Reconsidering the Inter- vs. Intra-Language Diversity Distinction

Esther Ploeger
Utrecht University, The Netherlands
e.ploeger@uu.nl

Relevant UniDive working groups: WG4

1 Introduction

Within the UniDive COST Action, diversity is commonly approached from two distinct angles: that of diversity *within* languages, and diversity *between* languages. These two perspectives are mentioned explicitly in UniDive's goals: to "quantify inter- and intra-linguistic diversity" (Savary et al., 2024). Recent relevant work by UniDive members usually aligns with either of these two perspectives. For example, `DiversUtils`[1] and `DELTA` (Estève and Dobrovoljc, 2026) address diversity within languages (i.e., on the level of text), while `LangDive` (Samardzic et al., 2024), `DistaLs` (Van Der Goot et al., 2025) and `TypDiv` (Ploeger et al., 2026) explicitly target inter-language diversity (i.e., on the level of languages as a whole). **However, the interaction between these two perspectives is commonly overlooked.**

The concept of neatly distinguishable, monolithic 'languages' has long been criticized in sociolinguistics (Jørgensen et al., 2015). In reality, 'languages' are often mixed, so the assumption of monolingual text is not necessarily representative of real-life language use; drawing boundaries between what counts as text in one language versus another is not always trivial. To illustrate: Example (1) and (2) show that questions written by native Danish speakers contain words commonly associated with English. In §2, we discuss why this is relevant for natural language processing (NLP) research generally. In §3, we discuss how a language's structural properties can influence diversity measurement in particular.

Ultimately, the goals of this extended abstract are to report on planned experimental work, as well as to open up a discussion among UniDive members: *Should we always separate these two perspectives on language diversity?*

[1] https://github.com/estevelouis/WG4

(1) *Er det okay at have crop top på til*
is it okay to have crop top on to
familie fødselsdag?
family birthday

Translation: 'Is it okay to wear a crop top to a family birthday?'

(2) *Hvad er en typisk brunch cocktail?*
what is a typical brunch cocktail

Translation: 'What is a typical brunch cocktail?'

2 Language Boundaries in NLP

Within sociolinguistics, it has been argued that 'languages' (such as German and Danish) are sociocultural constructs, which are not necessarily representative of how people actually use language. Jørgensen et al. (2015) show that speakers do not confine themselves to neatly separated codes such as 'German' or 'Danish'. Instead, they draw on whatever resources are available to them, regardless of how these resources are traditionally classified. Since it has been estimated that 'most of the human language users in the world speak more than one language', monolingualism is the exception rather than the norm (Auer and Wei, 2008).

This is in stark contrast with common practice in multilingual NLP research, which often relies on strict language boundaries: 'multilingual' datasets such as `ParaCrawl` (Esplà-Gomis et al., 2019; Bañón et al., 2020) and `mC4` (Xue et al., 2021) provide data separately per language.[2] Yet, the sociolinguistic realities of language use are increasingly important in NLP as well. Firstly, as data-hungry methods for language modelling remain popular, increasingly large multilingual datasets are mined from the web. Rather than being extensively curated and standardized, these corpora may increasingly include text from social media and blogs, more closely reflecting real-life linguis-

[2] In practice, such corpora may contain language misclassifications and other issues (Kreutzer et al., 2022).

tic practice. However, relying on strict language boundaries to process such data risks imposing artificial divisions onto linguistic behavior, thereby ignoring the cross-resource mixing that characterizes actual language use. Secondly, as more lay people interact with language technology via natural language interfaces, i.e., with chatbots based on large language models (LLMs), dealing with such language use becomes crucial. For example, Müller-Eberstein et al. (2025) asked Danish speakers to interact with an LLM and prompt it with questions about Danish culture. Example (1) and (2), sourced from the DAKULTUR dataset (Müller-Eberstein et al., 2025), shows that such natural Danish includes words that are commonly associated with English ('crop top', 'brunch cocktail').

3 How Do Language Properties Affect Diversity Measurement?

In practice, in NLP research, measuring diversity *within* a language is often synonymous with measuring diversity *within English*. Yet, structural characteristics of a language, such as morphological complexity, can influence diversity measurement. For example, Kettunen (2014) found that type-token ratio (TTR), one of the most popular measures of lexical diversity, is correlated with morphological complexity. More recently, Samardzic et al. (2024) used the mean word length of a corpus in a language as 'an approximation of aggregated morphological features'.

Consider again Example (1). While Danish exhibits productive nominal compounding, the word 'crop top' is split[3], as well as the Danish compound *familiefødselsdag* into *familie fødselsdag*. This pattern aligns more closely with English orthographic conventions than with Danish compounding practices. While such variations are not problematic from a sociolinguistic standpoint, they can have a considerable impact on a lexical diversity metric such as TTR.

As another example: Shaib et al. (2025) introduce a toolkit for measuring text diversity in NLP, including metrics such as Compression Ratio (the ratio between the size of a file compressed with gZip, versus its original size) and self-BLEU (Zhu et al., 2018). A limitation of this work, however, is that all its evaluations were conducted for English. As a small case study, we take 100 lines of parallel text for two languages, from WMT24 (Kocmi et al., 2024), and calculate the two aforementioned metrics. Table 1 shows that the compression ratio for the Hindi text is higher than that for the Spanish text (indicating lower diversity), while using the self-BLEU metric shows the opposite. Different metrics for lexical diversity may thus be sensitive to different language properties, but this has, to the best of our knowledge, not been evaluated systematically thus far.

Language	**C.R. (↓)**	**self-BLEU (↓)**
Spanish	2.5440	0.1220
Hindi	4.7970	0.0110

Table 1: Lexical diversity metrics (Compression Ratio; C.R. and self-BLEU) for 100 parallel lines in Spanish and Hindi, calculated using Shaib et al. (2025)'s `diversity` toolkit.

4 A Starting Point for Discussion

Working Group 4, and the UniDive COST Action in general, have made considerable progress in measuring intra- and inter-linguistic diversity in NLP. One of the outcomes is a range of tools (including `DiversUtils`[4], `DELTA` (Estève and Dobrovoljc, 2026), `LangDive` (Samardzic et al., 2024), `DistaLs` (Van Der Goot et al., 2025) and `TypDiv` (Ploeger et al., 2026)), which can be used to measure diversity within and between languages. With the end of the action nearing, it could be of interest to discuss **how our toolkits apply to real-world language use**, in which the boundaries between intra- and inter-language variation are perhaps less distinguishable. How can we compare within-language diversity metrics fairly *across* languages? How should our choice of diversity metrics be influenced by properties of the language under study? When is it necessary to assume strict language boundaries, and when is it possible to include multilingual realities? Do we necessarily need to distinguish between diversity within and between languages, or could we unify these perspectives?

[3]The Danish online dictionary includes the word *croptop*, https://ordnet.dk/ddo/ordbog?query=croptop

[4]https://github.com/estevelouis/WG4

References

Peter Auer and Li Wei. 2008. Introduction: Multilingualism as a problem? monolingualism as a problem? *Handbook of multilingualism and multilingual communication*, pages 1–12.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, et al. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4555–4567.

Miquel Esplà-Gomis, Mikel L Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. Paracrawl: Web-scale parallel corpora for the languages of the eu. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119.

Louis Estève and Kaja Dobrovoljc. 2026. Delta: A toolkit for measuring linguistic diversity in dependency-parsed corpora. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 75–85.

Jens Normann Jørgensen, Martha Sif Karrebæk, Lian Malai Madsen, and Janus Spindler Møller. 2015. Polylanguaging in superdiversity. In *Language and superdiversity*, pages 147–164. Routledge.

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

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Max Müller-Eberstein, Mike Zhang, Elisa Bassignana, Peter Brunsgaard Trolle, and Rob Van Der Goot. 2025. DaKultur: Evaluating the cultural awareness of language models for Danish with native speakers. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 50–58, Albuquerque, New Mexico. Association for Computational Linguistics.

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2026. A principled framework for evaluating on typologically diverse languages. *Computational Linguistics*, 52(1):237–269.

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.

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine De Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, et al. 2024. Unidive: A cost action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 372–382.

Chantal Shaib, Venkata S Govindarajan, Joe Barrow, Jiuding Sun, Alexa Siu, Byron C Wallace, and Ani Nenkova. 2025. Standardizing the measurement of text diversity: A tool and comparative analysis. In *Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 36–46.

Rob Van Der Goot, Esther Ploeger, Verena Blaschke, and Tanja Samardzic. 2025. Distals: a comprehensive collection of language distance measures. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 307–318.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.