UniDive Tools and Methods for Measuring Linguistic Diversity in NLP Datasets

Tanja Samardžić
IDSIA USI-SUPSI
Lugano

Rob van der Goot
IT University
Copenhagen

Louis Esteve
Université
Paris-Saclay

Olha Kanishcheva
Friedrich-Schiller
Universität Jena

Wessel Poelman
KU Leuven

Agata Savary
Université
Paris-Saclay

Sercan Karakas
University
of Chicago

Kaja Dobrovoljc
Jozef Stefan
Institute, Ljubljana

Voula Giouli
Aristotle University
of Thessaloniki

Luka Terčon
University
of Ljubljana

Marie-Catherine de Marneffe
UC Louvain

Relevant UniDive working groups: WG4

1 Introduction

Multilingual tasks or benchmarks are being increasingly created in the NLP community for measuring the performance of contemporary language models across languages. An obvious question posed when creating a multilingual benchmark is how do we choose the languages and the text samples to be included in it. While the availability of annotated data is the most important criterion, researchers are increasingly aware of the need for linguistic diversity, which is sometimes set as a sampling objective before even producing a new multilingual dataset. For instance, the authors of the XCOPA benchmark (Ponti et al., 2020) proposed a typological index (TI) as an objective quantification of the linguistic diversity of their dataset. Otherwise, the most common diversity quantification method is the number of language families. For instance, the TyDiQA benchmark includes 11 languages coming from 10 language families illustrating what is considered to be a highly diverse dataset. On the other hand, Universal Dependencies (de Marneffe et al., 2021; Nivre et al., 2020) have been criticised in their beginning for being dominated by one language family (Indo-European).

While the number of language families might serve as an approximative estimation of linguistic diversity, there is a need for more precise measures that would take into account the fact that some language families might contain diverse languages. Hence, one of the prominent objectives of the UniDive WG4 has been to develop tools and methods for measuring linguistic diversity at different levels. As a result, several Python libraries have been created in the activities related to WG4 addressing various aspects of measuring linguistic diversity. In this proposed presentation, we report on the work in progress aiming at demonstrating the use of these libraries for measuring linguistic diversity in NLP datasets at scale.

2 Levels of Linguistic Diversity

The tools and methods included in our overview address linguistic diversity at two levels, sometimes confounded in the literature.

The first level concerns **meta-linguistic categories**, that is the properties of a language as a whole. For example, some languages allow sentences without an explicit subject, while others do not. This is a binary meta-linguistic feature known as "pro-drop". Features of this kind are typically listed in typological databases such as WALS,[1] Grambank[2] or PHOIBLE.[3] In addition to the information about the structure of languages, meta-linguistic categories can encode the information about the location of a language in a genealogical tree, its geographical location, script, the number of speakers or the endangerment status. Popular sources of such features, also called *language metadata* are Glottolog[4] and, more recently, LinguaMeta.[5] These features can be queried from the sources to calculate distances between the languages or the distribution of feature values in any given set of languages.

The second level concerns **in-text categories**,

[1] https://wals.info/feature
[2] https://grambank.clld.org/parameters
[3] https://phoible.org
[4] https://glottolog.org
[5] https://github.com/google-research/url-nlp/tree/main/linguameta

that is frequency distributions of a given lexical or grammatical phenomenon within a given text sample. For instance, in a language that allows sentences without subjects, one text genre can be characterised by an increased use of subjects compared to the language base rate. At this level, we can observe and measure how varied lexical and grammatical choices are in any given text sample of any language. This is a more fine-grained notion of linguistic diversity that can be applied to characterise the diversity of a single language, but such measures can also be used to compare different languages provided that comparability is directly ensured, e.g. by controlling the text data size.

While the tools reviewed here tend to focus on one or the other category of linguistic diversity, they might use both kinds.

3 Published Python Libraries

The collection of UniDive tools and methods for measuring linguistic diversity at both levels includes the following items:

DiversUtils measures the text-level diversity of linguistic resources as well as the diversity of NLP systems' predictions (https://github.com/estevelouis/WG4).

DistaLs (Goot et al., 2025) is focused on measuring distances between languages using different features: metalinguistic information, typological data (phonology, morphology, syntax, lexicon) and features derived from textual data (https://bitbucket.org/robvanderg/distals/src/master/).

LangDive (Samardzic et al., 2024) assesses the level of linguistic diversity in a given set of languages with respect to the reference diversity of the WALS 100L sample using structural and text-based features (https://github.com/ICEF-NLP/jmm_diversity/tree/langdive-lib?tab=readme-ov-file).

TypDiv (Ploeger et al., 2025) is a tool for sampling diverse sets of languages using structural linguistic features (https://github.com/esther2000/typdiv-sampling).

DELTA (Estève and Dobrovoljc, 2026) DELTA measures linguistic diversity across lexical, morphological, syntactic, and other CoNLL-U-encoded phenomena by extracting dependency-based patterns of any complexity and applying a broad range of diversity metrics (https://gitlab.lisn.upsaclay.fr/esteve/delta).

Qwanqwa (Poelman et al., 2026) is a language metadata toolkit to make it easier to work with a large variety of metadata from a single interface. This library does not calculate measures, but is an essential tool for creating the input language lists for the distance/diversity libraries (https://github.com/WPoelman/qwanqwa).

These tools deal with different aspects of linguistic diversity and related topics. For instance, to identify the languages over which we calculate the measures with these libraries, we need to know their exact ISO-codes, which can be provided by the Qwanqwa library. These codes can be given as input to DistaLs or LangDive. LangDive and DiversUtils also take text data as input, while DELTA works with syntactically parsed texts.

4 Case Studies and Analyses

To better understand the complementarity of these tools, as well as their potential overlap, we carry out a number of case studies. In each case study, we take several sets of languages included in a multilingual NLP dataset and we measure the diversity of this set of languages using the appropriate tool.

These case studies include resources associated with UniDive (UD, PARSEME, AdMIRE, ELEXIS-WSD, Universal NER) and various multilingual NLP benchmarks (FLORES+, MMLU-ProX, Okapi, BenchMAX, mBERT, XTREME, XGLUE, XNLI, XCOPA, TyDiQA, XQuAD). In addition to these resources, the plan is to screen some other benchmarks more recently published in the ACL Anthology.

Taking various diversity measures over these datasets will allow us to assign objective linguistic diversity rankings taking into account multiple criteria. This will also allow us to validate the measures themselves by identifying possible overlaps and complementarity between them. For example, we should expect the Typological Index, the number of language families and the mean language distance to be correlated. If not, we can look for causes and revise the measures to be more reliable. We will also compare text-level and language-level features to see whether resources that include diverse languages also include diverse text samples. A benchmark may cover typologically distinct languages yet exhibit little structural diversity. If so,

this might point to limitations in the design of the datasets.

Based on these analyses, we aim to provide recommendations on how to compose diverse datasets. For instance, including a language or family with specific feature values might increase the diversity score of a new dataset. Another strategy could be to to use TypDiv for initial sampling which would be refined to accommodate other criteria.

Finally, various diversity measures are necessary to investigate potential gains in cross-linguistic generalisation of NLP systems thanks to the exposure to diverse linguistic phenomena.

References

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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, Rabat, Marocco. Association for Computational Linguistics.

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, Suzhou, China. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2025. A Principled Framework for Evaluating on Typologically Diverse Languages. *Computational Linguistics*, pages 1–36.

Wessel Poelman, Yiyi Chen, and Miryam de Lhoneux. 2026. QQ: A toolkit for language identifiers and metadata.

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.

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.