UraLex basic vocabulary dataset

Kaj Syrjänen, Jyri Lehtinen, Outi Vesakoski, Mervi de Heer, Toni Suutari, Michael Dunn, Urho Määttä & Unni-Päivä Leino

Overview

The UraLex basic vocabulary dataset has its origins in the basic vocabulary cognacy dataset collected by the research initiative BEDLAN (Biological Evolution and the Diversification of Languages), funded by the Kone Foundation between 2009-2013. The data has since been revised and expanded in follow-up research projects, including SumuraSyyni (2014-2016), UraLex (2014-2016) and AikaSyyni (2017-2020). The dataset has been compiled especially for the purposes of quantitative language classification/historical linguistics, such as Bayesian Inference of phylogeny.

The bulk of the data was originally compiled by Jyri Lehtinen, the principal compiler of the BEDLAN research initiative's cognacy dataset. It has since been expanded by the UraLex project, with the bulk of the editing work done by Kaj Syrjänen and Jyri Lehtinen. A number of individuals have made contributions to the dataset at its various stages, including Hilkka Ahola, Zoltán Balogh, Rogier Blokland, Natalia Chinaeva, Robert Forkel, Nikolett F. Gulyás, Mikko Heikkilä, Katri Hiovain, Terhi Honkola, Sulev Iva, Santra Jantunen, Eino Koponen, Svetlana Lumme, Luke Maurits, Eve Mikone, Arto Moisio, Larisa Ponomareva, Michael Rießler, Tapani Salminen, Merja Salo, Olga Titova, Marja Torikka, Judit Varga, Susanna Virtanen, Kaarina Vuolab-Lohi and Evgenia Zhivotova.

The dataset mainly covers lexical reflexes (e.g. words and expressions) denoting 313 meanings. Most of these meanings (226) come from standardized *basic vocabulary* lists, which cover meanings that are fairly universal, culturally neutral, and generally expressed by semantically and morphologically simple words which are relatively stable over time. Basic vocabulary words also tend to be resistant to being replaced by borrowing or semantic shift (McMahon & McMahon 2005). The 226 basic vocabulary meanings of the dataset cover the contents of three standardized basic vocabulary lists: the Swadesh200 list (Swadesh 1952), the Swadesh100 list (Swadesh 1955), and the Leipzig-Jakarta list (Tadmor 2009). The 87 meanings that are not in the 226 basic vocabulary meanings cover WOLD401-500, a list of 'less basic vocabulary' first introduced in Lehtinen *et al.* (2014).

The lexical reflexes representing each meaning in the dataset have been chosen to be as semantically neutral and context-independent as possible; in other words, they are contemporary general-purpose words or expressions for each meaning. Simple words or expressions have been favoured over complex ones. A language may also have multiple lexical reflexes documented for a single meaning, generally in cases where there are several equally appropriate expressions to denote that meaning.

To ensure consistent quality of the dataset, one individual has been responsible for drawing up the list of representative words for each language, and this list for most of the languages has then been double-checked by one or more language experts or native speakers. The compilers and language-checkers are listed in the `Language_compilers` table of the dataset.

In addition to the recorded gloss, the phonetic transcription in the Uralic Phonetic alphabet and/or International Phonetic Alphabet is included for some of the lexical reflexes.

Each lexical reflex is also associated with multistate characters that reflect what are often referred to as *cognate relationships* but should more accurately be called *root-meaning relationships* (Chang *et al.* 2015). Root-meaning relationships indicate that the neutral words (i.e. words whose use is not bound to a specific context) occupying a given meaning slot originate from the same ancestral root, initially acquired through inheritance. The words are considered to belong to the same root-meaning set even if they represent different derivations, as long as they share the same historical root and occupy the same meaning slot. For instance, reflexes of the meaning 'animal' in several Saami (North, Inari, Skolt) and Finnic languages (Finnish, Karelian, Ingrian, Western Votic, Estonian and Võro) are based on the same ancestral root, the Proto-Uralic stem **elä-* 'to live' (SSA, 102), using different derivational suffixes, and are consequently grouped to the same 'root-meaning' set. In contrast, cognate relationships (in the accurate sense of the word) require the full word, rather than just the root, to share a historical connection through inheritance. Cognates also do not need to occupy the same meaning slot, unlike root-meaning relationships. For example, Komi Zyryan *bi* 'fire', Finnish *päivä* 'day' and North Saami *beaivváš*, *beaivi* 'sun, day' are considered to be cognate words that originate from a shared Proto-Finno-Ugric form, possibly the Proto-Uralic stem **päjwä*, **päjwa* 'warm, heat, fire' (SSA, 457). However, they do not form a common root-meaning trait because they do not occupy the same meaning slot. Likewise, the aforementioned reflexes for 'animal', which count as root-meaning forms, would not be counted as cognates, as they are different derivations of a shared root rather than full words with an ancestral connection.

In addition to inheritance from an ancestral language, linguistic items are also transmitted between languages through borrowing. While borrowed words share a root with their source language, they are not considered to belong to the same root-meaning set as their source, even if the donor language belongs to the same language family, since root-meaning sets represent inherited connections rather than lateral connections. For example, the words in the meaning slot for 'left' in Finnic languages (e.g. Est. *kura* 'left') are considered to be of common origin, while the North Saami *gurut* is a borrowing from Finnic defining a different set (EES 193). Similarly borrowings from other language families are separated into their own root-meaning sets. However, the dataset also includes another multistate character column besides the root-meaning sets to reflect *correlate* relationships; these group together words of the same meaning in different languages that share a root-form through either borrowing or inheritance.

The root-meaning relationships and correlate relationships in the current version of the dataset are based on published etymological references, mostly those published before 2014.

The data currently covers 26 Uralic languages: Finnish, Ingrian, Karelian, Estonian, South Estonian (Võro), Veps, (Western) Votic, (Courland) Livonian, Inari Saami, Kildin Saami, North Saami, Pite Saami, Skolt Saami, South Saami, Ume Saami, Erzya, Meadow Mari, Komi-Zyrian, Komi-Permyak, Udmurt, Hungarian, (Sosva) Mansi, (Vakh-Vasyugan) Khanty, Tundra Nenets, Nganasan and Northern Selkup. The data also includes a partial reconstruction of Proto-Uralic.

The dataset also provides several subsets of the included meanings (meaning lists). These include the lists that serve as the basis for the chosen meanings - that is, Swadesh100 (Swadesh 1955), Swadesh200 (Swadesh 1952), Leipzig-Jakarta (Tadmor 2009) and WOLD401-500 (a list of less basic vocabulary from Lehtinen *et al.* 2014). In addition, the dataset also includes a Swadesh207 list (which combines the meanings of Swadesh200 and Swadesh100), the Fullbasic list (which combines the 226 basic vocabulary meanings from Leipzig-Jakarta, Swadesh100 and Swadesh200 as one list) and Ura100 (a Uralic basic vocabulary list introduced in Syrjänen *et al.* 2013). Notably, the Ura100 list, which is based on an older 17-language version of the Uralic dataset, should be regarded as obsolete, and we do not encourage people to use it. In addition to the aforementioned meaning lists, the dataset also includes Leipzig-Jakarta ranks for the meanings that belong to the Leipzig-Jakarta list and the WOLD401-500 list.

Information from the UraLex basic vocabulary dataset or its predecessor, the BEDLAN cognate database, has featured in Honkola *et al.* (2013), Syrjänen *et al.* (2013), Lehtinen *et al.* (2014), List *et al.* (2017) and Tambets *et al.* (2018).

The dataset is released under Creative Commons Attribution 4.0. The authors welcome all contributions for future versions.

Contents

The raw version of the dataset (found in the `raw` folder of the repository) is organized into seven tables, provided both as a single ODS (OpenDocument spreadsheet) file and also separate TSV (tab-separated values) files. The overall data structure of each table in the basic vocabulary dataset is described below, with separate sections for each table of the dataset. Fields in parentheses represent information that has been duplicated from another table of the dataset using the VLOOKUP function (see the ODS file) to make the contents more human-readable.

Citable bibliographical references from the `Citation_codes` table are also provided as a separate BibTeX file. The dataset folder also includes maintenance scripts for update purposes.

In addition to the raw version the repository also includes a CLDF conversion of the data produced by Robert Forkel and Luke Maurits.

`Data`

This table contains the main data of the basic vocabulary dataset, including each recorded lexeme along with their correlate and root-meaning sets.

1. `(language)`

   Language name.

2. `(definition)`

   Verbose definition of a meaning.

3. `(uralex_mng)`

   Meaning name in UraLex/LexDB database form.

4. `mng_item`

   BEDLAN dataset numerical meaning code.

5. `lgid3`

   BEDLAN dataset numerical language code.

6. `item`

   Lexeme data. Contains a lexeme or [No equivalent] (no suitable equivalent for a meaning exists), [Form not found] (no suitable equivalent was found) or [Not reconstructable] (non-recontructable meanings in Proto-Uralic).

7. `item_UPA`

   Phonetic transcription in Uralic Phonetic Alphabet (included for 11 languages).

8. `item_IPA`

   Phonetic transcription in International Phonetic Alphabet (included for 16 languages).

9. `form_set`

   Correlate set (historical connection based on borrowing or cognacy), marked with positive integers. For [No equivalent] items the field is marked with ‘0’; for [Form not found] and [Not reconstructable] items the field is marked with ‘?’.

10. `cogn_set`

    Cognate set (historical connection based on cognacy), marked with one-letter or two-letter codes. For [No equivalent] items the field is marked with ‘0’; for [Form not found] and [Not reconstructable] items the field is marked with ‘?’.

11. `age_term_pq`

    The earliest likely age of the cognate set (root form).

12. `age_term_aq`

    The latest likely age of the cognate set (root form).

13. `borr_source`

    Borrowing source of lexeme.

14. `borr_qual`

    Likelihood of borrowing (*possible*, *probable* or *clear*).

15. `etym_notes`

    Notes related to etymology of the lexeme.

16. `glossing_notes`

    Notes related to the meaning of the lexeme.

17. `general_notes`

    Other notes related to the lexeme.

18. `ref_abbr`

    Bibliographical references related to lexeme. Multiple references separated by comma and space.

`Meanings`

This table provides information related to separate meanings, including the meaning codes used to refer to separate meanings, and verbose descriptions of each meaning.

1. `mng_item`

   BEDLAN dataset numerical meaning code.

2. `LJ_rank`

   Leipzig-Jakarta rank, included for meanings belonging to either WOLD401-500 or Leipzig-Jakarta and marked “-” for the remaining meanings.

3. `uralex_mng`

   Meaning name in UraLex/LexDB database form.

4. `definition`

   Verbose definition of a meaning.

`Languages`

This table provides language-related information. Notably, it also includes codes and information for Uralic languages that are currently not covered by the main contents of the dataset.

1. `lgid3`

   BEDLAN dataset numerical language code.

2. `language`

   Language name in verbose form.

3. `ASCII_name`

   Language name in simplified form.

4. `ISO-639-3`

   ISO-639-3 code for the language (if available).

5. `Description`

   Additional language description.

6. `Subgroup`

   Traditional subgroup of the language.

`Meaning_lists`

This table specifies the meaning lists and related information. The meanings found on each of the meaning lists is marked with boolean values 0 (absent) and 1 (present).

1. `mng_item`

   BEDLAN dataset numerical meaning code.

2. `(uralex_mng)`

   Meaning name in UraLex/LexDB database form.

3. `(LJ_rank)`

   Leipzig-Jakarta rank, included for meanings belonging to either WOLD401-500 or Leipzig-Jakarta and marked "-" for the remaining meanings.

4. `Ura100`

   Ura100 list (see Syrjänen *et al.* 2013).

5. `Swadesh100`

   Swadesh100 list.

6. `Swadesh207`

   Swadesh200 + Swadesh100 list.

7. `Leipzig-Jakarta`

   Leipzig-Jakarta list. Notably, the UraLex version of the list covers 101 meanings instead of 100. This is because "foot" and "leg" in UraLex are represented as separate meanings, in the style of the Swadesh200 list, and not as a combined meaning "foot/leg", as they are on the official Leipzig-Jakarta list.

8. `WOLD401-500`

   WOLD401-500 list (see Lehtinen *et al.* 2014).

9. `Fullbasic`

   Swadesh200 + Swadesh100 + Leipzig-Jakarta list.

10. `Swadesh200`

    Swadesh200 list.

`Meaning_list_descriptions`

This table provides more verbose descriptions for the meaning lists found in the `Meaning_lists` table.

1. `list`

   Meaning list name.

2. `description`

   Meaning list description.

Citation_codes

1. `ref_abbr`

   Bibliographical reference key.

2. `original_reference`

   Bibliographical reference information related to key. Citable references (publication, URL) are recorded as BibTeX entries, whereas experts are recorded in plain text. The BibTeX entries recorded here are also provided as a separate BibTeX file (`Citations.bib`).

3. `type`

   Reference type (P = publication, U = URL, E = expert).

Language_compilers

This table records the word list collectors and double-checkers of each language.

1. `lgid3`

   BEDLAN dataset numerical language code.

2. `(language)`

   Language name.

3. `collected_by`

   Who has collected the data for a language.

4. `final_wordlist_checked_by`

   Who has checked the wordlist for a language.

References

Chang, W., C. Cathcart, D. Hall & A. Garrett. (2015). “Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis.” *Language* 91(1): 194-244. DOI: 10.1353/lan.2015.0005.

EES = Metsmägi, I., M. Sedrik & S. Soosaar. (2012). “Eesti etümoloogiasõnaraamat.” Tallinn: Eesti Keele Sihtasutus.

Honkola, T., O. Vesakoski, K. Korhonen, J. Lehtinen, K. Syrjänen & Niklas Wahlberg. (2013). “Cultural and climatic changes shape the evolutionary history of the Uralic languages.” *Journal of Evolutionary Biology* 26: 1244-1253. DOI: 10.1111/jeb.12107.

Lehtinen, J., T. Honkola, K. Korhonen, K. Syrjänen, N. Wahlberg, & O. Vesakoski. (2014). "Behind Family Trees: Secondary Connections in Uralic Language Networks." *Language Dynamics and Change* 4: 189-221. DOI: 10.1163/22105832-00402007.

List, J.-M., S. J. Greenhill & R. D. Gray. (2017). "The potential of automatic word comparison for historical linguistics." *PLOS One.* DOI: 10.1371/journal.pone/0170046.

McMahon, A. & R. McMahon (2005). *Language Classification by Numbers.* Oxford: Oxford University Press.

SSA = Suomen sanojen alkuperä I–III (1992-2000). Suomalaisen Kirjallisuuden Seuran Toimituksia 556. Kotimaisten kielten tutkimuskeskuksen julkaisuja 62. Suomalaisen Kirjallisuuden Seura, Helsinki.

Swadesh, M. (1952). "Lexicostatistic dating of prehistoric ethnic contacts." *Proceedings of the American Philological Society* 96, 452-463.

Swadesh, M. (1955). "Towards greater accuracy in lexicostatistic dating." *International Journal of American Linguistics* 21, 121-137.

Syrjänen, K., T. Honkola, K. Korhonen, J. Lehtinen, O. Vesakoski, & N. Wahlberg. (2013). "Shedding more light on language classification using basic vocabularies and phylogenetic methods. A case study of Uralic." *Diachronica* 30(3), 323-352. DOI: 10.1075/dia.30.3.02syr.

Tadmor, U. (2009). "Loanwords in the world's languages: Findings and results." In M. Haspelmath & U. Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 55-75. Berlin: Walter de Gruyter.

Tambets, K., B. Yunusbayev, G. Hudjashov, A.-M. Ilumäe, S. Rootsi, T. Honkola, O. Vesakoski, Q. Atkinson, P. Skoglund, A. Kushniarevich, S. Litvinov, M. Reidla, E. Metspalu, L. Saag, T. Rantanen, M. Karmin, J. Parik, S. I. Zhadanov, M. Gubina, L. D. Damba, M. Bermisheva, T. Reisberg, K. Dibirova, I. Evseeva, M. Nelis, J. Klovins, A. Metspalu, T. Esko, O. Balanovsky, E. Balanovska, E. K. Khusnutdinova, L. P. Osipova, M. Voevoda, R. Villems, T. Kivisild & M. Metspalu. (2018). "Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations." *Genome Biology* 19(139). DOI: 10.1186/s13059-018-1522-1.