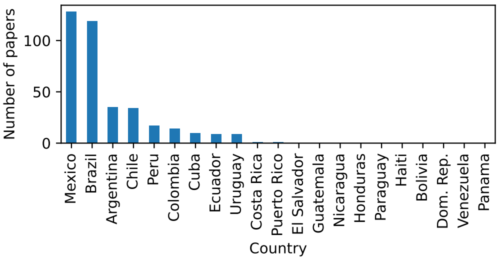

Figure 1: Distribution of LatAm publications in CL.

Country	Researcher	Pub.
Mexico	Alexander Gelbukh	12
Argentina	Luciana Benotti	12
Brazil	Thiago Salgueiro	10
Chile	Jocelyn Dunstan	8
Uruguay	Luis Chiruzzo	8
Colombia	Fabio González	6
Cuba	Suilán Estévez-Velarde	6
Peru	Arturo Oncevay	4
Ecuador	Josafá Aguiar	2
Costa Rica	Guillermo González	1
Puerto Rico	Manuel Pérez-Quiñonez	1

Table 1: Most productive LatAm researchers in CL.

representation. We comprehensively analyzed CL publications, estimating the number of publications by LatAm researchers country-by-country and displaying the people making CL research possible in LatAm.

2 Data and methods

We constructed a metadata dataset comprising all publications in the ACL Anthology repository, a comprehensive database of computational linguistics (CL), and natural language processing literature comprising main conference and workshop papers (Gildea et al., 2018). To gather author data for each article, we linked this dataset with the OpenAlex database, a large-scale bibliographic catalog of scientific papers (Priem et al., 2022), enhancing the metadata with detailed author information. We queried both sources on 2024-08-30.

To compile the list of LatAm publications, we retrieved the works where at least one of the authors was affiliated with an LatAm institution at the time of publication. The list of LatAm countries was extracted from Wikidata and is shown in Figure 1 (Wikidata, 2024). Finally, we summarized the results with the number of publications per country and created a list of the most productive authors per country.

3 Results

Our analysis revealed that the metadata dataset contained 40,997 publications, of which 978 originated from LatAm-affiliated researchers, constituting a modest 2.4 % of the total publication volume (starting from around 2014; check the Limitations paragraph). This finding highlights the underrepresentation of LatAm researchers in the CL field, with LatAm-affiliated authors contributing only a small fraction of the overall output. This disparity underscores the pressing need to enhance the presence and representativeness of LatAm researchers in CL research.

Figure 1 shows the distribution of the publications by country, and Table 1 highlights the most productive researchers by country. In the poster form of this publication, we will compile and show photographs of each of the most productive researchers (after authorization) to increase the visibility of LatAm researchers. According to the distribution of publications, there is an evident inequality where some of the largest countries accumulated most of the publications, whereas others did not publish any work. Regarding the most productive researchers, there is a significant gender difference, where only three of the most productive researchers are women.

4 Conclusion

Our study reveals a pressing concern regarding the underrepresentation of LatAm researchers in the CL community. Our analysis highlights that some countries in the region have not contributed to any CL publications, and a significant disparity exists in the number of publications per country. Despite these challenges, our results also showcase talented researchers who have significantly contributed to the field.

By shedding light on the underrepresentation of LatAm researchers in CL, this study aims to promote greater visibility and inclusivity within the community. We hope these findings will catalyze change, encouraging institutions and organizations to support researchers from underrepresented regions.

Limitations The data sources may be incomplete or inaccurate. To join the ACL Anthology and OpenAlex publications, they must have a DOI assigned; however, not all publications have a DOI. Furthermore, the affiliations extracted from OpenAlex may be incomplete or incorrect, as the database uses heuristics to extract this information.

References

Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. 2018. The ACL Anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia. Association for Computational Linguistics.

José Torres Ortega, Jorge Ortega De La Rosa, Esperanza Díaz Arroyo, and Luis Alberto Bolaño Melo. 2022. Education, research, and development expenditure is the best way to competitiveness—a panel data approach for latin american countries. *Procedia Computer Science*, 203:651–654.

Thiago Pardo, Caroline Gasperin, Helena de Medeiros Caseli, and Maria das Graças Nunes. 2010. Computational linguistics in Brazil: An overview. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 1–7, Los Angeles, California. Association for Computational Linguistics.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *Preprint*, arXiv:2205.01833.

Mukund Rungta, Janvijay Singh, Saif M. Mohammad, and Diyi Yang. 2022. Geographic citation gaps in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1371–1383, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

United Nations. 2024. Human Development Index. `https://hdr.undp.org/data-center/human-development-index`. [Accessed 06-09-2024].

Wikidata. 2024. Latin america (q12585). [Online; accessed 7-September-2024].