Linguistic Typology for NLP researchers: Methods and Resources in the 21st century

Proposal for UniDive winter training school, Jan 20-24, 2026 (Yerevan)

Harald Hammarström
Department of Linguistics and Philology, Uppsala University

Title:	Linguistic Typology for NLP researchers: Methods and Resources in the 21st century
Level:	Introductory
Necessary infrastructure:	-
Prerequisites:	Python
Bibliography:	No reading preparations required. Materials by the teacher plus reference materials (Dixon 2012, Eberhard et al. 2025, Forkel 2025, Hammarström et al. 2025, Haspelmath 2023, McCarthy et al. 2020, Seifart et al. 2024) .
Instructors:	Professor Harald Hammarström, Uppsala University, harald.hammarstrom@lingfil.uu.se, Background in NLP now working on Linguistic Typology and Quantitative Methods

Abstract

The *Linguistic Typology for NLP researchers: Methods and Resources in the 21st century* aims to introduce linguistic diversity, resources and principles for language metadata, language-independent methods for linguistic analysis, and discuss the most important implications of linguistic variation for NLP. We first present the full extent of language diversity as it is known

today and by what principles language can be defined. We then show how current data on languages, dialects, families, countries, coordinates etc can be obtained and used in practice. We present modern linguistic theory from typology whereby concepts that must exist by necessity (e.g., morpheme) are separated from those that exist in specific languages and discuss how languages can be compared. Finally we explain the virtues and limits of currently available NLP resources for linguistic diversity and their implications for language-neutral NLP techniques.

Draft Schedule

Session	Duration	Content
1	2h	Languages of the world: The language situation in the world, types of languages (modality, genre, …), defining and discretizing languages
2	2h	Hands-on working with language diversity resources: Classification, speaker numbers, maps
3	2h	Linguistic analysis: morphemes, words, clauses, IGT, grammars and comparative concepts
4	2h	Linguistic diversity and NLP: Resource poverty, word order, morphological typology and their implications for NLP

References

Dixon, R.M.W. 2010, 2010, 2012. *Basic Linguistic Theory*. Oxford: OUP. 3 vols.

Eberhard, David M., Gary F. Simons & Charles D. Fennig. 2025. *Ethnologue: Languages of the World*. 28th edn. Dallas: SIL International.

Forkel, Robert. 2025. Pyglottolog and Glottolog-CLDF. `https://github.com/glottolog/pyglottolog`, `https://github.com/glottolog/glottolog-cldf`.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2025. Glottolog 5.2. Leipzig: Max Planck Institute for Evolu-

tionary Anthropology. Available at http://glottolog.org. Accessed on 2025-05-30.

Haspelmath, Martin. 2023. Defining the word. *WORD* 69(3). 283–297.

McCarthy, Arya D. , Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post & David Yarowsky. 2020. The johns hopkins university bible corpus: 1600+ tongues for typological exploration. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 2877–2885. Marseille, France: European Language Resources Association.

Seifart, Frank, Ludger Paschen & Matthew Stave. 2024. *Language Documentation Reference Corpus (DoReCo) 2.0*. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).