PanLex: Building a Resource for Panlingual Lexical Translation

David Kamholz, Jonathan Pool, Susan M. Colowick
The Long Now Foundation
San Francisco, California, U.S.A.
kamholz@panlex.org, pool@panlex.org, smc@panlex.org

Abstract

PanLex, a project of The Long Now Foundation, aims to enable the translation of lexemes among all human languages in the world. By focusing on lexemic translations, rather than grammatical or corpus data, it achieves broader lexical and language coverage than related projects. The PanLex database currently documents 20 million lexemes in about 9,000 language varieties, with 1.1 billion pairwise translations. The project primarily engages in content procurement, while encouraging outside use of its data for research and development. Its data acquisition strategy emphasizes broad, high-quality lexical and language coverage. The project plans to add data derived from 4,000 new sources to the database by the end of 2016. The dataset is publicly accessible via an HTTP API and monthly snapshots in CSV, JSON, and XML formats. Several online applications have been developed that query PanLex data. More broadly, the project aims to make a contribution to the preservation of global linguistic diversity.

Keywords: lexicon, language, translation, dictionary, database

1. Introduction

PanLex[1] aims to enable panlingual lexical translation—the translation of lexemes among all human languages in the world.

Initiated in 2005 as a research project at the University of Washington Turing Center[2] and since 2012 sponsored by The Long Now Foundation[3], PanLex has developed a lexical database covering literary and colloquial varieties of living natural languages, as well as extinct, reconstructed, artificial, and controlled languages. Most of the data for the project are based on bilingual, multilingual, and monolingual publications, including dictionaries, word lists, wiktionaries, wordnets, and thesauri.[4]

PanLex proceeds from the assumption that one can achieve the greatest return on investment in panlingual translation by investing in purely lexical (more precisely lexemic) translation. There are several advantages to this strategy. (1) Lexemic translation data, unlike grammatical and corpus data, are widely available for thousands of languages. (2) The problems of procuring lexemic translation data, imposing a uniform structure on them, and linking them to outside data are relatively tractable. (3) Lexemes play a critical role in expressing propositional content. (4) In some contexts, purely lexemic translation is sufficient (e.g., online profile headings, catalogs, tags, search, and user interfaces).

The PanLex project focuses on procuring content, maintaining a high-quality dataset, and making its data available to researchers and developers. It leaves to others the task of developing sophisticated user-facing applications. It also does not attempt to make computationally tractable the complex problem of translation inference, i.e., inferring unattested translations from the existing dataset.

As of March 2014, the database contains about 20 million lexemes (or *expressions*) in about 9,000 language varieties, with 1.1 billion pairwise translations among lexemes.

This paper describes the essential features of PanLex. An overview of related work (§2) is followed by discussion of the database design (§3), compliance with existing standards (§4), the project's strategy in acquiring new data (§5), the current level of PanLex's language coverage (§6), public access to the dataset (§7), research on the dataset (§8), software applications (§9), and the project's next phase (§10).

2. Related work

The developers of standards and resources have increasingly recognized and accommodated the multiplicity of languages in the world.

The ISO 639 standard (SIL International, 2014) began in 1988 with identifiers for 185 languages. As of 2014, its codes identify 8,330 individual languages.

The Unicode standard (The Unicode Consortium, 2013), when first published in 1991, standardized the encoding of 19 scripts. With version 6.3.0, published in 2013, the standard has grown to include 98 scripts.

The Rosetta Project[5] was founded in 1996. A project of The Long Now Foundation with which PanLex closely collaborates, it aims to build a publicly accessible, long-term digital archive of all human languages.

The Wikimedia Foundation's Wiktionary project[6] was launched in 2002. English was its sole source language until 2004. By early 2014, 171 source languages were represented.[7]

Automatic translation services have also become available in a growing number of languages. In 2001, Google's translation service offered translation among 5 languages.[8]

[1] http://panlex.org

[2] http://turing.cs.washington.edu

[3] http://longnow.org

[4] Language informants have also contributed a small amount of data.

[5] http://rosettaproject.org

[6] http://www.wiktionary.org

[7] http://meta.wikimedia.org/wiki/Wiktionary#List_of_Wiktionaries

[8] https://web.archive.org/web/20011212060213/http://www.google.com/language_tools

Now called Google Translate[9], it performs "most of the translation on the planet" (Och, 2012). By early 2014, its language count had increased to 80.

The Automated Similarity Judgment Program (Wichmann et al., 2014), which began in 2008, has published a database of expressions for 40 concepts in about 7,000 languages.

The World Atlas of Language Structures Online (Dryer and Haspelmath, 2013) is a database covering phonological, morphological, syntactic, and lexical properties of 2,679 languages based on about 7,000 sources.

Rosenfelder (2014) has compiled lexemes for the numbers from 1 to 10 in about 5,000 languages.

The UWN project (de Melo, 2014) has produced a taxonomy of meanings labeled with about 1.5 million lexemes in about 200 languages.

As the above examples illustrate, projects that are designed to be panlingual tend to have specific and limited objectives, such as the identification of all languages, their character sets, or expressions for a finite set of concepts.

PanLex, with its objective of documenting only the lemmatic forms of lexemes (§3), is no exception. However, it is distinguished from other projects in maximizing both lexical and language coverage. All projects known to us with comparable lexical scope cover fewer than 800 languages, while those covering thousands of languages have substantially narrower lexical scope.

PanLex and the related projects listed above, though distinct, can and do benefit from one other. PanLex editors have discovered millions of lexical translations in related projects' publications, and PanLex data, in turn, are available for use by related projects.

3. Design

The PanLex database is designed to capture any written lexical translation from any source.[10] Thus, it mandates only minimal information for any given lexical translation. At least one *expression* (i.e., single- or multi-word lexeme) must be translated, generally into another expression. If no expression is available, it may be translated into an explanatory *definition*. The only required attributes of an expression are its language variety and a text string constituting its *lemma* (citation form). The lemma is ideally in a standard orthography, but may instead be a phonetic transcription or other string, so as to accommodate data for languages which lack a standard orthography. A newly documented expression is identified with an existing expression if the language variety and text string match; otherwise, a new expression is created.

Language varieties are identified with their three-character ISO 639 *language code* (e.g., *eng* for English), plus a three-digit *variety code*, starting at *000*. Dialects are treated as distinct varieties, as are forms of a language written in distinct scripts. "Controlled languages" that are not varieties of any other language, such as the numbered items in Swadesh lists (12 = 'two', etc.) or the abbreviations in the periodic table of the elements (H = hydrogen, etc.), are given the language code *art*, the generic code for an artificial language. New language varieties are added whenever necessary to document new data.

Editors are individuals who add and modify PanLex data. They attribute each translation to a *source*, which may be a document or (less commonly) an informant.

Meanings are arbitrary numbers assigned to each set of intertranslated expressions in a source. They are automatically generated whenever a translation is added to the database. The pairing of a meaning with an expression is called a *denotation*. Thus, a source's claim of semantic equivalence among a set of expressions is represented as the assignment of a shared meaning to each of the expressions. When sources differ on how an expression should be translated, the alternative translations coexist in the database, with each translation traceable to its source(s).

Because meanings are automatically generated as translations arrive, editors do not need to discover existing meanings that are semantically identical. Translation inference must address the problem of *meaning consolidation*, i.e., identifying distinct sources' meanings that should be considered alike.

PanLex has a symmetric design: if two or more expressions share a meaning, the expressions are not classified as a source or headword and one or more targets. Instead, each expression is treated as a translation of each of the others; the translation graph is undirected. This permits inferred translations from any lexeme into any language variety.

Some additional data that are common in lexical resources are incorporated into the design as optional elements. These include three properties that may be attributed to meanings: (1) *meaning identifiers*, strings used by sources to uniquely identify their meanings; (2) *domains*, expressions that classify meanings but do not express them, such as *chemistry* or *music*; (3) *definitions*, explanatory strings (not expressions) identified as being in particular language varieties. Two properties may characterize denotations: (1) *word classes*, parts of speech such as *noun* or *verb*, drawn from a closed set; (2) *metadata*, arbitrary pairs of strings representing attributes and values.

The most important entities and relationships in the database are represented in Figure 1 on the following page.

4. Standards

The PanLex project seeks to achieve compliance with prevailing international standards, so long as these do not impair its mission. The database design (§3) incorporates several standards, listed below.

Textual data are constrained to comply with the Unicode standard. All text strings are subjected to a single normalization form (NFC) and a single encoding form (UTF-8). For example, the small Latin letter *a* with ogonek (*ą*) is always encoded with its Unicode hexadecimal codepoint 0105 and stored in text strings with the bytes `C4A5` as required by UTF-8, regardless of how it may be encoded in a source. It is possible that future conflicts may arise between this constraint and the documentation of some low-density languages and sign languages.

[9] `http://translate.google.com`

[10] For a more detailed and technical overview of the database, see `http://dev.panlex.org/wp-content/uploads/2014/03/panlex-db-design.pdf`.

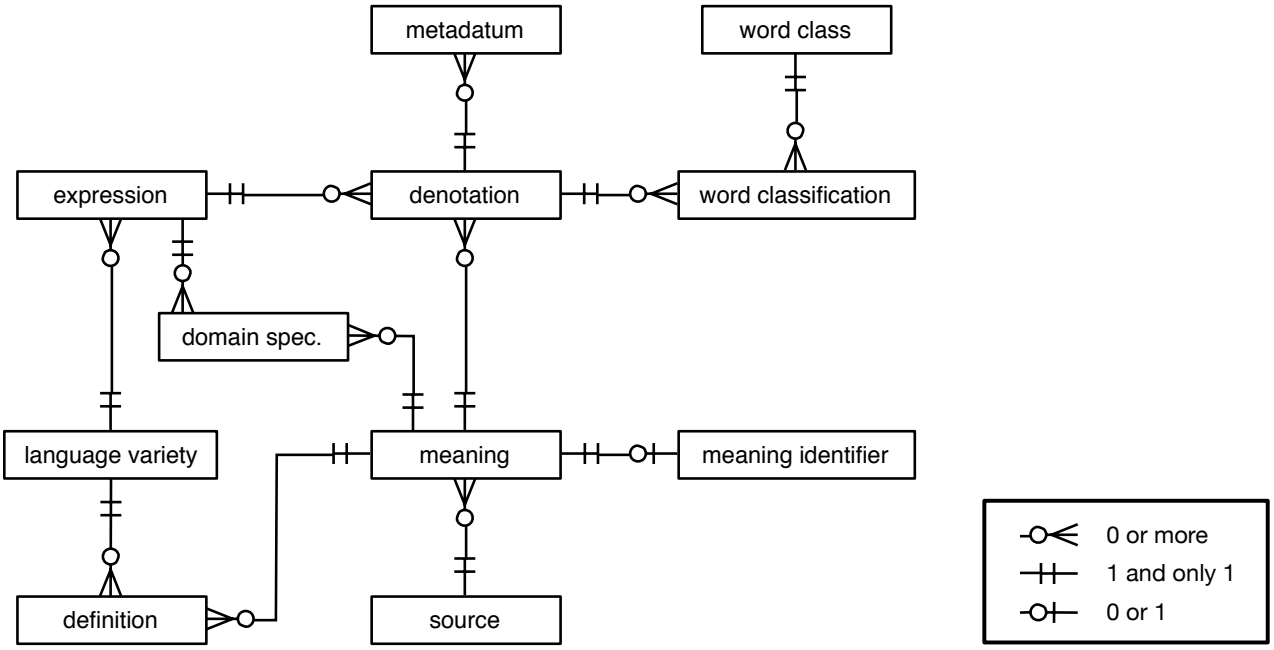

Figure 1: Entity-relationship diagram of the PanLex database, using Crow's Foot notation.

Language codes are limited to the three-character ("alpha-3") codes of ISO 639-2, ISO 639-3, and ISO 639-5. In the future, PanLex may adopt other standards of language and language-variety identification, such as those developed by the LINGUIST List[11] and the Internet Engineering Task Force[12].

The closed set of word classes used in PanLex is an extension of the set found in the OLIF standard[13].

PanLex documents each original source's ISBN and URL, when they exist.

The project also cooperates with efforts to make the entire translation graph accessible under existing linked data standards. Researchers at the University of Leipzig have developed an RDF interface[14] for PanLex, have made the data conform to the lemon and GOLD data models, and have linked PanLex data to Lexvo and DBpedia (Westphal et al., forthcoming).

5. Acquisition

The PanLex content acquisition strategy aims at broad, high-quality lexical and language coverage at moderate cost. We prefer sources meeting the following criteria:

(1) Availability. Before recruiting language informants, we make use of available lexicographic publications and manuscripts. About 2,700 such sources are currently in the pipeline, ready to be consulted.

(2) Tractability. We prefer sources amenable to *analysis*, i.e., semi-automated computational extraction of usable data. Such sources are generally born digital and exhibit Unicode compliance or encoding convertibility, as well as clear and consistent structures. They disaggregate and mark all objects and properties, rather than (for example) combining an entry's lemma, part of speech, and pragmatic force into a single text item.

(3) Comprehensiveness. Analysis typically has a fixed cost, so we prefer sources that document large numbers of lexical translations.

(4) Coverage of *low-density languages*, i.e., those with meager corpora and lexical documentation.

(5) Rare language connections. In order to support robust translation inference, we are building an any-to-any rather than hub-and-spoke graph, where no single language acts as an indispensable pivot. For example, a Hungarian–Yoruba dictionary would be valuable: Hungarian and Yoruba are individually well represented in PanLex, but there are few direct translations between them.

(6) High quality. Works of expert lexicography tend to be more reliable than sources open to public editing or generated algorithmically from corpora.[15] PanLex editors estimate the quality of each consulted source on a 0-to-9 scale. Users can weight competing translations by their sources' qualities.

These criteria can conflict, so editors must exercise judgment. For example, sources induced automatically from corpora tend to be tractable, but low-quality. Lexicons of low-density languages tend to be less tractable because of non-digital formats or legacy character encodings, and their

[11] `http://linguistlist.org/forms/langs/find-a-language-or-family.cfm`

[12] `http://tools.ietf.org/html/bcp47`

[13] `http://www.olif.net/formalization/values/olifValuesFixJuly2001.htm`

[14] `http://ld.panlex.org/rdf.html`

[15] Sharoff et al. (2013) summarize efforts to induce translation lexicons from corpora. Vulić and Moens (2012) discuss barriers to achieving quality in such lexicons.

lexical coverage is often not comprehensive.

The project has collected and developed tools to make source analysis more efficient.[16] These include programs that extract data from some of the more common formats of lexical resources, such as DICT and MDF; libraries to parse HTML, XML, and PDF files; and routines to check submitted data for validity. Optical character recognition (OCR) programs have also been used to convert page images to manipulable text, but with limited success.[17]

Source analysis typically involves *tabularization* followed by *serialization*. In tabularization, we transform source data into well-defined tables with one entry per row. In serialization, we normalize these data (e.g., for punctuation, letter case, and word class specification) and convert the tables into a format that can be validated and ingested into the database.

Because of the complexity of the process, most editorial work has so far been performed by about 20 collaborators. Outside contributions have mainly taken the form of raw sources, not analyzed data.

Pilot studies are conducted as needed to evaluate competing acquisition models that may offer improved efficiency or accuracy. One study (Baldwin et al., 2010) explored an approach to automating the structural analysis that a human editor performs on a PanLex source. In 2013, twelve student interns were trained in about a week to analyze sources with the project's software tools and spent six more weeks applying this training.

During 2014, the project is experimenting with the outsourced transcription of page images from printed sources. We have hired contractors to transcribe material and to train an OCR program to recognize the layout and characters of a particular source. Human transcription tasks range from simple text entry to the production of data compliant with PanLex requirements. Another approach being tested is for the contractor to find only expressions with particular meanings in a source (namely, those found in an empirical concepticon, described in §10) instead of documenting all of the source's translations. This experimentation will guide future acquisition efforts (§10).

6. Language coverage

Despite our effort to prioritize low-density languages (§5), there are vast differences in coverage among language varieties in the PanLex database. As Table 1 shows, by three different measures of density, the representation of language varieties is far from uniform.

A language variety's *expression count* is the total number of its expressions documented in the database. Its *meaning count* is the total number of PanLex meanings assigned to one or more of its expressions. To obtain its *translation count*, one computes the number of translations and synonyms that each of its expressions has and sums the computed numbers.[18] Thus, for example, there are 8,703 language varieties with at least 2 expressions, but only 369 with at least 2,000 expressions.

	Language varieties with ≥ the minimum:		
Minimum	Expressions	Meanings	Translations
2	8,703	8,707	8,617
20	6,854	8,162	8,617
200	2,364	2,811	8,557
2,000	369	434	8,479
20,000	87	108	8,285
200,000	23	37	5,568
2,000,000	1	4	30

Table 1: Number of language varieties with the specified minimum counts of expressions, meanings, and translations.

The language varieties that are richest in expressions, meanings, and translations largely coincide. The top 25 of each set, as of March 2014, are shown in Table 2. It contains 33 distinct language varieties, of which 18 occur in all three groups.

Rank	Expressions	Meanings	Translations
1	English	English	English
2	S. Mandarin	Russian	French
3	Russian	S. Mandarin	German
4	French	French	Russian
5	German	German	S. Mandarin
6	P. Mandarin	Italian	Spanish
7	Spanish	P. Mandarin	Italian
8	Italian	Spanish	Portuguese
9	T. Mandarin	Czech	Dutch
10	Japanese	Esperanto	Polish
11	L. Uyghur	Dutch	Czech
12	A. Uyghur	T. Mandarin	Japanese
13	Czech	Turkish	Swedish
14	Dutch	Japanese	Finnish
15	Yoruba	Croatian	Hungarian
16	Portuguese	Hungarian	Esperanto
17	Arabic	Finnish	Arabic
18	Hungarian	Estonian	Turkish
19	Polish	L. Uyghur	Slovak
20	Esperanto	Portuguese	B. Norwegian
21	Vietnamese	Arabic	P. Mandarin
22	Turkish	Vietnamese	Danish
23	Finnish	A. Uyghur	Catalan
24	Hindi	Polish	Thai
25	Thai	Swedish	Romanian

Abbreviations: A. = Arabic script, B. = Bokmål, L. = Latin script, P. = pinyin, S. = Simplified, T. = Traditional.

Table 2: Highest-density language varieties, measured in counts of expressions, meanings, and translations.

[16] The software tools and workflow used in source analysis are documented in detail at `http://dev.panlex.org`.

[17] Effective use of OCR is difficult for many PanLex sources because of pervasive mixtures of languages and scripts, structural complexity, and the need to determine source-by-source whether text flows within or across columns (Kanungo and Mao, 2003; Karagol-Ayan, 2007; Mabee, 2012).

[18] The translation counts in Table 1 are estimates based on a 5% sample of expressions.

7. Access

Access to the PanLex dataset is available in live and snapshot form. Live access is provided via an API[19] that accepts JSON queries and delivers JSON results over an HTTP endpoint. The API is intended for small-scale, experimental applications. Snapshots[20] are generated monthly in CSV, JSON, and XML formats. Other snapshot formats (e.g., PostgreSQL dumps) may be requested.

The Long Now Foundation, the PanLex project sponsor, offers snapshots of the data in accordance with the terms of Creative Commons CC0 1.0 Universal.[21] Because the project consults thousands of sources and engages in "mass digitization" (Borghi and Karapapa, 2013), these sources' license terms and ownership claims may ultimately limit the project's uses of sources, as well as some uses of PanLex snapshots by others. Each source's database record includes a license category and a claim or grant of permission quoted from the source's documentation, if present.

8. Research

PanLex began as an effort to develop and test methods of lexical translation inference. Given a set of resources, each providing translations among a set of languages, could novel translations, not documented by any of the original resources, be inferred from the aggregate of their translations?

Researchers at the University of Washington Turing Center built TransGraph, the initial version of the PanLex database, and enlarged it to 60 million translations among 10 million lexemes, based on about 600 sources. They demonstrated that high-precision translations not present in any of the original resources could be inferred from the translation graph (Mausam et al., 2010). Further research has shown that such attested and inferred lexical translations are useful for web search (Etzioni et al., 2007) and translingual interpersonal communication (Everitt et al., 2010).

9. Applications

The PanLex project is primarily focused on the creation of its dataset, leaving the development of applications to others. A complete, user-friendly query interface to PanLex is not yet publicly available.[22] However, the project staff has deployed a few demonstration applications that use the public API (§7).[23] TeraDict, InterVorto, and TümSöz are applications that translate lexemes from English, Esperanto, and Turkish, respectively. The PanLex Tattoo Generator provides translations of words popularly used in tattoos.[24] PanLinx creates a hyperlink to each expression in the database, so that search engines can index all expressions and translations in PanLex.

Global Glossary[25] offers user-friendly access to a subset of translations in the PanLex database and additional translations automatically inferred from them.

Researchers at the University of Washington Turing Center developed an early demonstration application called PanImages, based on the original version of PanLex. PanImages translated search queries to facilitate cross-language image retrieval (Etzioni et al., 2007; Colowick, 2008; Colowick and Pool, 2010). Users submitting queries in low-density languages could thereby retrieve many more images than otherwise. To obtain culturally specific images, users could submit queries and select the languages they would be translated into (e.g., one could enter "breakfast" and select Turkmen).

10. Future work

During 2014–2016, the PanLex project plans to consult its entire backlog of sources. This will include an anticipated 1,300 sources to be acquired during this time, in addition to the 2,700 sources currently in the pipeline, for a total of about 4,000 sources. Of these 4,000 sources, about 1,000 will consist of page images rather than encoded text. Processing this backlog will demand increased productivity in source analysis, and may require additional contributors and volunteers.[26] We anticipate that our 2014 pilot studies (§5) will indicate how best to achieve this goal.

One potential application of PanLex is the discovery of concepts that are commonly expressed and translated, such as those found in Swadesh lists, ontologies, wordnets, and thesauri. A recently coined generic name for such a concept inventory is *concepticon* (Poornima and Good, 2010). As an alternative to expert curation of concepticons, commonly translated concepts can potentially be derived empirically from PanLex data. Project researchers are beginning to investigate the viability of empirical concepticons based on PanLex.

11. Conclusion

The Long Now Foundation, PanLex's sponsor, encourages long-term thinking, viewing "now" not as the last and next few minutes or days, but rather as humanity's last and next 10,000 years. It endorses the analogous "big here" with respect to places and people (Eno, 1995). The foundation designs and builds ambitious artifacts to demonstrate the feasibility of these concepts. PanLex is one of these artifacts.

PanLex encourages one to contemplate a world in which speaking a minority language is not a liability. By including all human languages, the project recognizes the unique contribution of each language. It serves as a reminder that the existence of thousands of languages has been the norm throughout human history.

PanLex will not single-handedly halt the current global trend towards language endangerment and death; only concerted efforts by individuals, societies, and governments

[19] `http://dev.panlex.org/api/`

[20] `http://dev.panlex.org/db/`

[21] `http://creativecommons.org/publicdomain/zero/1.0/legalcode`

[22] PanLex developers use a custom web interface for most of their retrieval and modification operations as they add and correct data.

[23] `http://panlex.org/try/`

[24] The PanLex Tattoo Generator debuted in San Francisco at the Exploratorium Market Days mini-festival in October 2013, where dozens of visitors got temporary tattoos in (sometimes exotic) foreign languages.

[25] `http://globalglossary.org/en/`

[26] Interested volunteers should fill out the form at `http://panlex.org/help/give-talent.shtml`.

can do that. Nonetheless, by enabling lexical interoperability, PanLex can make basic communication between speakers of any two languages feasible. We therefore believe that PanLex can make a contribution to the emerging effort to safeguard global linguistic diversity.

12. Acknowledgements

PanLex is supported by a fund at The Long Now Foundation, created with a gift from Utilika Foundation.

The authors thank Steven Bird, Jeremy Pool, Megan Williams, and three anonymous reviewers for valuable comments on drafts of this paper.

13. References

Baldwin, T., Pool, J., and Colowick, S. M. (2010). PanLex and LEXTRACT: Translating all words of all languages of the world. In *Coling 2010: Demonstration Volume*, pages 37–40. http://turing.cs.washington.edu/papers/BaldwinEtAl-Lextract.pdf.

Borghi, M. and Karapapa, S., editors. (2013). *Copyright and Mass Digitization*. Oxford University Press.

Colowick, S. M. and Pool, J. (2010). The functionality of PanImages. Technical report, The Long Now Foundation, San Francisco, CA. http://dev.panlex.org/wp-content/uploads/2014/03/panim-doc.pdf.

Colowick, S. M. (2008). Multilingual search with PanImages. *Multilingual*, 19(2). http://turing.cs.washington.edu/PanImMultilingual.pdf.

de Melo, G. (2014). UWN/MENTA: Towards a universal multilingual wordnet, March. https://www.mpi-inf.mpg.de/yago-naga/uwn/.

Dryer, M. S. and Haspelmath, M., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/.

Eno, B. (1995). The Big Here and the Long Now, January. http://longnow.org/essays/big-here-and-long-now/.

Etzioni, O., Reiter, K., Soderland, S., and Sammer, M. (2007). Lexical translation with application to image search on the web. In *Proceedings of Machine Translation Summit XI*. http://turing.cs.washington.edu/papers/EtzioniMTSummit07.pdf.

Everitt, K., Lim, C., Etzioni, O., Pool, J., Colowick, S., and Soderland, S. (2010). Evaluating lemmatic communication. *trans-kom: Journal of Translation and Technical Communication Research*, 3(1):70–84. http://vg05.met.vgwort.de/na/184af264d4b84f3d9f3a735ebaf5aa36?l=http://www.trans-kom.eu/bd03nr01/trans-kom_03_01_03_Everitt_et_al_Lemmatic_Communication.20100531.pdf.

Kanungo, T. and Mao, S. (2003). Stochastic language models for style-directed layout analysis of document images. *IEEE Transactions on Image Processing*, 12(5):583–596. http://lhncbc.wip.nlm.nih.gov/files/archive/pub2003017.pdf.

Karagol-Ayan, B. (2007). Resource generation from structured documents for low-density languages. Dissertation, University of Maryland, College Park, MD. http://hdl.handle.net/1903/7580.

Mabee, C. (2012). Report on internship project: Optical character recognition in multilingual text. Technical report, The Long Now Foundation, San Francisco, CA. http://dev.panlex.org/wp-content/uploads/2014/03/ocr-intern-report.pdf.

Mausam, Soderland, S., Etzioni, O., Weld, D. S., Reiter, K., Skinner, M., and Bilmes, J. (2010). Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, 174:619–637. http://turing.cs.washington.edu/papers/aij-2010-panlingual.pdf.

Och, F. (2012). Breaking down the language barrier—six years in. http://googleblog.blogspot.com/2012/04/breaking-down-language-barriersix-years.html.

Poornima, S. and Good, J. (2010). Modeling and encoding traditional wordlists for machine applications. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, ACL 2010*, pages 1–9. http://aclweb.org/anthology//W/W10/W10-2101.pdf.

Rosenfelder, M. (2014). Numbers from 1 to 10 in over 5000 languages, March. http://www.zompist.com/numbers.shtml.

Sharoff, S., Rapp, R., and Zweigenbaum, P. (2013). Overviewing important aspects of the last twenty years of research in comparable corpora. In Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P., editors, *Building and Using Comparable Corpora*, pages 1–17. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-20128-8_1.

SIL International. (2014). ISO 639-3, March. http://www-01.sil.org/iso639-3/default.asp.

The Unicode Consortium. (2013). The Unicode Standard. http://www.unicode.org/standard/standard.html.

Vulić, I. and Moens, M.-F. (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, pages 449–459. http://dl.acm.org/citation.cfm?id=2380872.

Westphal, P., Stadler, C., and Pool, J. (forthcoming). Countering language attrition with PanLex and the Web of Data. *Semantic Web Journal*. http://www.semantic-web-journal.net/content/countering-language-attrition-panlex-and-web-data-3.

Wichmann, S., Müller, A., Velupillai, V., Wett, A., Brown, C. H., Molochieva, Z., and Bishoffberger, J. (2014). The Automated Similarity Judgment Program, March. http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm.