PanLex Swadesh Lists

Jonathan Pool, editor

URL: http://dev.panlex.org/db/panlex_swadesh.zip
PanLex documentation: http://dev.panlex.org

Background

The following pages contain 719 Swadesh Lists drawn from the PanLex Swadesh List Corpus.

PanLex is a project of The Long Now Foundation that aims to make every word or multiword expression in every language translatable into every other language. To accomplish this goal, PanLex is consulting thousands of dictionaries, word lists, glossaries, and other sources of lexical translations, and storing all of the translations in a free public database. The PanLex database contains over a billion direct translations, from which billions more can be indirectly inferred.

A Swadesh List is a numerically ordered list of basic concepts. The PanLex Swadesh Corpus contains two variants of the Swadesh List: one with 110 concepts, and one with 207 concepts. The 207-concept list is the one used here. This variant of the Swadesh list (PanLex language variety code art-012) is described in: Luís Morgado da Costa, Francis Bond, and František Kratochvíl, "Linking and Disambiguating Swadesh Lists: Expanding the Open Multilingual Wordnet Using Open Language Resources", *GLOBALEX 2016: Lexicographic Resources for Human Language Technology* (02016), pp. 29–36. Each concept is identified with a number from 1 to 207.

Each of the 719 lists is titled with its Language Variety Code and Language Variety Name. Languages in PanLex are identified with a three-letter ISO 639 language code. Since the language code does not differentiate all forms of a language, the language variety is further distinguished with a numeric identifier. These identifiers are typically used to differentiate dialects and orthographies (mainly those written in different scripts). A PanLex language variety code consists of the language's three-letter ISO 639 language code, a hyphen, and a three-digit numeric identifier. For example, "eng-000" indicates English, and "cmn-000" indicates Mandarin written in Simplified Han characters. The PanLex

Language Variety Name is the variety's autoglossonym (name of the variety in the variety itself), if available, or a commonly-used name in another language variety, if not. For example, "English" is the name of English and "现代标准汉语" is the name of Mandarin written in Simplified Han characters.

There are over 10,000 language varieties in the PanLex database, but only 719 Swadesh Lists here. In order for a Swadesh 207 list to be included in the PanLex Swadesh Corpus, at least 75% of the 207 expressions must have a translation attested in the PanLex Database. This results in 796 lists in the original corpus. From this list we have removed artificial languages (used in PanLex mainly for international standards and other terminology sets) and less prominent varieties of a language, which include (usually) secondary writing systems or (occasionally) minority dialects of that language.

PanLex is compiled from thousands of sources. These sources are of varying quality and are not always consistent with each other. This messiness is inevitably reflected in the Swadesh Lists here. The complete list of PanLex sources is available at https://panlex.org/tech/plrefs.shtml.

Accessing the Swadesh Lists via NLTK

NLTK (Natural Language Toolkit) is a Python module for natural language processing. It contains convenient interfaces to numerous corpora, including the PanLex Swadesh Corpus. If you would like to access the PanLex Swadesh Lists via NLTK, you should install NLTK (http://www.nltk.org/install.html) and download the PanLex Swadesh Corpus (http://www.nltk.org/data.html). To use the corpus from NLTK, follow the instructions at http://www.nltk.org/book/ch02.html#comparative-wordlists, replacing "swadesh" with "swadesh110" or "swadesh207".

Credits

The PanLex Project staff are:

Jonathan Pool, project director
David Kamholz, lexical data specialist
Susan Colowick, research associate
Alex DelPriore, source analyst
Gary Krug, source analyst

Benjamin Yang, source analyst
Julie Anderson, source acquisition specialist

The PanLex Steering Committee is:

Emily Bender, University of Washington
Steven Bird, University of Melbourne
Laura Welcher, The Long Now Foundation

The Swadesh Lists were compiled and edited for the Rosetta Disk by Caroline Glazer, PanLex intern for Summer 02016.