Revisiting Joshi et al. (2020): Methodological Gaps in the Analysis of Language Resource Distribution and NLP Conference Inclusion

Polina Kuznetcova* **and Annika Spirgath***
Institute of Computational Linguistics
Heidelberg University
{kuznetcova,spirgath}@cl.uni-heidelberg.de

Abstract

Despite over 7,000 languages existing worldwide, NLP research remains concentrated on a small, privileged subset. The influential taxonomy of Joshi et al. (2020) has shaped how this imbalance is understood — yet its methodology contains significant inconsistencies that distort the picture. In this paper, we identify and address three key shortcomings: the absence of standardized language naming, inconsistent treatment of macrolanguages versus individual varieties, and an imprecise measure of language inclusion in NLP conferences. We introduce `ISO639-3`-based standardization across all language catalogs, a principled two-tier approach to macrolanguage handling, and a refined method that distinguishes languages genuinely researched from those merely mentioned. Applied to all major NLP conference papers through 2024, our corrected taxonomy reveals a more balanced resource distribution than previously reported, with 162 languages shifting to higher-resource classes. Nevertheless, 96% of the world's languages remain without any computational resources, reinforcing the urgency of the language gap. Our methodology and data are fully reproducible, providing a stronger foundation for future work on linguistic equity in NLP.

1 Background and Motivation

There are more than 7,000 languages spoken worldwide, yet NLP research and current large language models (LLMs) have concentrated on only a small subset of them. This imbalance is what Peppin et al. (2025) call the "language gap," a divide that limits the scientific reach of NLP and reinforces existing inequalities. Models trained primarily on English and other Western-centric data reproduce these cultural perspectives, performing poorly, and at times unsafely, for other languages, generating harmful, biased, or misleading outputs where training representation was lacking (Peppin et al., 2025).

This disparity is not limited to model quality, but also affects the cost of deployed systems. Ahia et al. (2023) show that commercial LLM APIs charge users different prices depending on language, since non-Latin-script languages require up to five times as many tokens to convey the same information as Latin-script languages. The authors attribute this gap to tokenizers being trained on data dominated by Latin-script, primarily English text. As a result, speakers of many under-resourced languages simultaneously pay more and receive worse performance.

Ensuring broader language coverage in NLP is therefore not only a technical challenge but an ethical responsibility, determining who benefits from AI and whose voices are left out. Addressing this responsibility, however, first requires an accurate picture of the problem: quantifying how (un)evenly resources and research attention are distributed across languages.

1.1 Related Work

Several studies have attempted to quantify the representation of languages in NLP.

Blasi et al. (2022) measured the gap between the global demand for language technology and how well that demand is actually met, combining demand indicators (speaker population, economic factors) with utility indicators (task performance). They found resources concentrated in a small subset of languages and concluded that this concentration is driven more by economic factors than by demographic demand, highlighting a systemic bias in language technology development.

Ranathunga and de Silva (2022) extend this line of analysis through language resources, NLP publications, multilingual web platforms, and pretrained multilingual models. Rather than classifying languages solely by resource availability, they incorpo-

*Equal contribution.

rate a categorization based on speaker population and vitality, showing that disparities persist even within language groups and are linked to family, region, speaker population, and economic factors.

Joshi et al. (2020) introduced an influential six-level taxonomy that categorizes languages by the availability of labeled and unlabeled resources, from resource-rich languages such as English (class 5) to extremely low-resource languages (class 0). Applying this taxonomy to major NLP conference papers, they found large imbalances: a few high-class languages receive most of the research focus, while many others with comparable speaker populations are ignored.

1.2 Limitations of Joshi et al. (2020)

We identified several areas in Joshi et al. (2020) that required clarification or improvement.

First, the study did not apply a unified naming convention across resources. This led to duplicate entries for the same language under different names (e.g. `Sinhala` vs. `Sinhalese`).

Second, their treatment of macrolanguages versus individual languages appeared inconsistent (this distinction is explained in Section 2.2). While resources in catalogs such as LDC and ELRA are sometimes listed at the macrolanguage level (e.g. `Arabic (macro)`) and at the level of specific varieties (e.g. `Arabic, Egyptian; Arabic, Gulf; Arabic, Levantine; Arabic, Mesopotamian`), Joshi et al. (2020) seemed to primarily count macrolanguages, as most of the documented varieties are missing in their language taxonomy. At the same time, some varieties, such as `Arabic, Moroccan`, are present in the data set, likely because they are represented as separate Wikipedia languages. Consequently, while all Wikipedia language editions were counted as sources of unlabeled data, labeled data was restricted to macrolanguages, producing an uneven representation of language families with multiple varieties.

Third, to evaluate the inclusion of certain languages in NLP conferences over the years, Joshi et al. (2020) checked only whether a language is mentioned at least once in a paper. This produces false positives, since a mention does not necessarily indicate that the paper conducts research on that language, particularly if it appears only in the related work section.

Finally, the class-wise inverse Mean Reciprocal Rank (MRR) metric they use to quantify the inclusion of each resource class is not comparable across venues, since a language's rank depends on how many languages a venue mentions overall. We revisit this issue in detail in Section 3.2.3.

1.3 Our Contribution

Our study builds on the framework of Joshi et al. (2020); however, a strict replication is not possible because the original data sets and collection procedures were never made publicly available. As a result, the widely cited six-level taxonomy has thus far lacked a fully verifiable empirical foundation.[1] Our contribution is therefore primarily methodological, with new empirical findings emerging as a consequence of the corrected approach.

Methodologically, we make five improvements. We (i) standardize language identification via `ISO639-3`, resolving naming inconsistencies that caused duplicate or missing entries; (ii) apply a principled method to distinguish between macrolanguages and individual varieties, reducing the inconsistencies that affected earlier counts; (iii) create a reproducible pipeline to collect data from the ELRA and LDC catalogs, making our resource counts more transparent;[2] (iv) introduce a targeted language-mention detection algorithm that distinguishes languages genuinely researched from those merely mentioned; and (v) adopt a class-wise share-based measure that, unlike inverse MRR, is directly comparable across venues.

Applying this corrected methodology on data through 2024, we obtain a substantially revised taxonomy and updated findings on conference language inclusion.

2 The Six Language Classes

2.1 Data Collection

As in Joshi et al. (2020), we draw on ELRA[3] and LDC[4] for labeled data and on Wikipedia for unlabeled data, treating each catalog or Wikipedia page as a single data point per language. Since the original study's own data and collection scripts were not available to us (see Section 1.3), we independently collected our own data from these same sources. Consequently, our counts reflect the state of the re-

[1] We contacted the corresponding author of Joshi et al. (2020), who kindly clarified several methodological points (e.g. the treatment of macrolanguages and the use of Ethnologue as the base language list), but the original data sets and code were no longer available.

[2] Code and data: GitHub repository.

[3] `http://catalog.elra.info/en-us/`

[4] `https://catalog.ldc.upenn.edu/`

sources at the time of collection (2025), including additions made since the 2020 study.

Concretely, we obtained ELRA catalog counts per language directly from the ELRA website. LDC does not provide catalog counts by language, so we extracted this information by scraping the LDC catalog ourselves. For unlabeled data, we retrieved Wikipedia page counts per language directly from the Wikipedia website.[5] Finally, speaker numbers for each language were integrated into the data sets, obtained from Wikidata using a publicly available SPARQL query script[6] on GitHub. Before proceeding to data analysis, all language names in our own collected data were standardized and processed to ensure consistency across sources.

2.2 Macrolanguages and ISO639-3 Standardization

During the examination of the LDC and ELRA catalogs, we observed that some resources are listed as macrolanguages (e.g. `Arabic (macro)`) while others are associated with specific varieties (e.g. `Arabic, Egyptian; Arabic, Gulf; Arabic, Levantine; Arabic, Mesopotamian`). This raises the fundamental question of what is to be considered a "language" in the first place.

According to `ISO639-3`, an international standard based on *Ethnologue* and seemingly followed by ELRA and LDC, languages are classified into macrolanguages and individual languages — a distinction not addressed by Joshi et al. (2020). A **macrolanguage** is defined as a set of closely related speech varieties that may not share a single traditional name, and which are often perceived as distinct languages by their users. In contrast, **individual languages** either form part of a macrolanguage or exist independently (ISO 639-3, 2025).

The mapping of macrolanguages is inherently fuzzy, as they may group varieties that are treated as dialects, closely related but distinct languages, or even a single language for primarily cultural or political reasons rather than linguistic ones (Morey et al., 2013). Due to this fuzziness, we decided to include macrolanguage counts and disregard their constituent varieties whenever possible. Additionally, resources on ELRA and LDC are sometimes listed solely under a macrolanguage label without specification of constituent varieties, making it impossible to rely exclusively on individual languages. As a result, we implemented a two-tiered approach.

Tier 1: Handling macrolanguages at catalog level

1. Check whether an entry with the tags (macro) or (macrolanguage) exists for a given language (e.g. `Arabic (macro)`);
2. If a macro count is given, take only this count and disregard individual varieties (e.g. `Arabic (Moroccan), Arabic (Tunisian)`);
3. If no macrolanguage is defined, count the individual language itself.

However, relying solely on macrolanguage tags fails in cases where related languages appear under different orthographic spellings or catalog conventions. For example, `Indonesian` is listed independently in ELRA, despite being classified as part of the Malay macrolanguage under `ISO639-3`. To ensure consistent counting, all language names were converted to `ISO639-3` in Tier 2.

Tier 2: ISO639-3 standardization

4. Replace language names with `ISO639-3` language tags;
5. Revise and manually fix problematic cases;
6. Check if any language IDs belong to a macrolanguage group;
7. If so, replace with the macrolanguage ID.

The results are summarized in Table 1. The list of problematic cases is available in Appendix A.

Step	ELRA	LDC	Wikipedia
After primary data collection	92	122	343
After disregarding varieties, no standardisation	91	96	342
After disregarding varieties, with standardisation	86	96	313

Table 1: Number of languages after each step across different catalogs.

Language name standardization reduced the number of languages by approximately 13% on average, while ensuring naming consistency and eliminating duplicates. Although Tier 1 steps could be considered redundant given Tier 2 standardization, we deliberately preserved them to highlight the discrepancies that arise when applying different counting methodologies and to underscore the critical role of standardization.

[5] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[6] https://gist.github.com/unhammer/3e8f2e0f79972bf5008a4c970081502d

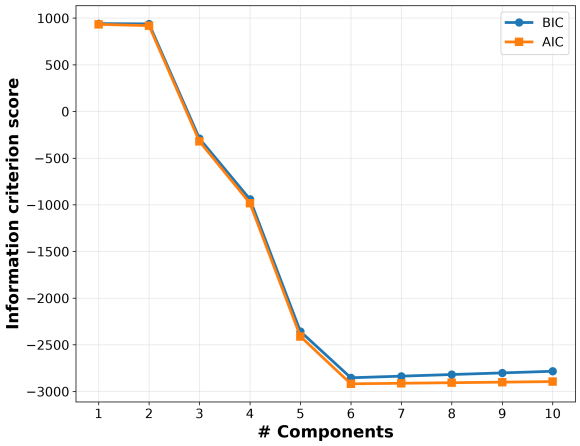

Figure 1: Clustering of the data set using a Gaussian Mixture Model (GMM).

2.3 Two Data Sets

The taxonomy of Joshi et al. (2020) included more than 2000 languages, while the number of languages with any linguistic resources in the listed catalogs did not exceed 350. How they arrived at their number remains unclear. To address this ambiguity, we analyzed **two data sets**: one comprising **only languages with available linguistic resources**, and one including **all languages** listed in *Ethnologue*. For the latter, the full language list was standardized following the same procedure, and languages with no available resources were assigned a value of 0 for both labelled and unlabelled sources.

2.4 Determining the Six Language Classes

A methodological question that Joshi et al. (2020) leave unaddressed is how the number of classes and the boundaries between them were originally chosen. The paper does not report a selection procedure or criterion for this choice, so it is unclear whether six classes reflect a principled cutoff or a more informal decision.

To verify that no alternative cluster solution was more appropriate, we conducted a Gaussian Mixture Model (GMM) analysis (Figure 1). Both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) converged on six clusters as the optimal solution (Figure 1). We take this as independent quantitative support for retaining Joshi et al. (2020)'s original six-class structure, even though we cannot verify whether their choice of six was reached through a comparable procedure.

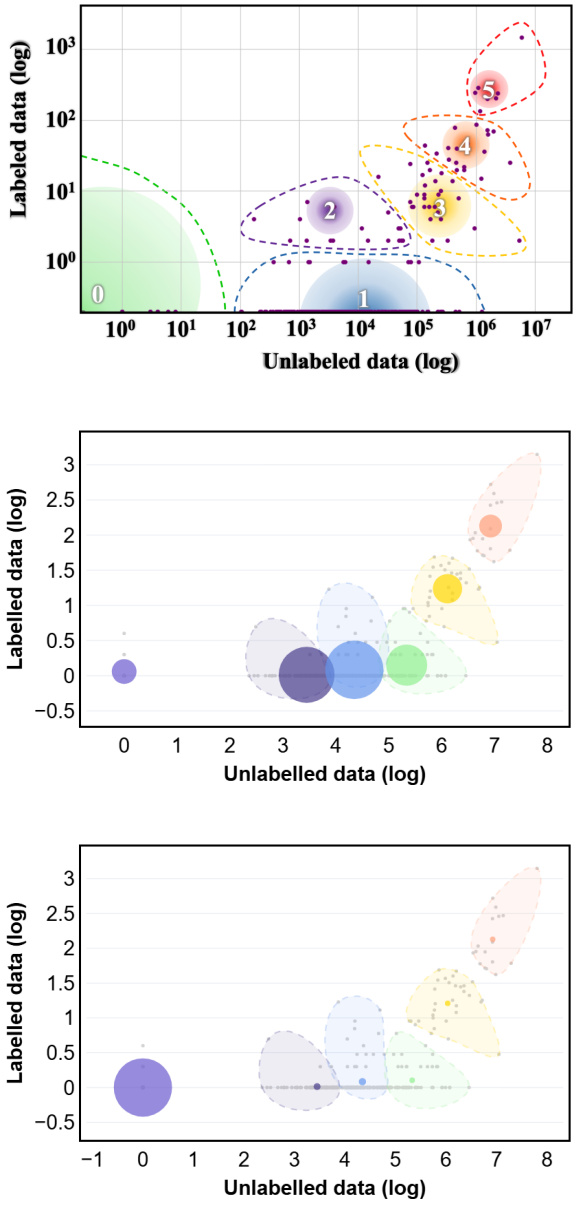

Figure 2: Language resource distribution in Joshi et al. (2020) taxonomy (top), our taxonomy of languages with resources (middle), and our taxonomy with all languages (bottom). The size of the gradient circle represents the # of languages in the class. The color spectrum VIBGYOR represents the total language user population size from low (violet) to high (red). Bounding curves demonstrate points covered by a language class.

Class	5 Example Languages	Languages with Resources			All Languages		
		#Langs	#LangUsers	% of Total	#Langs	#LangUsers	% of Total
0	Filippino, Luo, Khasi, American Sign Language, Middle English	20	6M	6.01%	7166	602.7M	95.81%
1	Hawaiian, Kabardian, Tahitian, Church Slavic, Farefare	102	354.7M	30.63%	102	354.7M	1.36%
2	Navajo, Neapolitan, Amharic, Turkmen, Sardinian	111	617.4M	33.33%	110	608.3M	1.47%
3	Malayalam, Nepali, Telugu, Afrikaans, Sinhala	55	735.1M	16.52%	53	683.8M	0.71%
4	Finnish, Czech, Norwegian, Turkish, Thai	28	773.7M	8.41%	31	834.1M	0.41%
5	English, Spanish, French, German, Arabic	17	4.6B	5.11%	17	4.6B	0.23%

Table 2: Number of languages, language users, and percentage of total languages for each language class in our taxonomies.

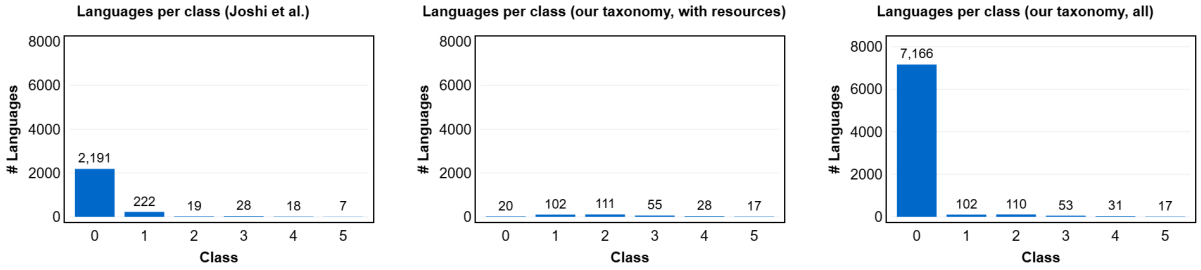

Figure 3: Language # per class in Joshi et al. (2020) taxonomy (left), our taxonomy of languages with resources (middle), and our taxonomy with all languages (right).

2.5 Results

Results are visualized in an interactive Streamlit dashboard.[7]

2.5.1 General Observations

Overall, our taxonomies show patterns broadly similar to those in Joshi et al. (2020)'s study (Figure 2). The same trend holds: resources are distributed in a highly skewed fashion, with only a small set of high-resource languages occupying the top-right corner, and the vast majority of languages concentrated below. Class 0, representing languages with negligible or no computational resources, remains far behind the remaining classes. This consistency confirms that the fundamental finding of Joshi et al. (2020) remains valid.

Notably, several languages in Classes 1 and 2 appear to have more labeled sources than in the original taxonomy. Possible causes may be either methodological differences in resource counting or genuine resource growth over time.

2.5.2 Comparison with Joshi et al. (2020) (Resources Only)

Reduction of Class 0. Only 20 languages (6.01%) fall into Class 0, compared with 2191 (88%) in the original taxonomy (Table 3). This discrepancy suggests that Joshi et al. (2020) may have included a large number of languages with no resources in their taxonomy, although this was never made explicit.

More balanced mid-resource classes. Most languages are distributed across Classes 1–3 (84%), with the largest clusters being Class 1 (102 languages, 30.63%) and Class 2 (111 languages, 33.33%) (Table 2). This contrasts with Joshi et al. (2020)'s taxonomy, where Classes 0 and 1 contain the absolute majority of all languages (Figure 3). This observation is most evident in Figure 4.

Less pessimistic outlook overall. Our Class 5 includes 17 languages — more than double the number in Joshi et al. (2020) (Table 3). The original "Winners" (English, Spanish, German, Japanese, French, Arabic, and Mandarin) were joined by Italian, Polish, Portuguese, Dutch, Korean, Vietnamese, Swedish, Serbo-Croatian, Russian, and Persian in our taxonomy.

2.5.3 Comparison with Joshi et al. (2020) (All Languages)

Class 0 dominance. Both taxonomies show that the absolute majority of languages have negligible or no computational resources. In our taxonomy, "The Left-Behinds" contain 7166 languages (95.81%), compared to 2191 (88%) in Joshi et al. (2020)'s study (Table 3).

High-resource languages remain few. Class 5 remains small (17 languages, 0.23%), similar to Joshi et al. (2020)'s statistics (7 languages, 0.28%) (Table 2).

[7] https://languagetaxonomy.streamlit.app/

Class	Joshi et al. (2020)	Our (Resources)	Our (All Languages)
0	2191	20	7166
1	222	102	102
2	19	111	110
3	28	55	53
4	18	28	31
5	7	17	17
Total	**2485**	**333**	**7479**

Table 3: Language # per class across taxonomies.

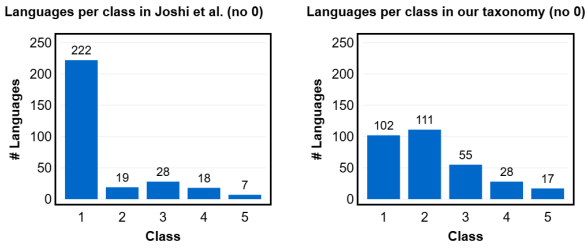

Figure 4: Language # per class in Joshi et al. (2020)'s taxonomy and our taxonomy after removing Class 0.

Resource inequality between languages and language users. From a language user perspective, the majority of people are concentrated in the well-resourced Class 5 languages (4.6B language users across only 17 languages) (Table 2).

2.6 Cluster Shifts

Our taxonomy comprises 333 languages, compared to 2485 in Joshi et al. (2020). Of these, 233 languages are common to both taxonomies. We attempted to standardize Joshi et al.'s language names to maximize matches, as described in Appendix B.

Notably, 162 languages experienced cluster changes between the two studies (Table 4). Specifically, 157 languages shifted upward (average shift of 1.37), while only 5 shifted downward. The predominance of upward shifts may indicate either methodological differences or genuine resource growth over the five-year period separating the two studies.

3 Conference-Language Inclusion

NLP conferences play an important role in influencing which languages receive attention and resources. We investigate whether major NLP conferences have become more inclusive of a broader set of languages, extending Joshi et al. (2020)'s work to all NLP conference papers through 2024 and introducing a more accurate method for identifying languages that are actually studied.

Direction	Count	Avg. Shift	% of Langs
Shifted Up	157	1.37	96.91%
Shifted Down	5	1.00	3.09%

Table 4: Cluster shifts between our taxonomy and Joshi et al. (2020).

Venue	Total Papers	From WS	Excluded	Papers Used
ACL	14046	2912	5 (0.0%)	14041
NAACL	5050	629	15 (0.3%)	5035
EMNLP	10549	1374	32 (0.3%)	10517
EACL	2802	507	5 (0.2%)	2797
TACL	713	0	0 (0%)	713
CONLL	10253	8983	103 (1.0%)	10150
COLING	7543	1398	13 (0.2%)	7530
LREC	9954	863	1078 (10.8%)	8876
SEMEVAL	2982	0	1 (0.0%)	2981
CL	2036	0	10 (0.5%)	2026
WS_other	7459	7459	154 (2.1%)	7305
Total	73387	24125	1416 (1.9%)	71971

Table 5: Data set statistics across all conferences.

3.1 Data Set and Statistics

We base our data set on the ACL Anthology Reference Corpus (Bird et al., 2008), which provides PDFs and metadata for conference papers. Our data set covers the same 10 conferences used by Joshi et al. (2020): ACL, NAACL, EMNLP, EACL, COLING, LREC, CONLL, SEMEVAL, TACL, and CL journals. We extend their data set by including all papers available in the ACL Anthology up to and including 2024, restricted to main-track proceedings (long and short papers only). Unlike Joshi et al. (2020), who treat every conference-affiliated workshop as a single "Workshops" category, we instead attribute each workshop paper to its host conference wherever the ACL Anthology records an explicit co-location. Workshops not co-located with any of our 10 venues are grouped into a WS_other category. Statistics are summarized in Table 5.

3.2 Analysis

3.2.1 Measuring Language Mentions

Joshi et al. (2020) measured language representation by checking whether a language appears anywhere in a paper — an approach that produces many false positives. To improve accuracy, we implement a more targeted method:

For each paper, we parse the abstract for language mentions and look for patterns indicative of cross-linguistic or multilingual research (keywords such as *typological*, *multilingual*, *cross-lingual*, *low-resource languages*, or numeric expressions like "(d+) languages"), tolerant of common spac-

ing and hyphenation variants (e.g., "multi-lingual," "cross lingual"). If such patterns are found, we additionally parse the first four pages of the full PDF for further mentions. We match language names using FlashText (Singh, 2017), a fast keyword-extraction library. When multiple candidate matches overlap at the same position (e.g., *German Sign Language* and *German*), it resolves them by preferring the longest one. In cases where our algorithm was not able to find any language mentions for a paper, we assume that the research focuses on English by default. Lastly, if the PDF could not be retrieved or parsed, or if no usable text could be extracted, the paper is excluded. See Table 5 for the exclusion statistics.

Overall, this procedure provides a more reliable measure of which languages are actively investigated rather than merely mentioned.

3.2.2 Language Occurrence Entropy

To measure how inclusive a conference is with respect to the languages in our taxonomy,[8] we adopt the language entropy measure from Joshi et al. (2020):

$$S = -\sum_{j=1}^{L} p_j \log p_j,$$

where p_j is the proportion of papers that mention language j, and L is the total number of languages considered. Results are shown in Figure 5.

3.2.3 The Problem of Class-wise Inverse MRR

Following Joshi et al. (2020), we compute class-wise inverse Mean Reciprocal Rank (MRR) for each venue and resource class:

$$\frac{1}{\text{MRR}} = \left(\frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \right)^{-1},$$

where Q is the set of languages belonging to a given class, and rank_i is language i's position in a single frequency-based ranking shared across all languages mentioned at that venue.

Rank is not comparable across venues. Since rank_i is a language's position in one ranking shared across all languages mentioned at a venue, its possible range grows with the number of unique languages that venue mentions. High-resource languages are rarely affected by this, as they are almost always ranked near the top regardless of how many languages a venue mentions overall. However, low-resource languages, are pushed further down the ranking as venues mention more languages, inflating their inverse MRR values. Appendix D confirms that this inflation is significant for low-resource classes. We therefore adopt a size-independent alternative for comparing venues.

3.2.4 Share of Language Mentions by Class

To allow direct comparison across venues, we compute, for each venue and class C, the share of that venue's language mentions belonging to C:

$$\text{Share}_C = \frac{\text{mentions in class } C}{\text{total language mentions in the venue}}.$$

Bounded in $[0, 1]$, this share is directly comparable across venues regardless of size. Results are shown in Figure 6.

3.3 Findings

3.3.1 Overall Language Occurrence Entropy

Our refined measurement generally resulted in lower language occurrence entropy compared to Joshi et al. (2020) (Figure 5), as expected: our approach reduces false positives by excluding superficial language mentions.

3.3.2 Conference Language Inclusivity over Time

Consistent with Joshi et al. (2020), all conferences show an overall rise in language inclusivity over time, although with substantial year-to-year fluctuation. Among the longer-established venues, this rise is most pronounced for ACL, EACL, COLING, and WS_other, which move from some of the lowest entropy values in early years to among the highest in recent years. In contrast, the more recently founded SEMEVAL and LREC begin around the year 2000 already at moderate entropy and continue to increase, suggesting they benefited from the inclusivity norms set by their predecessors. Nevertheless, since 2020, many conferences (ACL, CL, EMNLP, SEMEVAL, LREC, and WS_other) show a mild decline, possibly reflecting a shift in attention toward English-centric large language models in the field. This suggests that pre-2020 gains in language inclusivity are not yet self-sustaining, underscoring the need for continued, active attention to linguistic diversity in NLP research.

[8] For the following analysis, we rely on the full taxonomy including languages without available resources. A small number of language entries were filtered out to reduce false positives, see Appendix C.

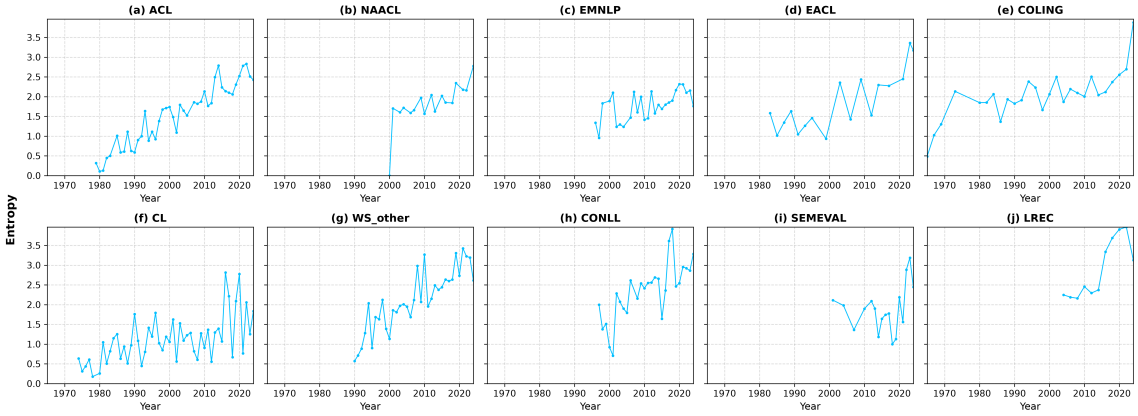

Figure 5: Trends in Language Occurrence Entropy across NLP Conferences (through 2024).

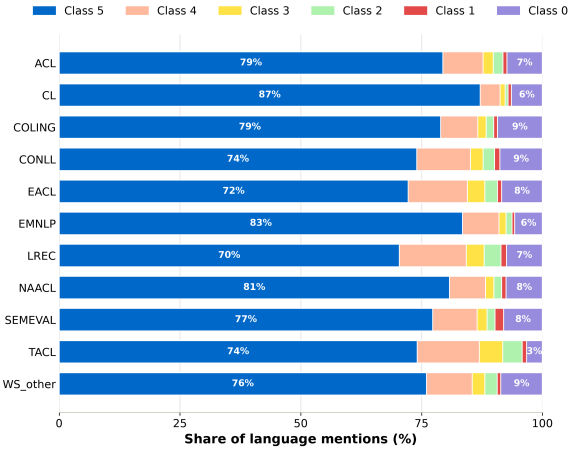

Figure 6: Class-wise share of language mentions by venue.

3.3.3 Class-wise Language Inclusion Patterns

While entropy captures the diversity of languages mentioned, it does not reflect how mentions are distributed across resource classes. Aggregated across the full period, Class 5 languages account for over 70% of language mentions at every venue, as shown in Figure 6. This indicates that overall research attention remains heavily concentrated in a small set of high-resource languages.

Among the remaining language mentions, Class 0 shows a higher share than Classes 1–3 at most venues, despite representing the least-resourced languages. This likely reflects the size of Class 0 in the taxonomy, which contains the vast majority of the world's languages, increasing the risk of more false-positive matches. Nevertheless, it may also reflect a preference for the most extremely under-resourced languages within low-resource language research.

4 Discussion and Conclusion

The analysis of our six resource-based language classes confirms that the overall patterns reported by Joshi et al. (2020) remain largely valid: resources are highly skewed, with a small number of high-resource languages dominating NLP, while the vast majority fall into low-resource classes.

Our taxonomy of languages with available resources reveals a more balanced distribution compared to the original study. Class 0 is much smaller, Classes 1–4 contain the majority of languages, and our Class 5 includes 17 languages — more than double the original count.

When considering all languages, however, 96% still fall into Class 0. This reinforces the "language gap" noted by Peppin et al. (2025) and the systemic bias documented by Blasi et al. (2022). At the same time, 157 of 162 shifted languages moved upward, suggesting a more optimistic trend than previously reported.

Extending Joshi et al. (2020)'s conference analysis through 2024 reveals that several major NLP venues exhibit a slight decline in language inclusivity after 2020, possibly reflecting the field's increasing focus on English-centric large language models. This suggests that earlier gains in language inclusivity are not yet self-sustaining and highlight the importance of actively monitoring and promoting linguistic diversity across all venues. Class-wise analysis further confirms the expected resource hierarchy, with Class 5 languages accounting for over 70% of language mentions across all conferences.

As the original data sets from Joshi et al. (2020) are unavailable, it is difficult to disentangle real progress from methodological differences. Regardless, this work introduces a more transparent and

robust methodology for such analyses.

Limitations

Several limitations of this study should be noted. First, because Joshi et al. (2020) did not release their original data sets, despite our efforts to obtain them directly from the authors (Section 1.3), a strict step-by-step replication was not possible. Our comparison therefore may conflate methodological differences with genuine changes over time.

Second, our resource counts are restricted to the three sources considered by Joshi et al. (2020) (ELRA, LDC, and Wikipedia), excluding large contemporary repositories such as HuggingFace. While this choice enables a more consistent comparison with the original work, it may not capture the full extent of current resource availability. As demonstrated by Ranathunga and de Silva (2022), incorporating additional repositories reveals substantially more resources for many languages.

Third, our decision to standardize resources at the macrolanguage level (Section 2.2) comes at the cost of losing regional-variant granularity. Since the underlying catalogs do not consistently specify the language variety covered by each resource, harmonizing counts across catalogs required collapsing varieties into their corresponding macrolanguages whenever variety-level attribution was unreliable. As a result, finer-grained distinctions between regional and dialectal variants are not reflected in our reported counts.

Fourth, our language-mention detection algorithm (Section 3.2.1) depends on abstract-level trigger keywords (e.g., *multilingual*, *cross-lingual*) to decide whether to extend parsing into the full text. This choice reduces false positives compared to scanning entire papers unconditionally, but means that papers discussing multiple languages without using these cues in the abstract may have some mentions missed, and are then defaulted to English by our fallback rule.

Fifth, our matching relies on each language's standardized ISO639-3 name, whereas research papers may refer to a language by a different name – its local name, an alternate English name, or an acronym (e.g. *Deutsche Gebärdensprache*, *German Sign Language*, or *DGS*) – so mentions using a name outside our taxonomy may be missed.

Finally, the Gaussian Mixture Model cluster boundaries (Section 2.4) are data-driven and may shift as the resource landscape evolves, making direct longitudinal comparisons harder over time.

Acknowledgments

We thank Prof. Dr. Katja Markert (Heidelberg University) for her supervision and guidance throughout this project. We also thank Pratik Joshi for his responsiveness and helpful clarifications regarding the methodology of Joshi et al. (2020).

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? Tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

ISO 639-3. 2025. Scope of denotation.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Stephen Morey, Mark W. Post, and Victor A. Friedman. 2013. The language codes of ISO 639: A premature, ultimately unobtainable, and possibly damaging standardization. Conference presentation and handout, http://hdl.handle.net/2123/9838. Presented at "Research, Records and Responsibility: Ten Years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures" (PARADISEC), University of Sydney.

Aidan Peppin, Julia Kreutzer, Alice Schoenauer Sebag, Kelly Marchisio, Beyza Ermis, John Dang, Samuel Cahyawijaya, Shivalika Singh, Seraphina

Goldfarb-Tarrant, Viraat Aryabumi, Aakanksha, Wei-Yin Ko, Ahmet Üstün, Matthias Gallé, Marzieh Fadaee, and Sara Hooker. 2025. The multilingual divide and its impact on global AI safety. *Preprint*, arXiv:2505.21344.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. 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 823–848, Online only. Association for Computational Linguistics.

Vikash Singh. 2017. Replace or retrieve keywords in documents at scale. *Preprint*, arXiv:1711.00046.

A Problematic Cases

Table 6 lists the cases that required special handling during standardization because their names did not match `ISO639-3` tags. Additionally, `Simple English`, `Berber languages`, and `Bantu languages` were removed from the data sets.

B Joshi et al.'s Taxonomy Standardisation Attempt

Joshi et al. (2020) did not standardize language names; their taxonomy entries are structured as `full language name, class number` (e.g. `língua gestual portuguesa, 0`). To improve comparability, we attempted to convert their language names into `ISO639-3` tags. Many entries proved problematic and could not be reliably mapped.

As a supplementary step, we experimented with minimum edit distance (MED) to match remaining names. In some cases this worked well (e.g. `shompen` → `Shom Peng`), but in many others it produced implausible matches (e.g. `Tasmanian` → `Dalmatian`, MED = 3; `langue des signes québecoise` → `Algerian Sign Language`, MED = 18). This method was therefore discarded. These limitations once again underscore the necessity of language name standardization for large-scale analyses.

Problematic Case	**Resolution**
Swahili	Swahili (macrolanguage)
Malay	Malay (macrolanguage)
Cantonese	Yue Chinese
Colognian	Kölsch
Dimli	Zaza
Doteli	Dotyali
Acehnese	Chinese
Alemannic	Swiss German
Arakenese	Rakhine
Aromanian	Macedo-Romanian
Bangla	Bengali
Banyumasan	Javanese
Emiliano-Romagnolo	Emilian
Frafra	Farefare
Fula	Fulah
Gan	Gan Chinese
Ghanaian Pidgin	Ghanaian Pidgin English
Greek	Modern Greek (1453–)
Guianan Creole	Guianese Creole French
Interlingua	Interlingua (IALA)
Kabiye	Kabiyè
Kyrgyz	Kirghiz
Limburgish	Limburgan
Livvi-Karelian	Livvi
Low Saxon	Low German
Māori	Maori
Mindong	Min Dong Chinese
Minnan	Min Nan Chinese
N'Ko	N'Ko
Nahuatl	Eastern Huasteca Nahuatl
Nepali	Nepali (macrolanguage)
Newari	Nepal Bhasa
Norman	Jèrriais
Northern Sotho	Pedi
Nupe	Nupe-Nupe-Tako
Chewa	Chichewa
Occitan	Occitan (post 1500)
Ossetic	Ossetian
Palatine German	Pfaelzisch
Pannonian Rusyn	Ruthenian
Pa'o	Pa'o Karen
Pashto	Pushto
Piedmontese	Piemontese
Punjabi	Panjabi
Saterland Frisian	Saterfriesisch
Tai Nuea	Tai Nüa
Tarantino	Neapolitan
Taroko	Sediq
Tongan	Tonga (Tonga Islands)
Uyghur	Uighur
Waray	Waray (Philippines)
West Flemish	Vlaams
Western Punjabi	Lahnda
Wu	Wu Chinese

Table 6: Problematic cases and their resolutions.

C Taxonomy for Conference-Language Inclusion Analysis

To reduce the risk of false-positive matches, we apply two filtering rules to the complete Ethnologue taxonomy of 7,479 languages, yielding a final taxonomy of 7,385 languages:

- All language names of two letters or fewer were removed, due to the high likelihood of matching extracted gibberish from older PDFs.
- Language names coinciding with common English words, personal names, place or institution names were removed: *e.g. even, label, mono, male, anal, fore, sake, bench, broken, thompson, she, aka, bit, day, con, duke, yale*

D Inverse MRR Results and Correlation Analysis

For completeness and comparability with Joshi et al. (2020), Table 7 reports raw inverse MRR values by resource class and venue, computed as described in Section 3.2.3. Within a venue, lower values indicate stronger inclusion, but across venues the results should be interpreted with caution. Rank is defined over a single shared ranking across all languages mentioned at a venue, so its magnitude depends on the total number of unique languages a venue mentions.

Conf/Class	0	1	2	3	4	5	#Langs	#Papers
ACL	243.4	200.2	131.5	99.1	30.4	5.4	642	14041
CL	54.7	71.3	65.7	56.4	32.0	4.9	136	2026
COLING	201.7	189.4	139.0	94.0	31.6	5.2	540	7530
CONLL	213.8	168.9	122.7	88.3	28.5	5.5	542	10150
EACL	144.9	128.6	105.2	75.8	28.1	5.4	321	2797
EMNLP	164.4	161.3	121.1	88.2	29.6	5.0	400	10517
LREC	305.0	222.7	138.9	90.1	28.2	5.2	713	8876
NAACL	148.2	148.9	109.0	90.9	30.6	5.1	371	5035
SEMEVAL	87.0	47.9	62.6	57.3	31.3	5.2	190	2981
TACL	92.8	111.4	65.7	63.8	25.4	5.4	144	713
WS_other	199.2	173.2	115.9	94.4	33.0	5.4	498	7305

Table 7: Class-wise inverse Mean Reciprocal Rank (MRR) by venue. #Langs = unique languages mentioned; #Papers = total papers in venue.

Figure 7 tests this dependence directly by plotting each venue's raw inverse MRR against its number of unique languages mentioned (#Langs), per class. Classes 0–3 correlate strongly with this count ($r = 0.88$–0.98, $p < .001$), while Classes 4–5 show no significant correlation ($r = 0.08$ and 0.33, $p > .3$), consistent with Section 3.2.3's explanation that high-resource languages stay near the top of a venue's ranking regardless of how many languages are mentioned overall. This asymmetry is why inverse MRR cannot be used to compare venues across resource classes, motivating the size-independent Share_C metric adopted in Section 3.2.4.

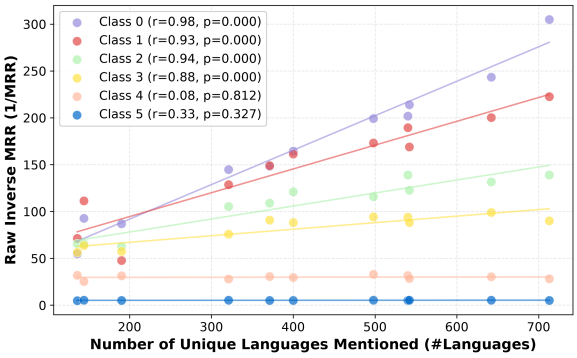

Figure 7: Raw inverse MRR by resource class vs. number of unique languages mentioned at each venue. Pearson correlations and p-values are labeled per class; trend lines shown.