DELTA: A Toolkit for Measuring Linguistic Diversity in Dependency-Parsed Corpora

Louis Estève
Université Paris-Saclay, CNRS, LISN
91400, Orsay, France
louis.esteve@lisn.fr

Kaja Dobrovoljc
University of Ljubljana, Slovenia
Jozef Stefan Institute, Slovenia
kaja.dobrovoljc@ff.uni-lj.si

Abstract

Despite growing interest in measuring linguistic diversity on the one hand and the increasing availability of cross-linguistically comparable parsed corpora on the other, tools for systematically measuring the diversity of specific linguistic phenomena on such data remain limited. To address this gap, we present DELTA, an open-source framework that integrates dependency tree querying with diversity computation, enabling systematic measurement across multiple linguistic levels (e.g., lexis, morphology, syntax) and multiple diversity dimensions (variety, balance, disparity). The pipeline processes CoNLL-U formatted corpora through configurable workflows, treating the format as a general-purpose tabular structure independent of specific annotation conventions. We validate DELTA on Parallel Universal Dependencies multilingual dataset, demonstrating its capacity for corpus profiling and cross-corpus diversity comparison.

1 Introduction

Natural Language Processing (NLP) has seen increasing interest in the concept of diversity in recent years. The year-wise share of papers in ACL Anthology containing "diversity" or "diverse" in their title or abstract has risen from less than 1% in the 2000s to over 10% in 2024 (Estève et al., 2025). Archetypical examples include research on generative models with concerns for diverse output, or dataset creation efforts prioritizing diverse content.

This growing interest reflects the variety of motivations, metrics, and target phenomena associated with diversity. Estève et al. (2025) identify two axes describing the motivations behind diversity: *goal versus means*, and *practical versus ethical*. For instance, ethical motivations include improved deontology (Song et al., 2024) and inclusiveness (Joshi et al., 2020), while practical motivations include meeting user expectations (Kumar et al., 2019) and improving model performance (Liu and Zeldes, 2023). The survey identifies 150 different equations for measuring diversity, and the target phenomena span multiple linguistic levels: lexis (Kosmajac and Keselj, 2019), morphology (Samir and Silfverberg, 2023), syntax (Guo et al., 2024), and semantics (Jolly et al., 2021).

In recent years, there has been growing interest in measuring diversity specifically on dependency-parsed corpora. The widespread adoption of Universal Dependencies (de Marneffe et al., 2021) – now covering hundreds of languages with consistent annotation (Zeman et al., 2025) – has made parsed data readily available for cross-linguistic analysis. Parsed corpora enable investigation of diversity across multiple linguistic levels—from lexis and morphology to syntactic and semantic patterns – opening possibilities for studying linguistic universals (Gerdes et al., 2021), typological differences (Levshina, 2019), and the impact of diversity on NLP systems (Savary et al., 2024).

However, while tools exist for querying parsed corpora and for computing diversity metrics, these capabilities have typically remained separate. Researchers must manually combine pattern extraction with statistical analysis, export intermediate results, and write custom code to bridge the two stages. This fragmented workflow hinders reproducibility, limits cross-study comparability, and presents barriers for researchers without programming expertise. What remains absent is an integrated framework enabling multi-level and multi-dimensional diversity measurement directly from dependency-parsed corpora, supporting comparative analysis across languages and datasets.

To bridge this gap, we present **DELTA**, a unified pipeline for measuring linguistic diversity in dependency-parsed corpora. DELTA integrates tree extraction and diversity computation, enabling multi-level and multi-dimensional diversity analysis across languages. In doing so, our main contributions are:

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 75–85

1. **An integrated pipeline** from annotated input to diversity scores, combining dependency graph querying with diversity metric computation on CoNLL-U formatted data, eliminating the need for manual data transformation between tools.

2. **A flexible framework** that enables multi-level and multi-dimensional measurement across linguistic levels (lexis, morphology, syntax) and diversity dimensions (variety, balance, disparity), with support for customizable queries and metrics, and adaptable to alternative annotation schemes beyond Universal Dependencies.

3. **Scalable infrastructure** with SLURM-based parallelization for large-scale analyses, along with pre-defined configurations for common diversity measurement tasks and plotting support.

We present the pipeline in the remainder of this paper and demonstrate it by measuring lexical, morphological and syntactic diversity across Parallel Universal Dependencies (PUD) treebanks (Zeman et al., 2017).

2 Related work

Numerous tools support structured exploration of dependency-parsed corpora, including online services such as Grew-match (Guillaume, 2021), PML-TQ (Štěpánek and Pajas, 2010), and INESS (Rosén et al., 2012), lightweight libraries such as `pyconll`[1] and `conllu`[2] for programmatic access, and STARK (Krsnik and Dobrovoljc, 2025) for quantitative subtree extraction. Separately, various libraries implement diversity indices from ecology and information theory, including the `diverse` R package (Guevara et al., 2016), `scikit-bio` (Rideout et al., 2025), and DiversUtils[3]. However, these querying and computation capabilities remain disconnected, requiring manual export and custom code to bridge the two stages.

Tools such as DistaLs (Goot et al., 2025), LangDive (Samardzic et al., 2024), and TypDiv (Ploeger et al., 2024) quantify diversity at the level of language samples—using typological databases and, in some cases, text-derived features – but focus on comparing languages or multilingual datasets rather than profiling specific phenomena within corpora. For profiling parsed corpora specifically, tools such as ComparaTree (Terčon and Dobrovoljc, 2025), Profiling-UD (Brunato et al., 2020), and Typometrics (Gerdes et al., 2021) support cross-linguistic comparison, but are limited to fixed feature sets and predefined comparison scenarios.

DELTA bridges these approaches within a single unified framework, enabling flexible, multi-level, and multi-dimensional diversity measurement of specific linguistic phenomena directly over dependency-parsed corpora.

3 System Architecture

DELTA takes any number of annotated corpora in CoNLL-U format as input and produces diversity measurements for each corpus as output. Built on two preexisting open-source tools – STARK[4] for pattern extraction and DiversUtils for diversity computation – the system provides a unified framework for flexible diversity analysis. Figure 1 illustrates the pipeline: users provide two configuration files specifying (1) which linguistic patterns to extract and (2) which diversity metrics to compute, and the system then extracts matching instances from each corpus and calculates their diversity.

In essence, DELTA's diversity measurement relies on the distinction between *categories* (types of linguistic patterns) and *elements* (individual occurrences of those patterns). For instance, if "noun phrase" is defined as a category of interest, then each individual noun phrase in the corpus constitutes an element of that category. This element/category dichotomy comes from ecology, where it is often termed the type/item dichotomy (Ramaciotti Morales et al., 2021; Solé et al., 2010): categories (types) often correspond to species, in which case elements (items) correspond to individual organisms.

The following subsections describe pattern extraction (§3.1), diversity computation (§3.2), output (§3.3), availability (§3.4) and scalability (§3.5).

3.1 Category extraction (STARK)

DELTA represents categories as dependency subtrees extracted from CoNLL-U formatted corpora. This tree-based representation means DELTA can compute diversity of any linguistic phenomenon expressible in tree-like form – from single-node

[1] `https://pyconll.github.io/`
[2] `https://pypi.org/project/conllu/`
[3] `https://github.com/estevelouis/WG4`
[4] `https://github.com/clarinsi/STARK`

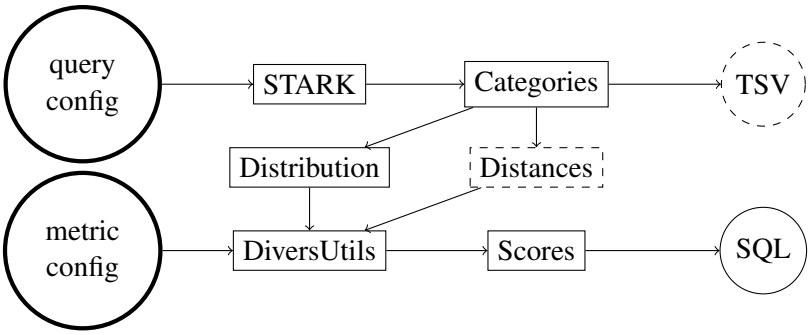


Figure 1: High-level representation of the main mechanisms in DELTA. Objects with dashed contour may not be generated depending on configuration.

subtrees (e.g., word forms, lemmas, POS tags) capturing lexical or morphological diversity, to multi-node subtrees (e.g., phrases, clauses, sentences) capturing (morpho)syntactic diversity.

STARK for pattern extraction. Category extraction is performed using STARK, an open-source toolkit for extracting dependency subtrees from parsed corpora. STARK extracts subtrees matching user-defined configurations, which can range from all attested subtrees to specific structural types. Each extracted subtree is counted and output with frequency information, providing the information for diversity computation.

Flexible pattern specification. The subtrees to be extracted can be defined flexibly along multiple dimensions. Users control tree representation by specifying which information appear on nodes (word forms, lemmas, part-of-speech tags, or combinations thereof), whether dependency labels are included, and whether linear word order is considered. Users also control tree filtering by specifying size constraints (e.g., only trees with a specific number of nodes), head constraints (e.g., only noun-headed structures), allowed or ignored dependency relations, or exact structural patterns via custom queries (e.g., only adjective-noun structures). These flexible and combinable parameters give users precise control over extraction granularity – from fully lexicalized constructions to abstract structural templates – making category extraction adaptable to diverse research goals. An overview of STARK's functionality is given by Krsnik and Dobrovoljc (2025), with detailed documentation also available online.[5]

Predefined configurations. To facilitate common use cases, DELTA also provides some predefined extraction configurations targeting specific linguistic phenomena. For single-node extraction, configurations include `forms.ini` (all word forms), `lemmas.ini` (all lemmas), `parts-of-speech.ini` (all POS tags), and `morphology.ini` (all POS and feature combinations). For multi-node structures, two general configurations extract all subtrees as proxies for all syntactic structures attested in a corpus (see Dobrovoljc (2025) for details): `syntactic-structures.ini` for extracting all dependency-labeled subtrees and `morphosyntactic-structures.ini` for extracting all dependency-labeled subtrees with POS tags as nodes.

As examples of using DELTA to measure diversity of very specific phenomena, we also include two specialized configurations targeting commonly analyzed syntactic structures: `svo.ini` for extracting delexicalized subject-verb-object patterns featuring *nsubj* and *obj* relations (Levshina, 2019) and `mwe.ini` for extracting lexicalized multi-word expressions featuring *fixed*, *flat*, and *compound* constructions (Savary et al., 2023). These configurations facilitate standard diversity measurements without requiring detailed parameter specification, but users can also define custom configurations, if needed.

Format flexibility. Crucially, STARK operates on the 10-column tabular structure without assumptions about its content, i.e., the type of values expected in the columns. While we use standard CoNLL-U column semantics throughout this paper, any categorical labels can populate the tag columns (UPOS, XPOS, FEATS), any directed relations can populate the dependency columns (HEAD, DEPREL), including semantic dependencies, and any unit can serve as a token (FORM, LEMMA), in-

[5] https://github.com/clarinsi/STARK/blob/master/settings.md

cluding multi-word sequences, e.g. for sequence-based diversity measurement.

3.2 Diversity computation (DiversUtils)

Diversity computation in DELTA is performed using DiversUtils, a C/Python library that implements diversity metrics from ecological and information-theoretic frameworks. DiversUtils takes the category frequencies extracted by STARK as input. Users specify which diversity metrics to compute via the DiversUtils configuration file. The library currently implements 32 diversity metrics, which can be understood along three complementary dimensions: *variety* (the number of categories), *balance* (the evenness of their distribution), and *disparity* (the degree of difference between categories). This framework, adapted from ecology (Ramaciotti Morales et al., 2021; Lion-Bouton et al., 2022), provides a conceptual structure for understanding what different metrics capture.

Variety measures how many distinct categories are present in the corpus. Simple variety metrics include richness which is just the number of categories, and "species count" which is richness minus one, such that in the minimum case where only one category is present, the diversity scores zero (Patil and Taillie, 1982). Higher variety indicates a greater number of distinct linguistic patterns in the data. Note that variety is sensitive to corpus size: larger corpora can account for a wider set of phenomena, especially rare ones (see the correlation between variety and treebank sentence count in Figures 3 and 4).

Balance measures the evenness of the frequency distribution – whether categories are equally represented or dominated by a few high-frequency categories. Balance is captured by metrics such as Shannon evenness (Ramaciotti Morales et al., 2021) for pure balance or entropies from the set of generalised entropies (Rényi, 1961; Patil and Taillie, 1982) for variety-balance hybrids. Higher balance indicates more uniform usage across categories.

Disparity measures the degree of structural difference between categories, capturing the dissimilarity between them. Unlike variety and balance, disparity requires a distance function. DELTA uses Zhang-Shasha tree edit distance by default (Zhang and Shasha, 1989),[6] which captures linguistically meaningful tree differences by considering both nodes and edges. For single-word trees, Word2Vec cosine distance can be specified as a semantic alternative (using `--w2v_path`).

Multi-dimensional metrics. Many diversity metrics encompass multiple dimensions simultaneously (Chao et al., 2014; Stirling, 2007). For example, Shannon-Wiener entropy in its original definition (Wiener, 1939; Shannon, 1948; Shannon and Weaver, 1949) is a hybrid of variety and balance, increasing when either more categories are present or when frequencies are more evenly distributed. Generalizations of entropy with varying parameters exhibit different weightings of variety and balance (Rényi, 1961; Patil and Taillie, 1982; Hill, 1973). The concept of entropy has been further generalized in ecology by incorporating distances between categories (Chao et al., 2014; Leinster and Cobbold, 2012; Scheiner, 2012), thus accounting for disparity in addition to variety and balance.

Methodological considerations. In the examples in this paper we make the choice of using easily interpretable metrics (richness for pure variety, and Shannon evenness for pure balance). For authors wishing to build a single unified ranking among datasets, a variety-balance hybrid is desirable. Based on the long history of diversity in ecology and biology, it is notably relevant to use Hill (1973) numbers rather than entropies, as Hill assesses that "The diversity numbers N_α, have therefore a natural intuitive interpretation, albeit rather a vague one. The corresponding generalized entropies H_α, being logarithmic, are harder to visualize.". Hill numbers are interpreted as the "effective number of species" if all species were equally probable, to give the same entropy. To add disparity, consider generalised Hill numbers such as that of Chao et al. (2014) as a start. Generalised Hill numbers are interpreted as the "effective number of equally common, equally distinct species or lineages". Conversely, the approach by Stirling (2007) to adding disparity has been criticized (Leydesdorff et al., 2019). Beyond the choice of the measure, when using datasets of non-commensurate sizes, consider averaging diversity scores over numerous samples of same sizes, so as to prevent biases due to dataset size.

3.3 Output and visualisation

DELTA produces two main artifacts from each analysis. First, STARK-produced category frequency lists are (optionally) stored in tab-separated for-

[6] We use the Python `zss` package.

mat (TSV), listing all extracted linguistic patterns with their frequencies (see example in Table 1). These files can be inspected manually to understand which categories were found, or reused independently in external analyses. Second, DiversUtils-produced diversity scores are stored in a SQLite database, providing a structured format optimized for querying and cross-corpus comparison. Each database record includes the corpus identifier, linguistic level analyzed, diversity metric applied, and the computed score.

The repository also includes preconfigured analysis scripts that operate directly on these databases, for example to plot variety against balance for a given collection of corpora (e.g., Figure 2), optionally also indicating the corpus size (e.g., Figures 3 and 4), enabling straightforward identification of diversity patterns across treebanks, languages, or linguistic levels. Users can also write custom queries against the SQLite databases to generate application-specific analyses or export results in alternative formats.

3.4 Availability and execution

DELTA is freely available as open-source software under the joint BSD-2 and CeCILL-B license.[7] The repository includes complete documentation, installation instructions, predefined configurations for common use cases, and example analysis scripts.

The system can be installed as a command-line program for Python, via instructions in the README.md. DELTA is executed via a command-line interface that takes three inputs: (1) a STARK configuration file specifying linguistic patterns to extract, (2) a DiversUtils configuration file specifying diversity metrics to compute, and (3) a list of input corpora in CoNLL-U format. A single command processes all specified corpora and produces the outputs described in Section 3.3.

The accompanying demonstration video,[8] illustrates the DELTA workflow by showing how the system computes and visualizes syntactic diversity over the full UD dataset. It reproduces the results shown in the Appendix (Figure 4), demonstrating the end-to-end pipeline from configuration to visualization.

[7] Hosted at `https://gitlab.lisn.upsaclay.fr/esteve/delta/`. CeCILL-B is the French equivalent of BSD-2, formed by French national scientific institutions.

[8] `https://gitlab.lisn.upsaclay.fr/esteve/delta/-/blob/main/video/DELTA-system-demo-video.mp4`

3.5 Scalability

For large-scale analyses, DELTA supports SLURM-based parallelization through array job submission, enabling efficient processing of multiple configurations across many treebanks simultaneously. In practice, the scale at which DELTA can work depends on the computational expense of both the tree querying and the diversity computation.

For tree querying, some queries may return as little as a constant number of elements per sentence $O(1)$ or as high as an exponential number of elements per sentence $O(x^s)$ where s is sentence size. Likewise, for diversity computation, variety and balance measures often take at most linear time $O(n)$, but disparity takes quadratic time $O(n^2)$, where n is the number of extracted categories.[9] Disparity computation is also sensitive to tree complexity: computing distance matrices for large category sets or complex trees can be intensive, so for the Zhang-Shasha tree edit distance, a configurable timeout (default 0.25s) limits computation time, producing exact or approximate results.

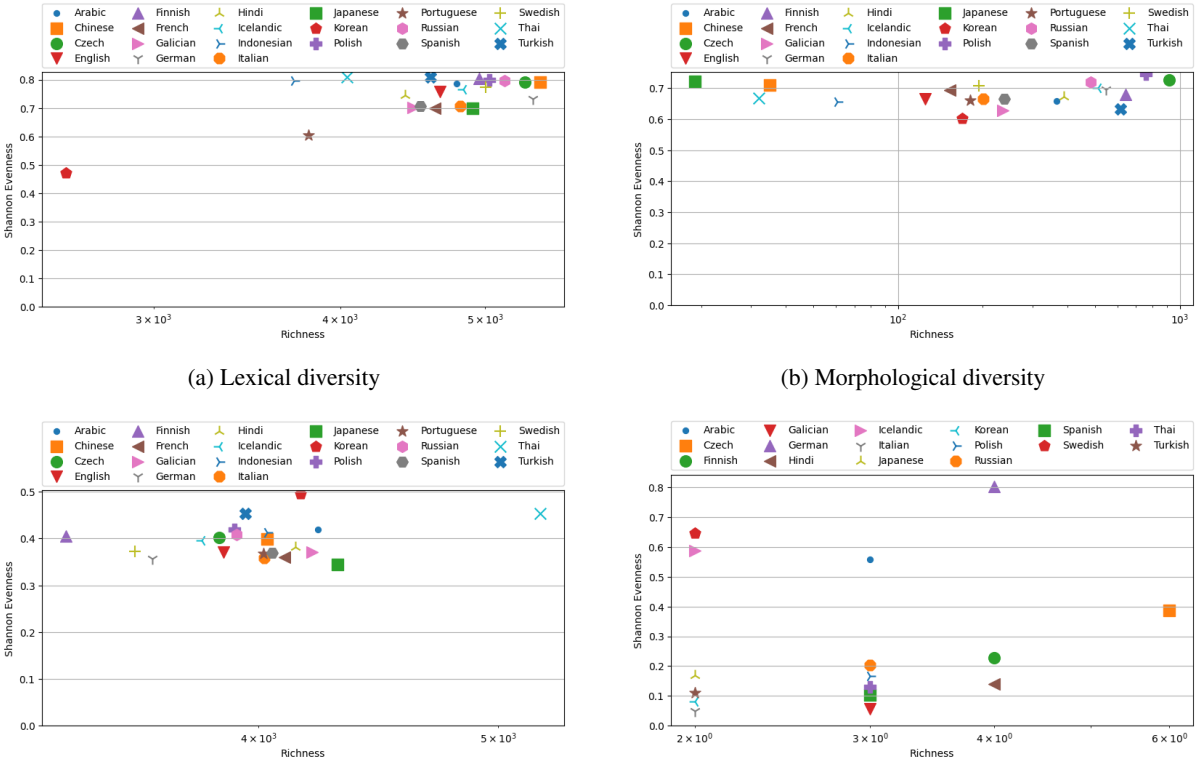
Empirically, inexpensive queries with linear diversity metrics can process billion-token datasets, with substantial time on reading/writing and parsing. In contrast, queries using disparity functions with tree edit distance become computationally intensive: even 1,000 categories can require multiple hours for matrix computation.

To provide a concrete benchmark: processing the entire UD v2.16 release for both lexical diversity (`lemmas.ini`) and syntactic diversity (`syntactic-structures.ini`) with linear metrics (`linear.ini`), to produce results shown in Figures 3 and 4, took 45 minutes total.

4 Evaluation

We validate DELTA's core capabilities through four experiments on Parallel Universal Dependencies (PUD), a collection of parallel treebanks consisting of 1,000 aligned dependency-parsed sentences across 24 languages (Zeman et al., 2017). By controlling for content and corpus size, PUD provides an ideal testbed for systematic cross-linguistic comparison.

[9] See the `linear.ini` and `quadratic.ini` configuration files for respectively at-most-linear, and at-most-quadratic diversity computations.



(a) Lexical diversity

(b) Morphological diversity

(c) Syntactic diversity

(d) SVO order diversity

Figure 2: Richness (number of distinct categories) versus balance (uniformity of their distribution) for 24 PUD treebanks across lexical (a), morphological (b), syntactic (c), and word-order (d) levels.

4.1 Experimental setup

We applied DELTA to measure diversity across four linguistic levels using the predefined configurations described in §3.1: (1) lexical diversity of lemmas; (2) morphological diversity of POS and feature combinations; (3) syntactic diversity of labeled subtrees; and (4) word order diversity of subject-verb-object placement.

For each linguistic level, we computed two complementary diversity dimensions. For variety, we use richness, which is simply the number of distinct categories n. For balance, we use Shannon evenness (Smith and Wilson, 1996; Ramaciotti Morales et al., 2021), which normalizes entropy by maximum entropy for n categories (Equation 1). This metric ranges from 0 (maximally uneven distribution) to 1 (perfectly even distribution).

$$H'(p) = \frac{H(p)}{\log_b(n)} = \frac{-\sum_{i=1}^{n} p_i \log_b(p_i)}{\log_b(n)} \quad (1)$$

4.2 Results

Figure 2 presents diversity scores across four linguistic levels for PUD treebanks.

Lexical diversity (Figure 2a) shows most languages clustering with similar numbers of distinct lemmas (4,000-5,000) that are evenly distributed (evenness 0.7-0.8). Korean stands out as a clear outlier with substantially lower richness and evenness, even in comparison to typologically similar Japanese. This pattern warrants further investigation, but is likely influenced by coarser tokenisation granularity in Korean UD, which retains more morphological material within tokens (Chun et al., 2018).

Morphological diversity (Figure 2b) exhibits the largest variation in richness (100–1,000 distinct morphological property combinations), reflecting expected typological differences in morphological complexity, with fusional languages (e.g., Czech, Polish) and agglutinative languages (e.g., Turkish, Finnish) exhibiting the highest values, though cross-treebank differences in how morphological information is encoded (e.g., feature inventory size and segmentation practices) may also contribute to the observed variation.

Syntactic diversity (Figure 2c), measured here as variety and balance of dependency subtree configurations, shows relatively tight clustering, with

most languages exhibiting comparable number and distribution of such configurations. Finnish and Thai emerge as the two most notable outliers, a pattern that may reflect morphology-syntax trade-offs but requires further investigation to disentangle typological properties from annotation-specific practices.

Subject-verb-object (SVO) order diversity (Figure 2d) reveals clear word order typology, with richness distinguishing fixed-order languages (few configurations)[10] from free-order languages like Czech (all six possible permutations). Evenness captures preference strength: languages with similar richness show very different distributions, from strong word order preferences (low evenness) to even distributions across all available patterns.

These results demonstrate DELTA's capacity for systematic linguistic diversity profiling within and across corpora. While many of the findings above align with established typological patterns, they also highlight the tool's ability to identify potential outliers or unexpected distributions, making it applicable not only to cross-linguistic comparison but to systematic comparisons across datasets more generally (e.g. between genres within a single language). DELTA thus enables researchers to identify and compare diversity patterns at multiple linguistic levels and from different diversity perspectives within a unified analytical framework.

5 Conclusion

We presented DELTA, a unified and configurable framework for computing linguistic diversity of various linguistic features in dependency-parsed corpora. By bridging expressive dependency-tree querying with a broad suite of diversity metrics, visualizations and pre-configured templates, DELTA provides the first integrated environment for systematic, reproducible measurement of diversity for any phenomenon expressible as a dependency (sub)tree – from individual words to complex syntactic patterns. While demonstrated here on standard UD treebanks, the framework's reliance on the CoNLL-U tabular structure rather than specific annotation content makes it also adaptable to alternative annotation schemes and, consequently, a broader range of linguistic phenomena.

Future work will focus on targeted linguistic research questions across specific languages, genres, and linguistic phenomena, as well as on validating the behaviour and interpretability of the adopted diversity metrics across varying corpus sizes and linguistic conditions. To support this, we will further improve computational efficiency, scalability to new formats and phenomena, and the overall user experience – and we welcome community feedback to guide these ongoing developments.

[10] Due to the definition of the balance metric used here, evenness can only be computed when at least two categories are attested. Languages with only one SVO order (e.g., English) therefore do not appear in Figure 2d.

Acknowledgments

We gratefully acknowledge financial support from the UniDive project (COST Action CA21167), the SELEXINI project (ANR-21-CE23-0033), the "Plan blanc" doctoral funding from Université Paris-Saclay (France), the SPOT project (ARIS Z6-4617), the LRTS research program (ARIS P6-0411), and AI4DH (EU HORIZON-WIDERA-2023-TALENTS-01-01, grant 101186647). We thank Agata Savary and Thomas Lavergne for their discussions on the topic of this tool. Generative AI tools were used by one author to support language editing during manuscript preparation.

References

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

Anne Chao, Chun-Huo Chiu, and Lou Jost. 2014. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45:297–324. Publisher: Annual Reviews.

Jayeol Chun, Na-Rae Han, Jena D Hwang, and Jinho D Choi. 2018. Building universal dependency treebanks in korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

Kaja Dobrovoljc. 2025. Counting trees: A treebank-driven exploration of syntactic variation in speech and writing across languages.

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.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics (2021-...)*, 6(1):17.

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.

Miguel R. Guevara, Dominik Hartmann, and Marcelo Mendoza. 2016. diverse: an r package to analyze diversity in complex systems. *The R Journal*, 8:60–78. Https://rjournal.github.io/.

Bruno Guillaume. 2021. Graph matching and graph rewriting: Grew tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175.

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.

Mark Oliver Hill. 1973. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, 54(2):427–432. Publisher: Ecological Society of America.

Shailza Jolly, Sandro Pezzelle, and Moin Nabi. 2021. EaSe: A diagnostic tool for VQA based on answer diversity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2407–2414, Online. Association for Computational Linguistics.

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.

Dijana Kosmajac and Vlado Keselj. 2019. Twitter bot detection using diversity measures. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 1–8, Trento, Italy. Association for Computational Linguistics.

Luka Krsnik and Kaja Dobrovoljc. 2025. STARK: A toolkit for dependency (sub)tree extraction and analysis. In *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 44–51, Ljubljana, Slovenia. Association for Computational Linguistics.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. 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 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom Leinster and Christina A. Cobbold. 2012. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572.

Loet Leydesdorff, Caroline S. Wagner, and Lutz Bornmann. 2019. Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*, 13(1):255–269.

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.

Juhani Luotolahti, Jenna Kanerva, and Filip Ginter. 2017. Dep_search: Efficient search tool for large dependency parsebanks. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 255–258, Gothenburg, Sweden. Association for Computational Linguistics.

Ganapati P. Patil and Charles Taillie. 1982. Diversity as a Concept and its Measurement. *Journal of the American Statistical Association*, 77(379):548–561. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2024. A principled framework for evaluating on typologically diverse languages.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphaël Fournier-S'Niehotta, Rémy Poulain, Lionel Tabourier, and Fabien Tarissan. 2021. Measuring diversity in heterogeneous information networks. *Theoretical Computer Science*, 859:80–115.

Jai Ram Rideout, Greg Caporaso, Evan Bolyen, Daniel McDonald, Yoshiki Vázquez Baeza, Jorge Cañardo Alastuey, Anders Pitman, Jamie Morton, Qiyun Zhu, Jose Navas, Kestrel Gorlick, Justine Debelius, Zech Xu, Matt Aton, llcooljohn, Joshua Shorenstein, Laurent Luce, Will Van Treuren, John Chase, charudatta-navare, Antonio Gonzalez, Colin J. Brislawn, Weronika Patena, Karen Schwarzberg, teravest, Jens Reeder, Igor Sfiligoi, shiffer1, nbresnick, and Dr K. D. Murray. 2025. scikit-bio/scikit-bio: scikit-bio 0.6.3.

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An open infrastructure for advanced treebanking. In *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29. Hajič, Jan.

Alfréd Rényi. 1961. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4.1, pages 547–562. University of California Press.

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.

Farhan Samir and Miikka Silfverberg. 2023. Understanding compositional data augmentation in typologically diverse morphological inflection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 277–291, Singapore. Association for Computational Linguistics.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. PARSEME meets Universal Dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9.

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 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 Underresourced Languages @ LREC-COLING 2024*, pages 372–382, Torino, Italia. ELRA and ICCL.

Samuel M. Scheiner. 2012. A metric of biodiversity that integrates abundance, phylogeny, and function. *Oikos*, 121(8):1191–1202. Number: 8.

Claude Elwood Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(4):623–656.

Claude Elwood Shannon and Warren Weaver. 1949. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Benjamin Smith and J. Bastow Wilson. 1996. A Consumer's Guide to Evenness Indices. *Oikos*, 76(1):70–82. Publisher: [Nordic Society Oikos, Wiley].

Ricard V. Solé, Bernat Corominas-Murtra, and Jordi Fortuny. 2010. Diversity, competition, extinction: the ecophysics of language change. *Interface*, 7(53):1647–1664.

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.

Jan Štěpánek and Petr Pajas. 2010. Querying diverse treebanks in a uniform way. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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

Luka Terčon and Kaja Dobrovoljc. 2025. ComparaTree: A multi-level comparative treebank analysis tool. In *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 129–139, Ljubljana, Slovenia. Association for Computational Linguistics.

Norbert Wiener. 1939. The ergodic theorem. *Duke Mathematical Journal*, 5(1):1–18.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher David Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia,

Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Daniel Zeman et al. 2025. Universal dependencies 2.16. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Kaizhong Zhang and Dennis Shasha. 1989. Simple Fast Algorithms for the Editing Distance between Trees and Related Problems. *SIAM Journal on Computing*, 18(6):1245–1262.

A Appendix

Tree	Freq.
NOUN <nsubj VERB >obj NOUN	142
VERB >nsubj NOUN >obj NOUN	13
NOUN <nsubj NOUN <obj VERB	9
NOUN <obj VERB >nsubj NOUN	5
VERB >obj NOUN >nsubj NOUN	2
NOUN <obj NOUN <nsubj VERB	1

Table 1: Example output TSV listing categories (subject-verb-object trees) for SVO diversity computation in Czech PUD treebank (Figure 2d). Trees are represented using a simplified query-like syntax inspired by the dep_search tool (Luotolahti et al., 2017).

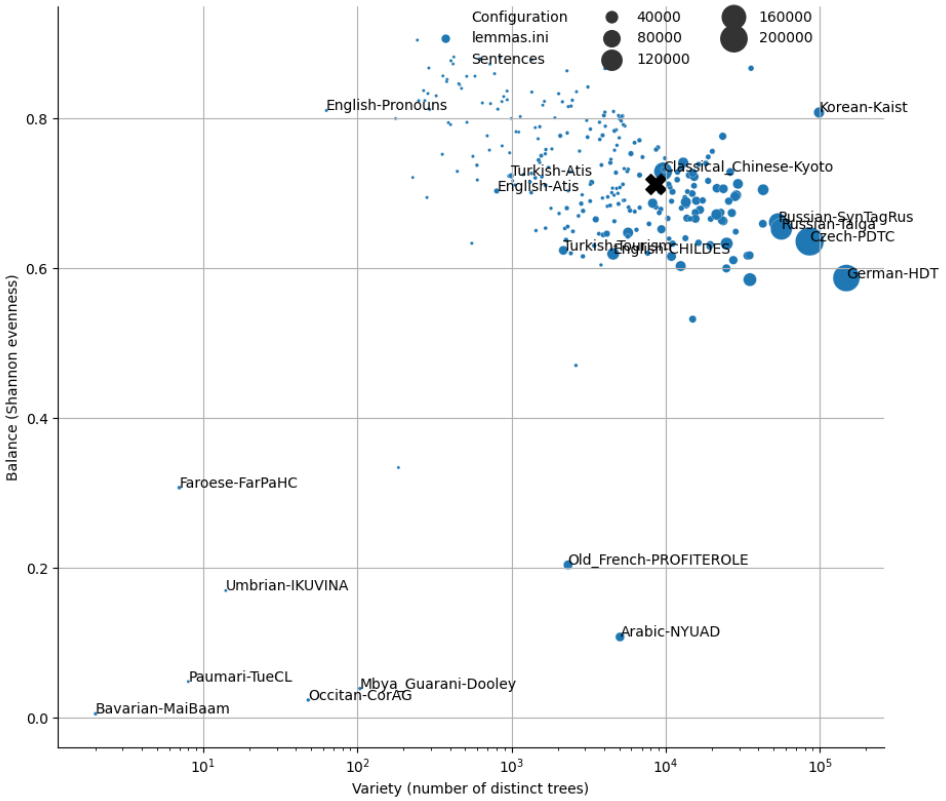


Figure 3: Lexical diversity across UD v2.16 treebanks. Each circle represents a treebank positioned by richness (x-axis) and Shannon evenness (y-axis), with size indicating sentence count. Cross marks (×) show the average diversity across plotted treebanks. For readability, the plot includes only treebanks with at least 100 sentences.

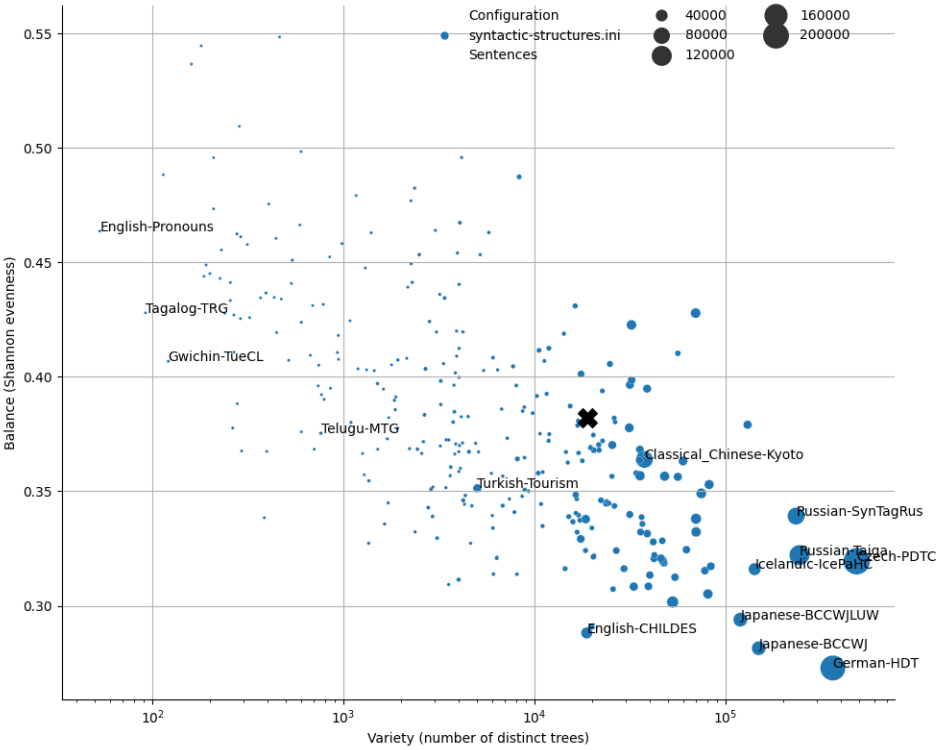


Figure 4: Syntactic diversity across UD v2.16 treebanks. Each circle represents a treebank positioned by richness (x-axis) and Shannon evenness (y-axis), with size indicating sentence count. Cross marks (×) show the average diversity across plotted treebanks. For readability, the plot includes only treebanks with at least 100 sentences.