A survey of diversity quantification in natural language processing: The why, what, where and how

Louis Estève
France
Univ. Paris-Saclay, CNRS, LISN

Marie-Catherine de Marneffe
Belgium
FNRS - UCLouvain

Nurit Melnik
Israel
The Open University of Israel

Agata Savary
France
Université Paris-Saclay, CNRS, LISN

Olha Kanishcheva
Germany
Heidelberg University

Relevant UniDive working groups: WG4

1 Introduction

Diversity has been gaining increasing attention in NLP in recent years (Figure 1). It has become an advocated property of datasets and systems in various NLP tasks and end-user applications, and many measures are used to quantify it. Nevertheless, there have been very few attempts to take a step back and understand the conceptualization of diversity in NLP and the motivations behind its endorsement. When such attempts were made, they were limited to particular areas (Tevet and Berant, 2021; Yang et al., 2025; Zhang et al., 2025) and diversity aspects (Lion-Bouton et al., 2022; Ploeger et al., 2024). Overall, NLP belongs to the "fields [...] where diversity is prominent in discussion, but remains undefined or analytically neglected" (Stirling, 2007, p. 707). The objective of this survey is to pave the way toward addressing these shortcomings by taking inspiration from studies outside of NLP where diversity has been systematically analyzed (Sarkar, 2010; Stirling, 1994). Our contributions are twofold: **an NLP-specific framework** for conceptualizing diversity quantification and **recommendations** for further conceptualization of diversity in the field.

This abstract summarizes our paper under review, also published as a pre-print (Estève et al., 2026).

2 Diversity in NLP: Four perspectives

We downloaded all papers from the ACL Anthology (2019-01-01 to 2024-07-26) that include "diverse" or "diversity" in their title. Our preliminary review of the 269 papers revealed several recurring issues in how diversity is conceptualized and measured. These observations motivated a more systematic investigation of the 188 papers in which diversity is formally quantified. Our analysis builds upon Stirling's (2007) unified cross-disciplinary framework for characterizing diversity, which views diversity as a property of a system whose *elements* can be apportioned into *categories* and which distinguishes three dimensions of diversity: *variety*, *balance*, and *disparity*. Our NLP-specific framework for conceptualizing diversity quantification is organized around four perspectives: *why*, *what*, *where*, and *how*. Seven researchers participated in the annotation of these papers. To ensure annotation quality, adjudication was conducted for challenging cases.

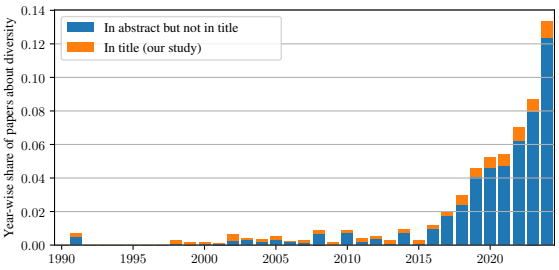

Figure 1: ACL Anthology papers from 1990 to 2024 with "diversity" or "diverse" in their title or abstract.

2.1 *Why* diversity is important in NLP

A large majority of papers in our corpus endorse diversity but do not justify this stand. Following Sarkar (2010), we uncover the normative assumptions behind the search for diversity. We find **ethical normative assumptions**, which focus on equality and inclusiveness (Joshi et al., 2020; Khanuja et al., 2023; Liu et al., 2024), protection of users, (Song et al., 2024; Yang et al., 2024), educational quality (Hadifar et al., 2023). and methodological rigor (Chen et al., 2023; Pradhan et al., 2022).

We also find **practical normative assumptions** e.g. meeting user expectations (Park et al., 2023), user engagement, (Akasaki and Kaji, 2019; Kim et al., 2023), improving performance in parsing (Liu and Zeldes, 2023), question answering (Yadav

et al., 2024), semantic role labeling (Tripodi et al., 2021), solving math problems (Shen et al., 2022), natural language generation (Zhang et al., 2021; Thompson and Post, 2020; Palumbo et al., 2020).

2.2 *What* objects are measured for diversity

The papers in our corpus address a wide range of "diversities" across NLP areas. **In-text categories** are associated with linguistic properties that are implicitly assumed to be inherent to a text. They are addressed and quantified in 119 papers in our dataset (e.g. Guo et al. (2024); Gao et al. (2019); Tevet and Berant (2021); Han et al. (2022)). **Meta-linguistic categories** relate to the classifications of texts within datasets. We found them in 51 papers. Multilingual datasets are often evaluated based on the languages they contain (e.g. Pouran Ben Veyseh et al. (2022); Sarti et al. (2022)). A third type of categories, addressed in 24 papers, are associated with the processes applied to the data. Thus, **processing categories**, such as annotations, annotators and training models, are external to the text itself (e.g. Weerasooriya et al. (2023); Parrish et al. (2024); Creanga and Dinu (2024); Greco et al. (2022); Kobayashi et al. (2022)).

2.3 *Where* diversity is measured

The *where* perspective addresses the NLP subfield that the paper targets and the pipeline stage where diversity is quantified. For the subfields, we used the main NLP areas defined in ARR.[1] The most frequent ARR areas are Machine Learning for NLP (31 papers), Generation (27), and Resources and Evaluation (27), followed by Dialogue & Interactive Systems (19) and NLP Applications (16). For the pipeline stages, Figure 2 shows the different stages of a standard machine learning pipeline in which we found diversity quantification. Output data is by far the most frequent stage (86 papers), followed by Data collection (37), while Annotation and Evaluation are rarely concerned with diversity quantification. The imbalance between ARR areas is not independent from the imbalance of the pipeline stage that diversity is quantified in. Some prototypical scenarii emerge, one notable is Generation being associated with quantification of diversity in the Output stage.

[1] https://aclrollingreview.org/cfp. The ARR area the paper was submitted to was unavailable, but we used the one in the title if present and our best guess if not.

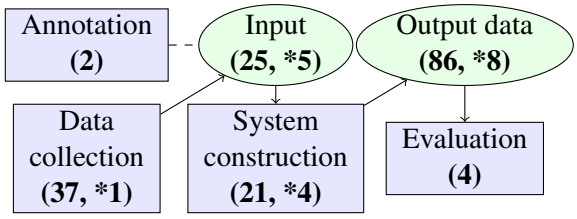

Figure 2: High-level stages in standard machine learning pipelines (rectangles for processes, ellipses for states), with numbers of papers quantifying diversity. Papers with two-stage quantification indicated with *.

2.4 *How* diversity is measured

Looking at all 188 papers quantifying diversity, we found more than 100 different names of diversity measures. Most are used in a handful of papers, making straightforward comparisons hard. Applying Stirling's (1994) framework, around 40 measures can be mapped onto the three dimensions of *variety*, *balance*, and *disparity*: richness (77 papers) and average pairwise distances (52 papers) are by far the most frequently used, while pure balance measures are rare (3 papers).

A further 35 measures cannot be cast into Stirling's framework. These include type-token ratios, which are not monotonic to the number of categories (43 papers); relative overlaps (9 papers); relative distributions, which quantify relative rather than absolute diversity (8 papers); and human judgment, which does not rely on the element/category dichotomy (12 papers). Some of these measures are likely to measure something other than diversity (at least in its cross-disciplinary understanding).

3 Discussion and recommendations

The state of NLP with respect to diversity quantification, in light of our survey, is characterized by a proliferation of measures and a lack of conceptual consensus – what Hurlbert (1971) called a *nonconcept* in the context of ecology. Yet some patterns emerge: our analysis reveals two prototypical scenarios. The first concerns corpus creation papers, which, motivated by ethical normative assumptions, quantify the meta-linguistic diversity of collected data using richness measures. The second concerns generation papers, which, motivated by user expectations, measure the in-text diversity of system outputs using Type-token ratio (TTR) or pairwise distance measures. Bringing order to this diversity of practices calls for a shared framework.

A few efforts, like Tevet and Berant (2021); Lion-Bouton et al. (2022); Yang et al. (2025); Zhang et al. (2025); Guo et al. (2025) and ours, take first steps towards comparability of different measures.

As a concrete step forward, we suggest that new papers addressing diversity quantification position themselves on the *why*, *what*, *where* and *how* perspectives, as defined in this survey, and consider answering the following questions: (i) Is diversity increase endorsed in your paper and if so, what are the motivating normative assumptions behind this stand? (ii) What are the categories and elements on which diversity is measured? (iii) At which stage of the processing pipeline do you quantify diversity? (iv) What diversity measure did you select? What is its relation to other existing measures, and to variety, balance and disparity? (v) If balance is concerned, how to characterize the distribution of the elements into categories? (vi) If disparity is concerned, what is the underlying distance measure and the distance aggregation method? (vii) What are your reasons for choosing this particular measure and why is it a diversity measure (and not a performance measure for instance)?

Acknowledgements

This research was funded by the CA21167 COST Action UniDive as well as the "*Plan Blanc*" (White Plan) doctoral grant from Université Paris-Saclay and a grant from the Israeli Ministry of Science and Technology (grant No. 0002336; Nurit Melnik, PI). Marie-Catherine de Marneffe is a research associate of the Fonds de la Recherche Scientifique - FNRS. We thank Giedre Valunaite Oleskevicienė and Irina Lobzhanidze for their participation to annotation, as part of the WG4 group of UniDive. This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD010615876 made by GENCI.

References

Satoshi Akasaki and Nobuhiro Kaji. 2019. Conversation initiation by diverse news contents introduction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3988–3998, Minneapolis, Minnesota. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2023. UniSumm and SummZoo: Unified model and diverse benchmark for few-shot summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12833–12855, Toronto, Canada. Association for Computational Linguistics.

Claudiu Creanga and Liviu P. Dinu. 2024. Designing NLP systems that adapt to diverse worldviews. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 95–99, Torino, Italia. ELRA and ICCL.

Louis Estève, Christophe Servan, Thomas Lavergne, and Agata Savary. 2026. A diversity diet for a healthier model: A case study of french modernbert.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1229–1238, Minneapolis, Minnesota. Association for Computational Linguistics.

Claudio Greco, Alberto Testoni, Raffaella Bernardi, and Stella Frank. 2022. A small but informed and diverse model: The case of the multimodal GuessWhat!? guessing game. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Veronique Hoste, Chris Develder, and Thomas Demeester. 2023. Diverse content selection for educational question generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 123–133, Dubrovnik, Croatia. Association for Computational Linguistics.

Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and improving semantic diversity of dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 934–950, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stuart H. Hurlbert. 1971. The nonconcept of species diversity: A critique and alternative parameters. *Ecology*, 52(4):577–586. Number: 4.

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, Online. Association for Computational Linguistics.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-Ho Lee, Jeonguk Kim, Donghoon Shin, and Yeonsoo Lee. 2023. Persona expansion with commonsense knowledge for diverse and consistent response generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1139–1149, Dubrovnik, Croatia. Association for Computational Linguistics.

Sosuke Kobayashi, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2022. Diverse lottery tickets boost ensemble from a single pretrained model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 42–50, virtual+Dublin. Association for Computational Linguistics.

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? An investigation into multicultural proverbs and sayings. 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 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? A thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

Enrico Palumbo, Andrea Mezzalira, Cristina Marco, Alessandro Manzotti, and Daniele Amberti. 2020. Semantic diversity for natural language understanding evaluation in dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 44–49, Online. International Committee on Computational Linguistics.

Jun-Hyung Park, Hyuntae Park, Youjin Kang, Eojin Jeon, and SangKeun Lee. 2023. DIVE: Towards descriptive and diverse visual commonsense generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9677–9695, Singapore. Association for Computational Linguistics.

Alicia Parrish, Susan Hao, Sarah Laszlo, and Lora Aroyo. 2024. Is a picture of a bird a bird? A mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 1–18, Torino, Italia. ELRA and ICCL.

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is "typological diversity" in NLP? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700, Miami, Florida, USA. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Nguyen. 2022. MINION: a large-scale and diverse dataset for multilingual event detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2286–2299, Seattle, United States. Association for Computational Linguistics.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wrightbettner, and Martha Palmer. 2022. PropBank comes of age – larger, smarter, and more diverse. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington. Association for Computational Linguistics.

Sahotra Sarkar. 2010. Diversity: A philosophical perspective. *Diversity*, 2(1):127–141.

Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7795–7816, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yibin Shen, Qianying Liu, Zhuoyuan Mao, Zhen Wan, Fei Cheng, and Sadao Kurohashi. 2022. Seeking

diverse reasoning logic: Controlled equation expression generation for solving math word problems. 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 2: Short Papers)*, pages 254–260, Online only. Association for Computational Linguistics.

Feifan Song, Bowen Yu, Hao Lang, Haiyang Yu, Fei Huang, Houfeng Wang, and Yongbin Li. 2024. Scaling data diversity for fine-tuning language models in human alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14358–14369, Torino, Italia. ELRA and ICCL.

Andrew Stirling. 1994. Diversity and ignorance in electricity supply investment. *Energy Policy*, 22(3):195–216.

Andy Stirling. 2007. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719. Number: 15 Publisher: Royal Society.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Rocco Tripodi, Simone Conia, and Roberto Navigli. 2021. UniteD-SRL: A unified dataset for span- and dependency-based multilingual and cross-lingual semantic role labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2293–2305, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.

Vikas Yadav, Hyuk joon Kwon, Vijay Srinivasan, and Hongxia Jin. 2024. Explicit over implict: Explicit diversity conditions for effective question answer generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6876–6882, Torino, Italia. ELRA and ICCL.

Yuming Yang, Yang Nan, Junjie Ye, Shihan Dou, Xiao Wang, Shuo Li, Huijie Lv, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Measuring data diversity for instruction tuning: A systematic analysis and a reliable metric. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18530–18549, Vienna, Austria. Association for Computational Linguistics.

Yuting Yang, Pei Huang, Feifei Ma, Juan Cao, and Jintao Li. 2024. PAD: A robustness enhancement ensemble method via promoting attention diversity. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12574–12584, Torino, Italia. ELRA and ICCL.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Tianhui Zhang, Bei Peng, and Danushka Bollegala. 2025. Evaluating the evaluation of diversity in commonsense generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24258–24275, Vienna, Austria. Association for Computational Linguistics.