Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle and Gerhard Jäger
(Project EVOLAEMP, Universität Tübingen)

A deep-coverage lexical database of Northern Eurasia

Our major lexical data collection project aims to provide a model of an entire geographical area which is detailed enough to feature most regular sound correspondences for each pair of languages, and for evaluating models of language contact. We therefore move beyond the Swadesh-style lists of basic concepts, and attempt to include a relevant subset of the cultural vocabulary as well. Coverage ranges from month names to names for different metals, and also includes some basic agricultural and political concepts. Altogether, we aim at complete coverage of a 1,016-concept list.

The database covers a large portion of the basic vocabulary across the many language families of Northern Eurasia. Among other massively cross-linguistic lexical resources, it can only be compared to the IDS (Intercontinental Dictionary Series) in depth. In comparison to the IDS, our database has the disadvantage of not consisting of expert contributions, but being based exclusively on dictionaries and other written sources available to us. At the same time, it has the advantages of spanning one continuous geographic area, and providing a unified phonetic format across all lexemes. The non-dependence on expert judgments also means that while early versions will contain more errors, continuous improvement will be easier to organize.

We already have near-complete data for 104 languages from the following language families: Indo-European (31 languages), Uralic (26), Turkic (7), Afroasiatic (7), Northeast Caucasian (6), Dravidian (4), Mongolic (3), Tungusic (3), Eskimo-Aleut (3), Chukotko-Kamchatkan (2), Yukaghir (2), Northwest Caucasian (2), Kartvelian (1), Yeniseian (1), Nivkh (1), Ainu (1), Korean (1), Japonic (1), Basque (1), and Burushaski (1). As of July 21st, we have managed to obtain some information for 96% of all language-concept pairs. For 85% of entries, we expect the data to be final. For the other 15% of the data, we are currently in the process of seeking expert and native speaker feedback, which will be incorporated during the next two years.

For almost every language (all except English, Danish, Irish, and French), we developed an automated grapheme-to-phoneme converter, which allows us to collect the data in orthographic form, while remaining able to derive phonetic representations in various formats as needed. While the quality of transcriptions is somewhat lower than could be achieved by manual transcription, this approach ensures consistency across all forms for a given language, and makes it much easier to revert a problematic decision for some symbol, or to extend coverage of a difficult phenomenon like unwritten epenthetic vowels. This flexibility again paves the ground for future incremental improvements in data quality.

Pre-release versions of the database are already in extensive use within our project, where we are using it to evaluate new methods for cognacy and sound correspondence detection, and determining the directionality of lexical influences between languages. The Uralic part of the database can now

be considered almost final, and has been available to the public for some time. The other parts of the database are scheduled for release under an open license until the end of 2017.