Quantitative Typology — New Directions in Assembling and Analyzing Cross-Linguistic Data

Annika Tjuka (Department of Linguistic and Cultural Evolution / Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany),
Christian Bentz (Chair for Multilingual Computational Linguistics, University of Passau, Germany)
Johann-Mattis List (Chair for Multilingual Computational Linguistics, University of Passau, Germany)

Description of the Workshop

Linguistic typology has long been a data-driven discipline, but it was not until the recent quantitative turn in linguistics that computational approaches have become integral to the field. Over the past decades, the use of computer-assisted methods has opened up new avenues for typological research, allowing researchers to combine data from different sources. This led to a steep increase in the availability of large typological databases (e.g., Rzymski et al. 2020; List et al. 2022; Skirgård et al. 2023; Zalizniak et al. 2024). The development of these databases allows researchers to ask novel questions and study phenomena in typologically diverse languages on a large scale. Our workshop aims to explore possible future directions arising from new methods, with a particular focus on assembling and analyzing cross-linguistic data using quantitative approaches. This quantitative typology brings with it new opportunities and challenges for research on language comparison (Carling & Verkerk 2024).

In our workshop, we concentrate on two key aspects that require careful attention and deliberate implementations:

- **Increased data availability**: With the increase of computational approaches and online resources, researchers now have access to a vast amount of linguistic data from various sources.
- **New research questions**: Quantitative methods enable researchers to ask novel questions about language, such as those related to psychology, lexicon, and other areas of linguistic inquiry.

By leveraging computer-assisted approaches for language comparison, researchers are now able to aggregate data from different sources. For example, data on words and concepts are available in the research fields of comparative linguistics and psychology. While most of these data are open- access, researchers from both disciplines were either not aware or not able to combine and compare word properties across languages. The development of the Cross-Linguistic Database of Norms, Ratings, and Relations (NoRaRe, Tjuka et al. 2022) showed that it was possible to aggregate data from various sources by utilizing the standards proposed by the Cross-Linguistic Data Formats (CLDF, Forkel et al. 2018). Another example is the aggregation of a large number of multilingual word lists compiled in Lexibank (List et al. 2022), or the collection of grammatical features in Grambank (Skirgård et al. 2023). These databases highlight the diversity and scope of the available linguistic data. Although these databases have drawbacks (Schapper & Koptjevskaja-Tamm 2022), they opened up new possibilities for typologist to explore linguistic features across diverse languages.

The availability of large amounts of data makes it possible to investigate linguistic features with a computational approach and ask new questions. By using statistical analysis and computational modeling, researchers can now examine complex phenomena such as semantic change (e.g., Xu et al. 2017), variation across semantic domains (e.g., Jackson et al. 2019), or language complexity (e.g., Bentz et al. 2023). The phenomenon of colexifications (François 2008) in particular has received considerable attention from cognitive scientists and data from the Database of Cross-Linguistic Colexifications (CLICS, Rzymski et al. 2020) have been used to analyze the interplay between language and cognition, for example, with respect to the universal forces that shape the lexicon (Brochhagen & Boleda 2022). In addition, colexifications have been utilized in studies of Natural Language Processing (e.g., Karidi et al. 2024), another area in which linguistic typology is becoming increasingly important (Bender 2016).

Central to this workshop is the focus on innovative data aggregation techniques and their application in typological studies. We invite contributions that not only demonstrate novel ways of gathering and integrating data but also employ quantitative analysis to tackle complex research questions. By showcasing these advancements, we hope to foster new directions for rigorous data standards and collaborative efforts in advancing the field.

References

Bender, Emily M. 2016. Linguistic Typology in Natural Language Processing. *Linguistic Typology* 20(3). 645–660. https://doi.org/10.1515/lingty-2016-0035.

Bentz, C., X. Gutierrez-Vasques, O. Sozinova & T. Samardžić. 2023. Complexity trade-offs and equi-complexity in natural languages: A meta-analysis. *Linguistics Vanguard* 9(s1). 9–25.

Brochhagen, Thomas & Gemma Boleda. 2022. When Do Languages Use the Same Word for Different Meanings? The Goldilocks Principle in Colexification. *Cognition* 226. 1–8. https://doi.org/10.1016/j.cognition.2022.105179.

Carling, Gerd & Annemarie Verkerk. 2024. Editorial: The Evolution of Meaning: Challenges in Quantitative Lexical Typology. *Frontiers in Communication* 9. 1–3. https://doi.org/10.3389/fcomm.2024.1476702.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics. *Scientific Data* 5(1). 1–10. https://doi.org/10.1038/sdata.2018.205.

François, Alexandre. 2008. Semantic Maps and the Typology of Colexification: Intertwining Polysemous Networks Across Languages. In Martine Vanhove (ed.), *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations*, 163–215. John Benjamins. https://doi.org/10.1075/slcs.106.09fra.

Karidi, Taelin, Eitan Grossman & Omri Abend. 2024. Locally Measuring Cross-lingual Lexical Alignment: A Domain and Word Level Perspective. In Yaser Al-Onaizan, Mohit Bansal & Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15877–15893. (12 November, 2024). Miami, Florida, USA: Association for Computational Linguistics. https://aclanthology.org/2024.findings-emnlp.932.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell D. Gray. 2022. Lexibank, a Public Repository of Standardized Wordlists with Computed Phonological and Lexical Features. *Scientific Data* 9(1). 1–16. https://doi.org/10.1038/s41597-022-01432-0.

Rzymski, Christoph, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, et al. 2020. The Database of Cross-Linguistic Colexifications, Reproducible Analysis of Cross-Linguistic Polysemies. *Scientific Data* 7(1). 1–12. https://doi.org/10.1038/s41597-019-0341-x.

Schapper, Antoinette & Maria Koptjevskaja-Tamm. 2022. Introduction to Special Issue on Areal Typology of Lexico-Semantics. *Linguistic Typology* 26(2). 199–209. https://doi.org/10.1515/lingty-2021-2087.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, et al. 2023. Grambank Reveals the Importance of Genealogical Constraints on Linguistic Diversity and Highlights the Impact of Language Loss. *Science Advances* 9(16). 1–15. https://doi.org/10.1126/sciadv.adg6175.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2022. Linking Norms, Ratings, and Relations of Words and Concepts Across Multiple Language Varieties. *Behavior Research Methods* 54. 864–884. https://doi.org/10.3758/s13428-021-01650-1.

Zalizniak, Anna A., Anna Smirnitskaya, Maksim Russo, Ilya Gruntov, Timur Maisak, Dmitry Ganenkov, Maria Bulakh, et al. (eds.). 2024. *Database of Semantic Shifts*. Moscow: Institute of Linguistics, Russian Academy of Sciences. http://datsemshift.ru.