A new pipeline for measuring diversity across various linguistic levels

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

Relevant UniDive working groups: WG3, WG4

1 Introduction

The growing emphasis on data-driven approaches in computational linguistics has led to increasing research on measuring dataset diversity (Estève et al., to appear) across various linguistic dimensions, such as lexis and syntax (Guo et al., 2023). However, there remains a notable absence of general-purpose tools that can facilitate these analyses within a unified framework, which would also enhance the comparability of results across studies.

To bridge this gap, we present a new pipeline[1]designed to measure dataset diversity from multiple perspectives, which integrates two tools (DiversUtils[2] and STARK[3]) to provide various measures of diversity for various linguistic phenomena. We present the pipeline in the continuation of this paper and demonstrate it by measuring lexical and morphosyntactic diversity of PUD treebanks.

2 Pipeline

In essence, the pipeline quantifies diversity by calculating the number of unique units (types) and their occurrences (items) in a dataset. A "type" refers to a distinct word or unit within the text, while an "item" denotes each individual instance of that type. Diversity is measured on types.

2.1 Type extraction (STARK)

Our pipeline first uses STARK (Krsnik et al., 2024) to extract dependency trees and sub-trees from UD-parsed corpora (*i.e.*, `*.conllu`-formatted data), treating each distinct tree as a type and its occurrences as items. The customizable parameters, specified in the configuration file, allow the user to control which token information to consider as the nodes of the tree (*e.g.*, word forms for extracting lexicalized trees or PoS tags for extracting non-lexicalized trees) and set restrictions to focus on specific types of trees, such as those headed by a particular PoS, those featuring certain dependency relations, or those pertaining to a specific pre-defined pattern. STARK thus enables users to define a wide range of syntactic types (from single tokens to complex structures) with various degrees of specificity (from specific predefined patterns to all possible structures) on multiple linguistic levels (from lexis to morphosyntax). We illustrate two such configurations in Section 3.

2.2 Diversity computation (DiversUtils)

Diversity may be understood through three dimensions (Morales et al., 2020; Lion-Bouton et al., 2022), each of which contains numerous equations: (1) *variety* focuses on the number of types, (2) *balance* focuses on the evenness in the distribution of types, (3) *disparity* focuses on the fundamental differences (or in practice distances based on some function) between types. It should be noted that diversity functions can encompass multiple dimensions at once (Chao et al., 2014; Stirling, 2007).

While all dimensions are of interest and DiversUtils implements dozens of functions, for conciseness we here discuss simple functions (*i.e.*, theoretically reasonable, understandable, and not encompassing multiple dimensions at once): richness for variety – the number of types n – and Shannon evenness for balance (eq. 1), which is entropy divided by maximum entropy for n types; thus it reaches 1 when the distribution is even, and goes towards 0 as the distributions gets increasingly uneven (Smith and Wilson, 1996; Morales et al., 2020).

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

2.3 Output

The pipeline first generates a tabular file containing the extracted types along with their number

[1] https://gitlab.lisn.upsaclay.fr/esteve/delta
[2] https://github.com/estevelouis/WG4
[3] https://github.com/clarinsi/STARK

of items and various metadata.It then computes different diversity functions, possibly with different parameters, and stores the scores in a database. The large number of scores in the database can be queried by other accompanying scripts to generate plots that facilitate data analysis.

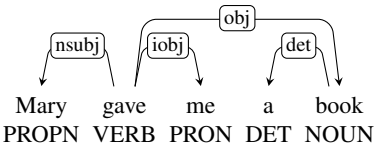
3 Selected use case

3.1 Dataset

Parallel Universal Dependencies treebanks (PUD) is a collection of parallel treebanks created as part of the CoNLL 2017 Shared Task (Zeman et al., 2017), which consist of 1,000 aligned sentences in 24 languages, with most sentences being originally in English and translated to other languages. It thus ensures the absence of size- and genre-related biases and allows us to compare the diversity of language describing the same semantic content.

3.2 Type definition

For **lexical diversity**, we extracted the list of all lemma types and instances from each of the treebank, which – in terms of STARK configuration (*cf.*, Section 2.1) – means extracting all incomplete trees of size 1 (single node) with lemma as the node type. In the example below, the resulting list of lexical types would be *Mary*, *give*, *I*, *a*, *book*.



For **morphosyntactic diversity**, we define our type as any labeled (sub-)tree occurring in the treebank, regardless of size, with the PoS category taken as the node type. The list of types extracted from the sentence in the example above would thus consist of the five trees we get when 'cutting' the tree at each word: the full sentence tree, the tree of the object ('DET <det NOUN') and the three single-token trees with terminal nodes ('PROPN', 'PRON', 'DET').

3.3 Results

We see in Figures 1 and 2 the richness and Shannon evenness of PUD treebanks for lexical and morphosyntactic diversity, respectively. Both plots reveal that most languages behave similarly, clustering around a specific area of variety and balance.

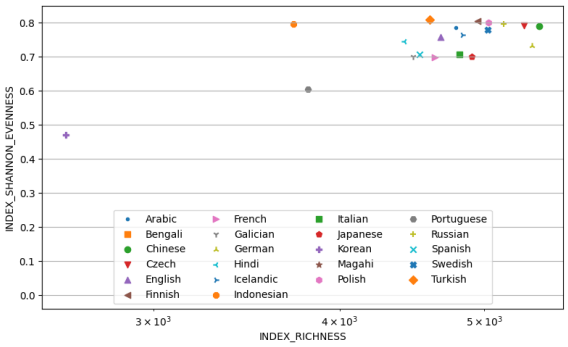


Figure 1: Lexical diversity with lemmas as types, on PUD. Variety (richness n) on the lower axis, and balance (Shannon evenness) on the upper axis.

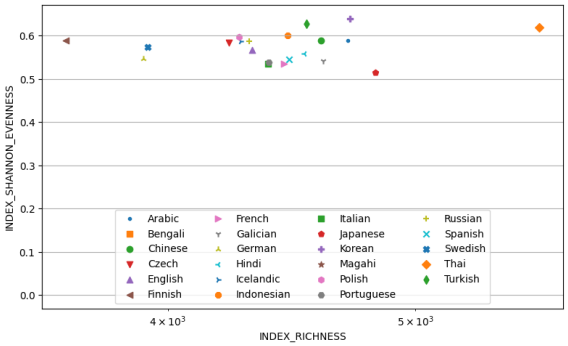


Figure 2: Morphosyntactic diversity with labeled trees as types, on PUD. Variety (richness n) on the lower axis, and balance (Shannon evenness) on the upper axis.

This indicates a comparable number of different lemmas and dependency trees, with an even distribution across their treebanks. However, some outliers exist: Indonesian, Portuguese, and Korean in terms of lexical diversity, and Finnish and Thai in morphosyntactic diversity. These are likely attributable to distinctive linguistic features, such as rich compounding, affixation, and/or inflection in agglutinating languages. However, further data analysis is needed to rule out treebank-specific annotation practices or other potential explanations.

4 Conclusion

We introduced a new configurable pipeline for measuring various dimensions of diversity in parsed data, along with two experiments that highlight the variation in lexical and morphosyntactic diversity across languages using semantically aligned data. In addition to the open-source release of the pipeline, our future work will explore the impact of genre on linguistic diversity using this framework.

Acknowledgements

This work was made possible thanks to the help of 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), and the LRTS research program (ARIS P6-0411).

References

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.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text. ArXiv:2311.09807 [cs].

Luka Krsnik, Kaja Dobrovoljc, and Marko Robnik-Šikonja. 2024. Dependency tree extraction tool STARK 3.0. Slovenian language resource repository CLARIN.SI.

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.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphael Fournier-S'niehotta, Remy Poulain, Lionel Tabourier, and Fabien Tarissan. 2020. Measuring Diversity in Heterogeneous Information Networks. ArXiv:2001.01296 [cs, math].

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

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.

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 D. 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.