Chances and Challenges in Compiling a Cross-Linguistic Database of Morpheme-Annotated Wordlists

Katja Bocklage
University of Passau

Keywords: morpheme annotation, cross-linguistic, database compilation, wordlists

In contrast to a dictionary, in which words for a given language are organized by headwords whose meanings are glossed and explained, wordlists start from a typically fixed list of concepts which are then translated into one or more languages. Wordlists are widely used in historical linguistics and lexical typology. In historical linguistics, they are used to search for cognate words in genetically related languages (List et al. 2018), or as the basis of phylogenetic reconstruction (Hoffmann et al. 2021). In lexical typology, they are used to study patterns of colexification in various forms (François 2008, Rzymski et al. 2020).

In the past decade, the compilation of wordlists has increased drastically, as reflected not only in various new wordlists that have been compiled to study different language families (List et al. 2022, Dellert et al. 2020), but also in new standards that have been proposed to handle various problems of annotation, including partial cognates (List 2016), sound correspondences (Wu et al. 2020), and patterns of lexical motivation (Hill and List 2017, Schweikhard and List 2020). Despite these efforts, however, there are only a few examples in which scholars have consistently tried to annotate larger wordlists at the level of the morpheme, indicating cognacy not only externally – across languages –, but also internally – inside one and the same languages. Such a collection of morpheme-annotated wordlists would be very useful in various respects. First, it would allow us to train methods for automated morpheme segmentation on low resource languages. Current methods for morpheme annotation make use of large dictionaries for languages that are generally well documented. In a low resource setting, morpheme annotation methods show a very disappointing performance (List 2023). Second, such a collection would allow us to study pathways of lexical motivations in unprecedented detail. So far, lexical motivation – the formal and semantic processes by which new words are derived from existing ones (Koch 2001) – has mostly been investigated in individual languages (Koch and Marzo 2007), and few studies have looked at typological motivation patterns (Urban 2012).

In the talk, we will present our efforts to employ newly developed annotation techniques in order to compile a cross-linguistic database of morpheme annotated wordlists. We will report on the formats that we test and present the results of an initial prototype that we are currently creating, consisting of morpheme-annotated wordlists in 10 different languages from four different language families.

References

Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, H. I., Baysarova, Z., Mühlenbernd, R., Wahle, J. & Jäger, G. (2020). NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Lang Resources & Evaluation*, 54, 273–301. DOI: 10.1007/s10579-019-09480-6.

François, A. (2008). Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In: M. Vanhove (Ed.), *From Polysemy to Semantic Change*. Towards a typology of lexical semantic associations (163-215). Amsterdam and Philadelphia: John Benjamins. DOI: 10.1075/slcs.106.09fra.

Hill, N. W. & List, J.-M. (2017). Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting*, 3(1), 47-76. DOI: 10.1515/yplm-2017-0003.

Hoffmann, K., Bouckaert, R., Greenhill, S. J. & Kühnert, D. (2021). Bayesian phylogenetic analysis of linguistic data using BEAST. *Journal of Language Evolution*, 6(2), 119–135. DOI: 10.1093/jole/lzab005.

Koch, P. (2001). Lexical typology from a cognitive and linguistic point of view. In M. Haspelmath, E. König, W. Oesterreicher & W. Raible (Eds.), *Language Typology and Language Universals* (Vol. 2/2)(1142-1178). Berlin/New York: de Gruyter. DOI: 10.1515/9783110194265-022.

Koch, P. & Marzo, D. (2007). A two-dimensional approach to the study of motivation in lexical typology and its first application to French high-frequency vocabulary. *Studies in Language*, 31, 2(2007), 259-291.DOI: 10.1075/sl.31.2.02koc.

List, J.-M. (2016). Beyond cognacy: historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*, 1(2), 119–136. DOI: 10.1093/jole/lzw006.

List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T. & Forkel, R. (2018). Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, 3(2), 130–144. DOI: 10.1093/jole/lzy006.

List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J. & Gray, R. D. (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(316), 1-16. DOI: 10.1038/s41597-022-01432-0.

List, J.-M. (2023). Open Problems in Computational Historical Linguistics. *Open Research Europe*, 3(201). DOI: 10.12688/openreseurope.16804.1.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V.,Bodt, T. A., Hantgan, A., Kaiping, G. A., Chang, S., Lai, Y., Morozova, N., Arjava, H., Hübler, N., Koile, E., Pepper, S., Proos, M., Van Epps, B., Blanco, I., Hundt, C., Monakhov, S., Pianykh, K. Ramesh, S., Gray, R. D., Forkel, R. & List, J.-M. (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies. *Scientific Data*, 7(13), 1-12. DOI: 10.1038/s41597-019-0341-x.

Schweikhard, N. E., List, J.-M. (2020). Developing an annotation framework for word formation processes in comparative linguistics. *SKASE Journal of Theoretical Linguistics*, 17(1), 2-26.

Urban, M. (2012). *Analyzability and semantic associations in referring expressions. A study in comparative lexicology* (PhD dissertation). Leiden University.

Wu M.-S., Schweikhard, N. E., Bodt, T. A., Hill, N. W. & List, J.-M. (2020). Computer-Assisted Language Comparison: State of the Art. *Journal of Open Humanities Data*, 6, 2. DOI: https://doi.org/10.5334/johd.12.