UniDive 2nd Training School

Course proposal on diversity quantification

Louis Estève, Marie-Catherine de Marneffe,
Nurit Melnik, Agata Savary, Olha Kanishcheva

Title of the course

Diversity quantification in natural language processing: The why, what, where and how

Abstract, including the topics to be covered

The notion of diversity has been gaining increasing attention in natural language processing (NLP) over the past few years. A survey of all papers in the ACL anthology between 1990 and the middle of 2024 found that of the 411 papers whose title included the terms “diverse” or “diversity”, 75% of them were recent papers, published since 2019 (Estève et al., under review). In parallel to its rising popularity, diversity has been addressed in an ad hoc manner in NLP, and with few explicit links to other domains where this notion is better theorized.

This course presents an attempt to adopt a unifying framework, inspired by ecology and applied in various scientific domains such as economics, physics and political science. In this framework, a system’s elements are assigned to categories, and diversity measures are positioned along three dimensions: variety, balance and disparity (Stirling, 2007). This seven hours and a half course provides training in measuring and understanding diversity within NLP, combining theoretical foundations with hands-on practical experience.

The first session introduces diversity quantification taxonomy through a 1.5-hour lecture that establishes the theoretical foundations. The lecture first introduces core concepts from ecological diversity theory including elements, categories, variety, balance, and disparity. It then presents findings from the survey regarding four central questions: **why** diversity is important in NLP, **what** types of “diversities” are measured, **where** diversity is measured across NLP areas and pipeline stages, and **how** diversity is quantified.

The second session transitions theory into practice through a 1.5-hour interactive workshop in which participants work in small groups and apply the taxonomy methodology to real research papers. The session ends with short group presentations.

The third and fourth sessions provide hands-on experience with actual diversity quantification through two 1.5-hour practical workshops focused on measuring diversity. Participants work with a number of datasets, and using existing software libraries, developed by one of the trainers in the framework of the UniDive COST action, they learn to establish diversity quantification pipelines, analyze resulting diversity scores, and compare the datasets with respect to these metrics. The third session focuses on in-text diversity, which involves categories that are associated with linguistic properties implicitly assumed to be inherent to a text (e.g. vocabulary or semantics). In the fourth session, the focus is on metalinguistic diversity, where categories are classifications of texts in datasets (e.g. their languages or domains);

Upon completion, participants will possess both theoretical understanding and practical skills in NLP diversity quantification. The course emphasizes the transition from intuitive diversity assumptions to rigorous quantitative assessment, preparing participants to contribute meaningfully to this rapidly growing field in computational linguistics.

Level (introductory, intermediate or advanced course)

intermediate (no pre-requisites about diversity, but necessary pre-requisites from linguistics and/or computation)

Draft schedule, with a brief description of each session

1. Session 1 (1h30min): *Diversity quantification taxonomy* - a lecture summarizing the Estève et al. (under review) survey of 300+ papers about diversity in NLP from 2019-2024.
    a. Learning outcomes:
        i. Understanding the importance of diversity framework in modern NLP
        ii. Understanding the unified taxonomy for diversity quantification
    b. Contents:
        i. Statistics about ACL Anthology showing a dramatic increase of interest in diversity
        ii. Foundations of the theory of diversity in ecology: elements, categories, variety, balance, disparity. Toy example.
        iii. Diversity taxonomy for NLP with 4 dimensions:
            1. Why diversity is important in NLP
            2. What diversity is measured on: in-text diversity, metalinguistic diversity and diversity of processing
            3. Where diversity is measured: NLP areas, NLP pipeline stages
            4. How diversity is measured - families of diversity measures based on: quantity, entropy, pairwise-distances, distribution distances, overlap and human judgement
            5. Discussion
2. Session 2 (1h30min): *Taxonomy in practice* – working on research papers (either given by the presenting staff or brought by trainees) to apply the methodology of the survey.
    a. Learning outcomes:
        i. Practical application of the diversity taxonomy to analysing research work
        ii. Understanding the benefits of this unified framework for comparability of approaches
    b. Contents:
        i. Trainers present how to process a paper with the methodology of the survey
        ii. Trainees form groups and work on their specific papers
        iii. Presenters walk between groups to discuss with them
        iv. Each group presents the result of casting the paper on the taxonomy
        v. The results are consolidated by the trainers before session 3 and displayed online

3. Session 3 (1h30min): *Practical session* – Measuring **in-text diversity** to familiarise trainees with the practice of actual diversity quantification.
    a. Learning outcomes:
        i. Be able to set up a pipeline for quantifying diversity in a dataset
        ii. Understand the importance of actual diversity quantification instead of just relying on intuitions of diversity
    b. Contents
        i. Use an existing library to quantify diversity in a dataset believed to be diverse but with no actual diversity quantification (the BNC corpus)
        ii. Analyse and discuss the diversity scores
        iii. Compare its diversity scores to another "random" dataset
4. Session 4 (1h30min): *Practical session* – Measuring **metalinguistic diversity** in a multilingual dataset
    a. Learning outcomes:
        i. Understand the difference between in-text and metalinguistic diversity
        ii. Understanding the relativity of the notion of diversity: how and why the diversity scores for the same data set change with the choice of elements and categories
    b. Contents:
        i. Comparing diversity rankings of treebanks in Universal Dependencies, based on different choices of categories/elements
        ii. Group work: pen-and-paper exercise on designing a maximally diverse dataset out of Universal Dependencies treebanks
5. Session 5 (1h30min): Trainees' presentations of their outcomes from sessions 2 and 4

Necessary infrastructure

Beamer, personal laptops, printed papers for the trainees, pens and papers for group work

Prerequisites for the trainees

- Reading a research paper about diversity (to be used in exercises)
- Reading the other required reading
- Installing a library (or alternatively, using a Jupyter notebook or Google collab) for the practical exercises
- Understanding the CoNLLU format of Universal Dependencies

Bibliography

Required readings

- Estève, L., de Marneffe, M.-C., Melnik, N., Savary, A., Kanishcheva, O., under review at TACL.
- Ploeger, E., Poelman, W., de Lhoneux, M., and Bjerva, J. (2024). What is "typological diversity" in NLP? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, Proceedings of

the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5681–5700, Miami, Florida, USA. Association for Computational Linguistics [link]

- A different diversity-related paper for each group of trainees - to be sent by the trainers shortly before the course.

Further readings

- Lion-Bouton, A., Ozturk, Y., Savary, A., and Antoine, J.-Y. (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.
- Stirling, A. (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.

Instructors, their affiliation, contact details and experience

- Louis Estève, Université Paris-Saclay, CNRS, LISN lab, France, louis.esteve@universite-paris-saclay.fr
  Louis Estève is a second-year computer science PhD student at Université Paris-Saclay. His topic is the quantification of diversity of linguistic phenomena in corpora and system predictions. He is a member of the UniDive COST action and has focused primarily on the Working Group 4 (WG4) aimed at defining relevant diversity functions.
- Olha Kanishcheva, Heidelberg University, SET University, kanichshevaolga@gmail.com
  Olha Kanishcheva is a researcher at Heidelberg University (Germany) with over 10 years of experience in the field of NLP. Her expertise covers both practical and academic aspects of the field. She teaches NLP to master's students and is a participant in the UniDive COST Action project (WG3 and WG4).
- Marie-Catherine de Marneffe, UCLouvain, marie-catherine.demarneffe@uclouvain.be. Marie-Catherine de Marneffe has taught introductory-level CL and Linguistics courses to mixed audiences both at the bachelor and master levels in the USA for 10 years. Since 2022, she has been teaching in Belgium at the master level. She is one of the principal developers of the UD framework and a co-leader of the WG4 on quantifying and promoting diversity in the UniDive COST Action.
- Nurit Melnik, The Open University of Israel, nuritme@openu.ac.il Nurit has written textbooks for courses in linguistics at the Open University during the past 15 years, and has previously taught courses in linguistics and corpus linguistics for language teachers. Her co-authored textbook titled "Computational Literacy for the Humanities" will be published by Routledge in June 2025. She is an active member of UniDive, mainly in WG4.
- Agata Savary, Université Paris-Saclay, CNRS, LISN lab, France <agata.savary@universite-paris-saclay.fr>. Agata Savary is full professor in computer science and has a long-standing experience in academic teaching. She was a trainer in 3 international summer schools: the 1st UniDive training school 2024 in Moldova, LCL 2019 in Pavia, Italy, ESSLLI 2018, Sofia, Bulgaria. She also served as co-organiser of 3 training schools in the PARSEME and UniDive COST action and as tutorial chair at ACL 2020.