

# PanLex: A Lexical Infrastructure Tool

**David Kamholz, Laura Welcher**

The Long Now Foundation, The Long Now Foundation  
P.O. Box 475668, San Francisco, CA 94147, USA  
kamholz@panlex.org, laura@longnow.org

## Abstract

An important piece of language technology infrastructure for under-resourced languages is lexical infrastructure. Word translations have various uses, and for many languages these are among the best data available. The PanLex Database is a lexical infrastructure tool that can translate any word in any language into any other language. It contains 25 million words in 5,700 languages, representing 1.3 billion translations from 2,500 multilingual sources. This practical tool is available now, provides an immediate benefit to linguistic communities, and helps enable future technology. It is an easy, low-cost way to make translation dictionaries available online and interoperable with other languages.

**Keywords:** translation, dictionary, database

## Résumé

[Balinese] Kepahan sané dahat mautama ring kawéntenan infrastruktur teknologi bahasa (panglimbak teknologi basa) ring basa sané nénten makéh maduwé sumber inggih punika infrastruktur leksikal. Artos sajeroning kruna-kruna makéh pisan kawigunanyané, taler ring makudang-kudang basa, data puniki wantah data sané pinih becik kawéntenannyané. Database PanLex inggih punika piranti infrastruktur leksikal sané prasida kaanggén ngartos sakancan kruna ri sajeroning basa sané kaartos dados basa lianan. Puniki madaging slaé (25) yuta kruna ring limang tali pitungatus (5700) basa, pinaka panyeledihi 1,3 miliar artos saking kalih tali limangatus (2500) sumber basa lianan. Piranti praktis puniki mangkin sampun wénten, prasida mawiguna ring paguyuban basa, miwah ngwantu, nyiagayang teknologi riwekas. Puniki dangan pisan, pamargi sané nénten akéh ngamedalang prabea prasida makarya kamus sané wanggunyané online miwah prasida kaoprasiang ring basa lianan.

## 1. Introduction

Realizing the goal of LT4All requires creating and improving language technology infrastructure for under-resourced languages (98% of all languages). An important piece of this is lexical infrastructure: word translations have a variety of uses, and for many under-resourced languages, lexical data is among the best data available.

The PanLex Database (Kamholz et al. 2014) is a lexical infrastructure tool: it can translate any word in any language into any other language. This practical and useful tool is available now, and can be improved quickly at low cost. It provides an immediate benefit to linguistic communities and helps enable future language technology, such as machine translation.

PanLex (panlex.org) is a project of The Long Now Foundation (longnow.org), a nonprofit that fosters long-term thinking. The PanLex Database has been continuously developed since 2005 as a sister project to the foundation's Rosetta Project which collects parallel documentation on the world's languages for very long-term archiving (see rosettaproject.org).

## 2. The PanLex Database

The PanLex Database is the world's largest lexical translation database, currently containing 25 million words in 5,700 languages. It has been designed from the start to accommodate lexical data from all languages, including dialects and other subvarieties, without privileging any one language. It has also been designed to accommodate a diverse array of multilingual sources. At minimum, a lexical data source need only provide lexemes in one language and corresponding lexical translations or

explanatory definitions in another language. This allows inclusion of basic wordlists and other limited sources. Richer information is included if available, for example part of speech, division into senses, semantic domain, and usage register. The database is available under the CC0 license. Data dumps and a live HTTP API are available (dev.panlex.org).

The PanLex Database incorporates data from more than 2,500 different sources. Sources run the gamut from print to digital, from basic wordlist to large database. The PanLex team particularly emphasizes sources from the least-resourced languages. Each source receives a quality score from 0 to 9.

The PanLex Database can generate both *direct* and *indirect* lexical translations. A translation is direct if one or more sources exist in the database that directly attest the translation of the desired lexical item into the target language. Depending on the language pair and the word being translated, a direct translation may not be available. An indirect translation first translates a word directly into all available languages, producing a set of intermediate translations, and then tries to translate each of these into the target language. Indirect translations are most effective when there are many intermediate translations, as these will effectively converge on the correct target translations. Every available language will be used for intermediate translations; there is no reliance on a single pivot language like English.

Direct and indirect translations are ranked according to translation score, which is assigned based on the number and quality of sources in the database that support the translation. This means that high-quality translations can be

produced from sources of varying quality, and that quality improves over time as new, independent sources are added to the database.

Figure 1 below, taken from PanLex’s online translator (translate.panlex.org), illustrates a direct and indirect translation from Breton *dour* ‘water’ to Cuzco Quechua *yaku*. The double-sided arrow indicates that a direct translation is available between the two words. The other lines show a subset of the available intermediate translations that support an indirect translation; these are in a variety of languages such as French *eau*, Swahili *maji*, and Georgian *წყალი*. Hundreds of intermediate translations may be used when generating indirect translations.

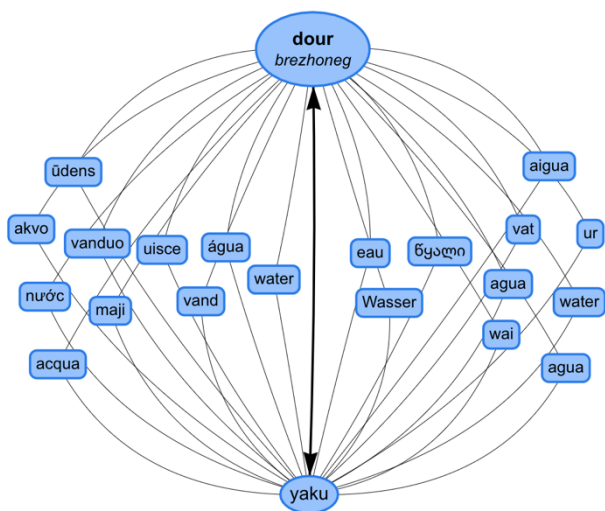


Figure 1: Translation of Breton *dour* ‘water’ to Cuzco Quechua *yaku*, with subset of intermediate translations.

### 3. An Enabling Technology

The PanLex Database currently contains 1.3 billion direct translations and billions more indirect translations. It is useful both on its own and as an enabler of other language technology.

Lexical translations have a variety of practical uses. End-users can use the PanLex Database directly in order to gloss words in unfamiliar languages in contexts such as travel, reading, education, and professional translation. It can help create multilingual glossaries for emergency preparedness. The database is an easy, low-cost way to make translation dictionaries available and interoperable with other languages.

Developers can use PanLex data to create their own apps and interfaces. A mobile keyboard app called Polyglot is currently under development which allows language learners to look up and check word translations interactively as they type, in any language for which PanLex has data. PanLex is especially useful where broad language support is needed.

Finally, the PanLex Database helps enable other efforts to bring language technology to under-served languages. It can support localization work by providing seed

translations for common terms. It can help improve machine translation quality and coverage as it is developed for more under-resourced languages.

### 4. Future Steps

PanLex has successfully linked more than 2,500 multilingual sources and made them interoperable, but much work remains to improve lexical infrastructure for under-resourced languages. Many languages need improved coverage, and PanLex already has more than 4,000 sources in its backlog.

The PanLex team is currently in discussions with Wikimedia Deutschland to make PanLex data available through Wikidata (wikidata.org). This will make the PanLex Database available through a well-known international platform and will allow crowd-sourced improvements. We are excited to continue developing this infrastructure to enable linguistic diversity and multilingualism worldwide.

### 5. Bibliographical References

Kamholz, D., Pool, J., and Colowick, S. (2014) PanLex: Building a Resource for Panlingual Lexical Translation. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 3145-3150, Reykjavik, Iceland. European Language Resources Association (ELRA).