We Need to Measure Data Diversity in NLP — Better and Broader

Dong Nguyen
Utrecht University
The Netherlands
d.p.nguyen@uu.nl

Esther Ploeger
Aalborg University
Denmark
espl@cs.aau.dk

Abstract

Although diversity in NLP datasets has received growing attention, the question of how to *measure* it remains largely underexplored. This opinion paper examines the conceptual and methodological challenges of measuring data diversity and argues that interdisciplinary perspectives are essential for developing more fine-grained and valid measures.

1 Introduction

Language models seem to exhibit remarkable natural language understanding and generation capabilities. Yet numerous studies have pointed towards serious flaws: they encode societal biases (Gallegos et al., 2024; Hofmann et al., 2024) and are sensitive to spurious correlations (Du et al., 2023). Many of these weaknesses have been traced back to the training data (Feng et al., 2023; McCoy et al., 2024). Moreover, as the size of these training datasets continues to increase, improvements in task performance show diminishing returns (Tirumala et al., 2023). Data size is often prioritized at the expense of data quality (Paullada et al., 2021). However, attention for the quality and composition of NLP datasets is growing (Albalak et al., 2024; Elazar et al., 2024; van Noord et al., 2025; Yu et al., 2024).

One aspect is the diversity of datasets, i.e., **data diversity**, which plays a key role in learning and generalization in both machine and human learning (Raviv et al., 2022). Smaller, more diverse datasets have enabled competitive model performance despite their size (Zhou et al., 2023; Schiller et al., 2024). Data diversity is also important for the fairness of models (Santy et al., 2023); for example, models trained on primarily standard language perform worse on dialectal data (Blodgett et al., 2018). For evaluation, limited diversity in test data can be a cause for poor performance estimates (Freitag et al., 2020) and hinders the ability to distinguish between different systems (Sugawara et al., 2022).

Unfortunately, the concept of "data diversity" is challenging to define and measure, and researchers are rarely explicit about what they mean with data diversity (Zhao et al., 2024). But without clear definitions and measurements, it is difficult to make progress. Previous papers that advocate for measuring data diversity have left important aspects unaddressed. Mitchell et al. (2023) emphasize the importance of measuring data characteristics, though diversity in NLP data was not their main focus. Zhao et al. (2024) highlight the need to measure data diversity; however, their discussion mostly focuses on considerations during dataset construction, rather than diversity measurement itself. Shaib et al. (2025) call for standardizing text diversity measurements; however, their work is mostly empirical and focuses on lexical diversity. Estève et al. (2025) recently surveyed approaches to measure diversity in NLP; however, they focused less on data diversity, and they did not focus on the broader challenges that we outline. Further, the interdisciplinary focus of these papers is limited.

Yet, data diversity measurement has long been of central importance in a wide range of fields, including language learning and assessment (Jarvis, 2013), corpus linguistics (Kilgarriff, 2001; Egbert et al., 2022), sociology (Steele et al., 2022) and ecology (Simpson, 1949). Across these and other fields, scholars have also grappled with how to define and quantify diversity. Our position is that an interdisciplinary approach is essential for improving data diversity measurement in NLP.

Contribution In this opinion paper, we provide a critical perspective on the fundamental challenge of **measuring dataset diversity in NLP**. We focus on language data and associated information (e.g., about the authors or annotators). We operationally define data diversity as *the variability of the data across specific dimensions (e.g., semantic, syntactic, lexical, sociolinguistic, genre, language, etc.)*.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 8812–8821
November 4-9, 2025

Cross-lingual diversity has received increased attention in NLP (Joshi et al., 2020; Ploeger et al., 2024), leading to standardized language diversity metrics (Ponti et al., 2020; Samardzic et al., 2024). However, since within-language diversity remains underexplored, we focus on measuring diversity in monolingual settings. Still, our recommendations are not specific to any language. We first motivate *why* data diversity should be measured (§2). Then, we structure our discussion around three core challenges: diversity dimensions (§3.1), their measurement (§3.2), and evaluation of measurement (§3.3). Throughout our paper we highlight connections with other fields. Rather than offering definitive answers, we hope to spark discussion and to encourage more rigorous and interdisciplinary approaches to measuring data diversity. We conclude with concrete recommendations for NLP research that uses or develops diversity measures (§4).

Related concepts Related to data diversity is a dataset's *representativeness*, which is widely discussed in corpus linguistics but still lacks a consensus definition (Egbert et al., 2022). In other fields, like statistics and fairness in machine learning, the term has also been used in many ways (Chasalow and Levy, 2021). One common view in corpus linguistics is that a representative corpus is a miniature of the target population's language; some other views align more with how we define data diversity (Egbert et al., 2022; Stefanowitsch, 2020).

Another related concept is data *bias*, which refers to systematic distortions in the data (Olteanu et al., 2019). It often refers to societal bias, e.g., when the data reflects historical prejudices, or statistical bias, e.g., due to issues like sampling biases and measurement errors (Mitchell et al., 2021). Although data bias and data diversity are related, they are not the same. For example, reducing statistical bias can increase diversity (e.g., by including more social groups), and some efforts to increase data diversity (e.g., oversampling minority groups) could increase statistical bias.

2 Why Measure Data Diversity?

Although many papers make qualitative judgements about the diversity of datasets (e.g., Zhou et al. 2023; Röttger et al. 2022), the exact nature and extent of the differences in diversity can remain unclear. We take the view that data diversity should be **measured** — that is, quantitatively characterized — along different dimensions (Lai et al., 2020; Mitchell et al., 2023; Zhao et al., 2024). Measuring data diversity is relevant for both creating and reusing existing datasets, whether they are task-specific or more general-purpose.

Measuring properties of datasets is essential for advancing NLP as a *science* In 2001, Kilgarriff wrote: "*In any science, one expects to find a useful account of how its central constructs are taxonomised and measured, and how the subspecies compare. But to date, corpus linguistics has measured corpora only in the most rudimentary ways, ways which provide no leverage on the different kinds of corpora there are.*" (p. 97). This critique is also relevant to NLP: Quantitative analysis of dataset-level properties has been underexplored, with exceptions of recent work like Lohr and Hahn (2023) and Ramponi et al. (2024). Among these properties, diversity has been highlighted in the literature (Lai et al., 2020; Mitchell et al., 2023).

Measuring data diversity has the potential to improve every stage of the NLP pipeline First, it would support more deliberate and informed data collection and curation (Rogers, 2021). Data collection, especially when requiring human annotation, can be resource-intensive. Recent work (Quattoni and Carreras, 2021; Su et al., 2023; Shi et al., 2021; Alcoforado et al., 2024) shows that human annotation can be employed more effectively when selecting diverse datapoints. Furthermore, quantitative metrics on dataset diversity could improve the documentation of new and existing datasets (McMillan-Major et al., 2024; Gebru et al., 2021) and contribute to dataset analysis tools (Lohr and Hahn, 2023; Ramponi et al., 2024). This could support early intervention, during data collection, instead of post hoc. For example, researchers could proactively collect more data from certain language varieties to improve the fairness of models. Second, leveraging data diversity measurements for training and prompting language models has led to better performing models, exhibiting improved robustness (Bukharin et al., 2024) and generalization (Levy et al., 2023). Lastly, when it comes to evaluation, insufficiently substantiated diversity claims have caused disagreement on the generalizability of results. For example, an evaluation dataset that covers "a diverse range of text genres" (Wang et al., 2018) may, by other standards, deemed "hardly diverse" (Raji et al., 2021). Systematic data diversity measurement could help researchers to assess and interpret the generalizability of evaluation results.

Measuring data diversity can clarify its impact and help explain and predict model behavior Current knowledge of data diversity in NLP is fragmented and coarse-grained, partly because it is often not measured or measured inconsistently across studies. Rigorous data diversity measurement could reveal which dimensions of data diversity most strongly impact model behavior (e.g., task performance, generation quality, robustness, fairness), and how its effects interact with dataset size and other properties. For example, Guo et al. (2024) study the effect of models being continuously trained on generated text through the lens of data diversity. Besides explaining model behavior, diversity measurements could also serve as predictive signals. For example, they could help predict a model's task performance (Xia et al., 2020), enabling researchers to better select datasets that align with their use cases, or anticipate fairness issues.

3 Challenges in Measuring Data Diversity

This section reflects on three key challenges when measuring data diversity: (i) What diversity dimensions should we consider? (§3.1); (ii) How should we measure them? (§3.2); and (iii) How should we assess the quality of the measures? (§3.3).

3.1 Dimensions of Data Diversity

NLP studies have considered a range of diversity dimensions. The most common are lexical diversity (Gandhi et al., 2024; Guo et al., 2024; Lohr and Hahn, 2023; Shaib et al., 2025) and semantic diversity (Lai et al., 2020; Li et al., 2024; Yu et al., 2022). Studies have also considered various other broad dimensions, including syntactic (Guo et al., 2024), topical (Schiller et al., 2024), and genre diversity (Liu and Zeldes, 2023), as well as more task-specific dimensions, including question types (Kim et al., 2023; Ghazaryan et al., 2025), key points in essays (Padmakumar and He, 2024) and summarization styles (Grusky et al., 2018).

A broader view: Consider multiple diversity dimensions The diversity of NLP datasets can be conceptualized along many dimensions. However, most studies focus on only one or at most two, with a few exceptions (e.g., Guo et al. 2024). This narrow focus risks limiting our understanding of what constitutes a truly diverse dataset. Considering multiple dimensions simultaneously can also reveal how they correlate. For instance, semantic diversity and lexical diversity may correlate considerably, as expressing a broader range of meanings typically requires a more varied vocabulary. Some dimensions may also be better viewed as multidimensional. For example, using an umbrella term like "instruction diversity" while only measuring semantic diversity of instructions (Bukharin et al., 2024) risks that the results are interpreted more broadly than what the metric actually captures.

We cannot — and need not — measure all possible dimensions; likewise, datasets cannot (and need not) be highly diverse along all of them. Instead, we should selectively focus on the dimensions most relevant to the task at hand. With this broader, yet targeted approach to data diversity measurement, we can clarify which dimensions matter for which tasks, and make more informed decisions about dataset construction.

A broader view: Social dimensions of diversity Although NLP has increasingly considered various dimensions of data diversity, there remains a need to pay more attention to the social dimensions of diversity—especially those related to the identities of people who produce or annotate data. These aspects are central to the fairness of NLP systems, yet they get repeatedly ignored (Santy et al., 2023).

First, research should consider the sociolinguistic dimensions of diversity. Data collection and curation practices tend to favor *standard varieties* of languages (Blodgett et al., 2020; Gururangan et al., 2022; Nguyen, 2025). Language, however, provides a multitude of ways (e.g., words, grammatical constructions) to express similar ideas, and this variation is shaped by social identity and context (Meyerhoff, 2018). NLP models primarily trained on standard language varieties often have lower performance on texts written in other language varieties or produced by speakers from underrepresented groups (Faisal et al., 2024). Yet measuring diversity along dimensions such as stylistic (Wegmann et al., 2022) or dialectal (Grieve et al., 2025) variation remains underexplored.

Second, research should also consider the background of a dataset's authors or annotators. Who produced or labeled the dataset will substantially influence what is included, excluded, or emphasized, with implications for fairness and bias in downstream models. For instance, only modeling the majority vote of annotators can exclude minority perspectives, on both highly subjective tasks like hate speech detection and seemingly more objective tasks like NLI (Sap et al., 2022; Cabitza et al., 2023; Santy et al., 2023). Annotator diversity is also better seen as multidimensional; besides social categories like race or age, researchers have highlighted the importance of lived experience for annotator diversity (Kapania et al., 2023).

A broader view: Consider the situational factors of texts NLP could draw from corpus linguistics by measuring the distribution of situational factors (Staples et al., 2015), such as the mode of communication (spoken or written, synchronous or asynchronous). More broadly, diversity could be measured over registers, e.g., fiction, news articles, and academic writing (Goulart et al., 2020). However, in NLP datasets, such information is often not available, and would need to be additionally annotated or inferred (Egbert et al., 2015).

3.2 How Should We Measure Diversity?

A single diversity dimension can often be measured in many ways. Consider semantic diversity, which is typically measured by first mapping text to embeddings (Lai et al., 2020; Li et al., 2024; Yu et al., 2022), followed by computations, like the average pairwise similarity between sentences or the entropy over clusters (Guo et al., 2024; Han et al., 2022). How the measure is formulated can influence both the interpretation of results and how the measure can be used (e.g., sampling algorithms, data collection interventions). However, researchers do not always motivate why they measure a data diversity dimension in a certain way.

A broader view: Components of diversity dimensions Data diversity dimensions can be decomposed into different components. Ecology, with its long tradition of diversity measurement, often recognizes two core components of biodiversity: richness (i.e., the number of species) and evenness (i.e., how balanced a distribution of species is) (Magurran, 2004). Other components are sometimes also considered, like disparity (taking the similarity between species into account) (Daly et al., 2018). While ideas from ecology may not map perfectly onto NLP, especially for dimensions that are not easily captured by categorical data, it offers a useful lens for thinking about data diversity (Friedman and Dieng, 2023; Estève et al., 2025; Lion-Bouton et al., 2022). For instance, in language learning, ideas from ecology have been applied to measuring lexical diversity (Jarvis, 2013). Similarly, for topical diversity, we could consider the number of distinct topics (richness), how the data is distributed over topics (evenness), and the semantic similarity between topics (disparity).

NLP researchers, however, are rarely explicit about which components of a diversity dimension are considered. This ambiguity can lead to inconsistent or misleading interpretations. For example, does *increasing dataset size lead to greater diversity*? If the diversity measure captures richness then adding more data generally helps—or at least doesn't hurt. However, if a balanced distribution across categories (i.e., evenness) is targeted, simply scaling up the dataset may not improve diversity and could even reinforce imbalances.

Better data diversity measures: Be critical of representations Many measures rely on mapping data to some representation, e.g., discrete categories or continuous embeddings. However, the choice of representation shapes what is visible and measurable. For example, consider measuring semantic diversity. The quality of embeddings may vary across topics or dialects, and they may capture information (e.g., style, Wegmann et al. 2022) beyond semantics. Furthermore, measuring diversity along social dimensions (e.g., annotator background) often involves categorizing people. However, attributes like gender, race and ethnicity are socially constructed and often oversimplified in datasets (Mickel, 2024; Steele et al., 2022). Social science research can offer useful insights and tools for more nuanced representations and measurement (Mickel, 2024; Steele et al., 2022).

A broader view: Beware of cross-lingual differences Diversity measures often rely on specific tools or models (e.g., tokenizers, embedding models), and the availability and quality of such tools varies between languages (Blasi et al., 2022). Some measures cannot be applied to all languages. For example, measuring syntactic diversity using a dependency parser (Guo et al., 2024) is not possible if such a tool is lacking. Even when tools are available, the quality of diversity measurements can vary across languages. For example, Lion-Bouton et al. (2022) showed that errors in automatic lemmatization can artificially inflate diversity scores. Structural properties of languages can also influence the distribution of diversity scores. For instance, type-token ratio (TTR), often used to measure lexical diversity, was found to correlate with morphological complexity (Kettunen, 2014). Therefore, special care should be taken when applying and comparing diversity measures across languages.

3.3 Desiderata and Evaluation

What makes a good diversity measure? This question is rarely addressed, and diversity measures are rarely evaluated (with a few exceptions; Tevet and Berant 2021; Han et al. 2022; Yang et al. 2025; Zhang et al. 2025). Yet without doing so, it is unclear whether a measure captures what is intended or improves on others.

Better and broader evaluation: Leveraging measurement theory In the social sciences, measurement theory is widely used to evaluate the quality of measurements for abstract concepts, like personality or self-esteem (Trochim et al., 2016). Its benefits have already been highlighted for measuring biases (Van der Wal et al., 2024) and evaluating AI systems (Wallach et al., 2025; Fang et al., 2025). It also holds potential for data diversity measurement (Zhao et al., 2024). Two key indicators of measurement quality are (construct) validity (i.e., are we measuring what we intend to measure) and reliability (i.e., the consistency or repeatability of the measurements). Although measurement theory provides a useful toolbox, translating its methods and concepts to NLP also poses challenges (Van der Wal et al., 2024).

To illustrate how measurement theory can inform the evaluation of a diversity measure's validity, consider the following three complementary sources of construct validity evidence (Trochim et al., 2016). Content validity involves assessing whether the full scope of a concept is adequately covered. Suppose we aim to measure dialect diversity in a dataset: if the chosen measure covers only a small subset of major dialects, this could point to a content validity problem. Convergent validity is another example, which examines whether multiple measures for the same diversity dimension produce similar diversity scores (Shaib et al., 2025). Finally, criterion validity tests whether the measure behaves as expected according to an external benchmark—for example, by constructing datasets with known differences in diversity and verifying that the measure correctly ranks them. Together, these three validity indicators provide complementary evidence of a measure's validity.

Better data diversity measures: Defining desiderata The NLP community should also engage more deeply with the question of what makes a good diversity measure. Desiderata can be theoretical, for example that a measure should be continuous or clarity on the conditions under which it is maximal (Daly et al., 2018). Desiderata can also be practical like the interpretability or actionability of measures (Delobelle et al., 2024). For example, measures based on pairwise similarities between embeddings are increasingly popular. These measures are often transparent in their formulation, but their dependence on embedding models complicates interpretation; furthermore, while they indicate how similar instances are, they do not indicate what is missing from a dataset, making it more difficult to intervene. The efficiency of a diversity metric may also be an important practical desideratum. Shaib et al. (2025) point out that some metrics are prohibitively slow and therefore not suitable for large-scale application.

4 Conclusion

We have examined key challenges in measuring data diversity in NLP and conclude with six immediately actionable recommendations. These recommendations relate to *what* diversity aspects should be measured (§3.1), *how* they should be measured (§3.2), and the *assessment* of such measures (§3.3): (i) Clearly define what is meant by data diversity in the context of a study; (ii) Make deliberate, transparent choices about how diversity is measured. They should be guided by the specific research goals and application; (iii) Develop measures that better capture the social dimensions of diversity; (iv) Critically examine what makes a good diversity measure. NLP needs clearer guidance on the desirable theoretical and practical properties of data diversity measures; (v) Evaluate the quality of data diversity measurement; and finally (vi) Look beyond NLP. Many fields have developed approaches to measure diversity; we should learn from them.

Limitations

Diversity is a rich and complex concept. Although we take the view that researchers should measure a dataset's diversity, we acknowledge that any quantification necessarily results in a loss of nuance and information. Moreover, not all dimensions of diversity are necessarily measurable. Engaging with qualitative perspectives on diversity is thus important; it could bring a critical and context-sensitive lens to diversity in NLP datasets. We also believe that combining manual and automatic analyses of datasets is an important path forward (Birhane and Prabhu, 2021). Relatedly, we acknowledge that quantifying data diversity may lead to diversity metrics becoming objectives on their own. We caution against the blind optimization of diversity metrics, as the quality of a dataset depends on various factors. Purely optimizing for a diversity metric might overlook other important factors of dataset quality, such as the relevance of texts.

Acknowledgments

We would like to thank the members of the NLP and Society Lab at Utrecht University and the anonymous reviewers for their feedback. This work is supported by the ERC Starting Grant DataDivers 101162980 and the Carlsberg Foundation under the *Semper Ardens: Accelerate* programme (CF21-0454). We have used ChatGPT to improve the grammar and phrasing of some sentences in this paper.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Transactions on Machine Learning Research*. Survey Certification.

Alexandre Alcoforado, Lucas Hideki Takeuchi Okamura, Israel Campos Fama, Bárbara Fernandes Dias Bueno, Arnold Moya Lavado, Thomas Palmeira Ferraz, Bruno Veloso, and Anna Helena Reali Costa. 2024. From random to informed data selection: A diversity-based approach to optimize human annotation and few-shot learning. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 492–502, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE.

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.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. 2024. Data diversity matters for robust instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3411–3425, Miami, Florida, USA. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Kyla Chasalow and Karen Levy. 2021. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 77–89, New York, NY, USA. Association for Computing Machinery.

Aisling J. Daly, Jan M. Baetens, and Bernard De Baets. 2018. Ecological diversity: Measuring the unmeasurable. *Mathematics*, 6(7).

Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. Metrics for what, metrics for whom: Assessing actionability of bias evaluation metrics in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21669–21691, Miami, Florida, USA. Association for Computational Linguistics.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Commun. ACM*, 67(1):110–120.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Jesse Egbert, Douglas Biber, and Bethany Gray. 2022. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge University Press.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations*.

Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik, Agata Savary, and Olha Kanishcheva. 2025. A survey of diversity quantification in natural language processing: The why, what, where and how. *Preprint*, arXiv:2507.20858.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Qixiang Fang, Daniel Oberski, and Dong Nguyen. 2025. PATCH! Psychometrics-AssisTed BenCHmarking of large language models against human populations: A case study of proficiency in 8th grade mathematics. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 808–823, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better synthetic data by retrieving and transforming existing datasets. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6453–6466, Bangkok, Thailand. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Gayane Ghazaryan, Erik Arakelyan, Isabelle Augenstein, and Pasquale Minervini. 2025. SynDARin: Synthesising datasets for automated reasoning in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6459–6466, Abu Dhabi, UAE. Association for Computational Linguistics.

Larissa Goulart, Bethany Gray, Shelley Staples, Amanda Black, Aisha Shelton, Douglas Biber, Jesse Egbert, and Stacey Wizner. 2020. Linguistic perspectives on register. *Annual Review of Linguistics*, 6(Volume 6, 2020):435–455.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7:1472411.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates. 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.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154.

Scott Jarvis. 2013. Capturing the diversity in lexical diversity. *Language Learning*, 63(s1):87–106.

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.

Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Jiwoo Kim, Youngbin Kim, Ilwoong Baek, JinYeong Bak, and Jongwuk Lee. 2023. It ain't over: A multi-aspect diverse math word problem dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14984–15011, Singapore. Association for Computational Linguistics.

Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. 2020. Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1739–1746, Marseille, France. European Language Resources Association.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Ruihang Li, Yixuan Wei, Miaosen Zhang, Nenghai Yu, Han Hu, and Houwen Peng. 2024. ScalingFilter: Assessing data quality through inverse utilization of scaling laws. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3209–3222, Miami, Florida, USA. 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.

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.

Christina Lohr and Udo Hahn. 2023. DOPA METER – a tool suite for metrical document profiling and aggregation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 218–228, Singapore. Association for Computational Linguistics.

Anne E. Magurran. 2004. *Measuring biological diversity*. Blackwell Science Ltd.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2024. Data statements: From technical concept to community practice. *ACM J. Responsib. Comput.*, 1(1).

Miriam Meyerhoff. 2018. *Introducing sociolinguistics*. Routledge.

Jennifer Mickel. 2024. Racial/ethnic categories in AI and algorithmic fairness: Why they matter and what they represent. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2484–2494, New York, NY, USA. Association for Computing Machinery.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. 2023. Measuring data. *Preprint*, arXiv:2212.05129.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(Volume 8, 2021):141–163.

Dong Nguyen. 2025. Collaborative growth: When large language models meet sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, Volume 2 - 2019.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations*.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

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.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Ariadna Quattoni and Xavier Carreras. 2021. Minimizing annotation effort via max-volume spectral sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2890–2899, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Alan Ramponi, Camilla Casula, and Stefano Menini. 2024. Variationist: Exploring multifaceted variation and bias in written language data. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 346–354, Bangkok, Thailand. Association for Computational Linguistics.

Limor Raviv, Gary Lupyan, and Shawn C Green. 2022. How variability shapes learning and generalization. *Trends in cognitive sciences*.

Anna Rogers. 2021. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tanja Samardzic, Ximena Gutierrez, Christian Bentz, Steven Moran, and Olga Pelloni. 2024. A measure for transparent comparison of linguistic diversity in multilingual NLP data sets. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3367–3382, Mexico City, Mexico. Association for Computational Linguistics.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, Andreas Waldis, and Iryna Gurevych. 2024. Diversity over size: On the effect of sample and topic sizes for topic-dependent argument mining datasets. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10870–10887, Miami, Florida, USA. Association for Computational Linguistics.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2025. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *Preprint*, arXiv:2403.00553.

Tianze Shi, Adrian Benton, Igor Malioutov, and Ozan İrsoy. 2021. Diversity-aware batch active learning for dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2616–2626, Online. Association for Computational Linguistics.

Edward H Simpson. 1949. Measurement of diversity. *Nature*, 163(4148):688–688.

Shelley Staples, Jesse Egbert, Douglas Biber, and Susan Conrad. 2015. *Register Variation A Corpus Approach*, chapter 24. John Wiley & Sons, Ltd.

Liza G. Steele, Amie Bostic, Scott M. Lynch, and Lamis Abdelaaty. 2022. Measuring ethnic diversity. *Annual Review of Sociology*, 48(1):43–63.

Anatol Stefanowitsch. 2020. *Corpus linguistics*. Number 7 in Textbooks in Language Sciences. Language Science Press, Berlin.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*.

Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel Bowman. 2022. What makes reading comprehension questions difficult? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6951–6971, Dublin, Ireland. Association for Computational Linguistics.

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.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2023. D4: Improving LLM pretraining via document de-duplication and diversification. In *Advances in Neural Information Processing Systems*, volume 36, pages 53983–53995. Curran Associates, Inc.

William MK Trochim, James P Donnelly, and Kanika Arora. 2016. Research methods: The essential knowledge base.

Oskar Van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. 2024. Undesirable biases in NLP: Addressing challenges of measurement. *Journal of Artificial Intelligence Research*, 79:1–40.

Rik van Noord, Miquel Esplà-Gomis, Malina Chichirau, Gema Ramírez-Sánchez, and Antonio Toral. 2025. Quality beyond a glance: Revealing large quality differences between web-crawled parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1824–1838, Abu Dhabi, UAE. Association for Computational Linguistics.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. In *ICML 2025*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. 2020. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8625–8646, Online. Association for Computational Linguistics.

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.

Xiao Yu, Zexian Zhang, Feifei Niu, Xing Hu, Xin Xia, and John Grundy. 2024. What makes a high-quality training dataset for large language models: A practitioners' perspective. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, ASE '24, page 656–668, New York, NY, USA. Association for Computing Machinery.

Yu Yu, Shahram Khadivi, and Jia Xu. 2022. Can data diversity enhance learning generalization? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4933–4945, Gyeongju, Republic of Korea. International Committee on 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.

Dora Zhao, Jerone T. A. Andrews, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Position: measure dataset diversity, don't just claim it. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.