Semi-automatically generated multi-language glossaries

Ilan Kernerman
K Dictionaries
8 Nahum Hanavi Street
Tel Aviv 63503 Israel
ilan@kdictionaries.com

Abstract

This paper describes the human-assisted automatic generation of multilingual glossaries, derived from the data of an existing English multilingual dictionary. The project began in 2014 and so far it covers 18 languages with 40+ language translations each, linked to each other through the original English dictionary core. Each language data is reverse-engineered from the English multilingual source, first producing a raw L2-English index that consists of the translations-turned-into-headwords and the English entries as their translation equivalents. This raw index is edited meticulously and then the translations of other languages from the English multilingual dictionary are attached to each sense of the entry, thus producing multilingual glossaries.

1 Introduction

The background of K Dictionaries is in semi-bilingual English learners' dictionaries (alias *Password*; Reif, 1987; Kernerman, L., 1994; Nakamoto, 1994). In 2000 it started to combine different language versions of *Password* into a first-of-its-kind English Multilingual Dictionary (EMD; Herpiö, 2001), and since then it continued to add more languages, revise and update the content, uniformize the data format and upgrade its structure. Today the English core features 30,000 entries with 40,000 senses, joined by a total 1.7 million translation equivalents in 46 languages covering each sense of every entry. This English multi-language network has evolved through manual input by lexicographers, translators and editors over nearly 30 years, and it is maintained in XML format.

In 2014, a new index editing tool (KIET) was introduced for editing L2-word-to-English-sense indices based on the original English-L2 pairs. It integrates a raw L2-English index produced by machine reversal of the English-L2 components – turning the L2 translations into candidate headwords and the former English headwords into their translations. Generating the raw index includes automatic data adjustments as well as importing the part of speech from the English entry to the new L2 headword. KIET serves to edit the L2 headwords, their parts of speech, and their English counterparts contained in specific senses that are re-arranged in a determined order. Since each English equivalent corresponds to the original entry in *Password*, with its translations in so many more languages, any or all of those language translations can be added automatically and serve to expand the English-L2 index entries multilingually. And since all the L2 multilingual datasets revolve around the same English multilingual pivot, each L2 can now operate from two ends – as source language and as target language. Such automatic leverage of the data is enabled by fairly modest human intervention that is focused on editing each L2-English pair. First tests are carried out to evaluate the L2-Ln relations, and converting the data to RDF will enable further interoperability with Linked (Open) Data.

2 The process

The multilingual glossary development process consists of three main steps, which are overviewed below:

1. Automatic extraction of bilingual data from the EMD and its reconstruction as a raw L2-English index;
2. Manual editing of the L2-English index;
3. Automatic insertion of the other language translations from the EMD alongside the English equivalents in the bilingual index.

2.1 Data extraction

The XML data of the EMD is parsed and basic tables are created. The program searches all the translation containers and compounds, and joins each one to its corresponding sense(s). The Sense set includes the following components:

- Translations for all the languages
- English definition (of the specific sense)
- English examples of usage (if appropriate)
- English headword and part of speech

The outcome of this parsing is illustrated in Figure 1.

Figure 1. Parsing the XML data and preparing translations in different languages.

The main characteristics of the Sense set are the Definition and the associated L2 Translation. Each sense has an identifier, which will serve later to generate the multilingual glossary. The software also generates Translation tables for all the languages, which will eventually serve the multilingualization process.

At this preliminary stage, the program can generate a raw L2-English index. First, it creates a temporary L2 index by parsing the Translations from the EMD and building a table that includes the following components:

- L2 Translation
- Part of Speech
- English Headword
- Definition (English)
- Example of usage (English; if appropriate)

As a result, the L2 Translation (from the EMD) becomes an L2 Headword. The program combines all the Senses in the EMD that were associated with it as an English Translation and lists them in alphabetical order (according to the original English Headword and Sense number). Subsequently, the L2 Headword is composed of the follow elements:

- Sense set 1
  - English Headword 1
  - Part of speech 1
  - Definition 1
  - Example of usage 1
- Sense set 2
  - English Headword 2
  - Part of speech 2
  - Definition 2
  - Example of usage 2
- Etc.

The main tables used in the index generation process are the following:

- English Headword table
- Senses table
- Translation table
- L2 Index table (i.e. L2 Headword table, generated from the English Headword, Senses and Translation tables)
- L2 Senses table (used for Tree and HTML preview, with the English Headwords, Definitions and Examples tables)

The automatically-generated raw L2-English index is now ready for editing.

2.2 Index editing

The index undergoes thorough manual editing, using the custom-designed KIET software tool, shown in Figure 2.

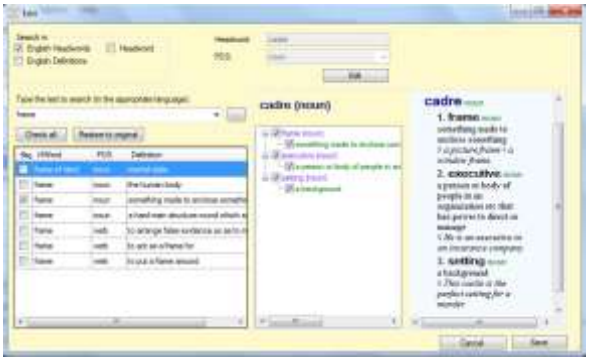

Figure 2. A screenshot of KIET.

The editor reviews the L2 translations-turned-into-headwords to decide which to keep intact, change into legitimate headwords, or remove if irrelevant, and revises the automatically allocated parts of speech. As for the English translation equivalents, the editor removes inappropriate ones and associates others, and re-arranges the default alphabetical order of the senses according to frequency and importance. A shortcoming of this framework is that the English equivalents are limited to those existing in the initial EMD.

A detailed account of the editorial process (for Russian-English) is available in Egorova (2015).

2.3 Translation import

Once the L2-English index is fully edited, the translation equivalents of the other languages in

the EMD are juxtaposed to their original English senses, which from here on function as bridges that automatically match each L2 headword to all other Ln counterparts. The basic table for this multilingual generation is the L2 Index table. The program searches through it and selects the appropriate set, as follows:

- L2 Headword and part of speech;
- All senses associated with the L2 Headword along with their sense identifier;
- Translations of all the languages selected by the sense identifier.

Samples of the outcomes appear in Figures 3 and 4. Figure 3 displays an edited entry from the German-English index, and Figure 4 displays its two first senses with the automatically associated translations from the EMD for 44 languages.

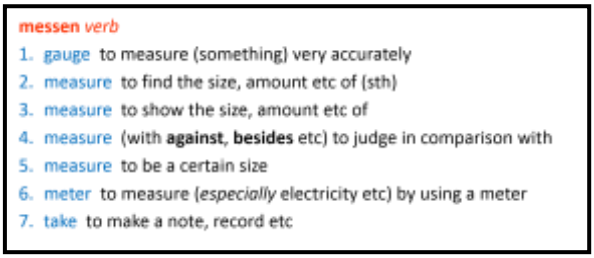

Figure 3. The entry *messen* in the German-English index.

Figure 4. The first two senses of *messen* in the automatically generated Geman multilingual glossary.

Unfortunately, the indirect juxtaposition of different languages through the English pivot is bound to include inaccuracies. Nevertheless, the results have merit for basic translation purposes and can serve as a base for improved matching, useful in particular for less-common language pairs and under-resourced languages. Assessing the rates of accuracy in the L2–Ln automatic matching is starting to be investigated.

3 Conclusion

The multilingualization process described in this paper benefits from relying on the well-formatted comprehensive EMD resource, as well as from chirurgical human editorial refinement embedded within the automatic extraction and generation of the internally linked lexical data.

The main drawback of this process concerns the uneasy balance created by the indirect multi-language associations, further enhanced by the restrictiion of associating L2 headwords in the editorial process only to senses already existing in the EMD. The effects on the precision and recall of the results are still to be explored.

The next steps include further interlinking of the ensuing L2 multilingual glossaries internally among themselves and externally with other Linked (Open) Data resources (based on RDF).

Acknowledgement

This paper features an extract of Kernerman, I. (2015), revised for the BabelNet Workshop in Luxembourg, 3 March 2016.

References

Egorova, Kseniya. 2015. Editing an automatically-generated index with K Index Editing Tool. In *Electronic Lexicography in the 21st Century: Linking lexical data in the digital age. Proceedings of eLex 2015, Herstmonceux Castle, 11–13 August 2015*, edited by I. Kosem, Miloš Jakubíček, Jelena Kallas, Simon Krek. Herstmonceux Castle, UK. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. Available at: https://elex.link/elex2015/.

Herpiö, Mika. 2001. *GlobalDix*: A unique multilingual dictionary for the worldwide market. *Kernerman Dictionary News* 9:12

Kernerman, Ilan. 2015. A multilingual trilogy: Developing three multi-language lexical datasets. In *Electronic Lexicography in the 21st Century: Linking lexical data in the digital age. Proceedings of eLex 2015, Herstmonceux Castle, 11–13 August 2015*, edited by I. Kosem, Miloš Jakubíček, Jelena Kallas, Simon Krek. Herstmonceux Castle, UK.

Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. Available at: https://elex.link/elex2015/.

Kernerman, Lionel. 1994. The advent of the semi-bilingual dictionary. *Password News* 1:1.

Nakamoto, Kyohei. 1994. Monolingual or bilingual, that is *not* the question: The 'bilingualised' dictionary. *Lexicon* 24. Tokyo: Iwasaki Linguistic Circle (Kenkyusha).

Reif, Joseph A. 1987. The development of a dictionary concept: An English learner's dictionary and an exotic alphabet. In *The Dictionary and the Language Learner: Papers from the Euralex Seminar at the University of Leeds, 1–3 April 1986*, edited by Anthony P. Cowie. Lexicographica Series Maior 17:140–158. Tübingen: Max Niemeier Verlag.