A multilingual trilogy:
Developing three multi-language lexicographic datasets

Ilan Kernerman

K Dictionaeies, 8 Nahum Hanavi Street, 63503 Tel Aviv
E-mail: ilan@kdictionaries.com

Abstract

This paper offers a brief overview of three multilingual developments by K Dictionaries and highlights the main editorial procedures involved and technical tools applied. The first regards an English multilingual dictionary bringing together 43 language versions of Password semi-bilingual dictionary. The second stems from the first, semi-automatically generating multilingual glossaries for any one of those languages to all others via detailed bilingual L2-English indexes. The third is part of the Global series and consists of monolingual datasets for over 20 languages that serve to create various bilingual and multilingual versions and multi-layered combinations. Further steps are anticipated in order to interlink and unify the different resources and processes, such as by associating translations in one lexicographic set to corresponding entries in others and thereby to more translations in other languages, and to converting the data to RDF format for interoperability with Linked Data and Semantic Web technologies.

Keywords: multilingual; dictionary; dataset; semi-automatic generation; linked data

1. Introduction

Multilingual linguistic resources are becoming exceedingly available, diversified and richly generated and used. Applying smart tools to their development and dissemination improves their quality and forms of usage, and increases their accessibility and popularity in a world opening up to cross-linking ever more languages. K Dictionaries (KD) first became involved in multi-language lexicography at the turn of the century with an English multilingual dictionary (EMD) project, and in recent years we have gone deeper into creating resources multilingually. This paper overviews three of our recent multilingual dictionary/lexicography processes, two of which are interrelated, and prospects for enhancing their interoperability both internally and externally for better technological application. First attempts to interconnect the KD data to Linked Data and integrate with Semantic Web technologies were undertaken last year, and more steps will include further multilingual adjustment of the different layers, resources and processes.

2. English multilingual dictionary

The first version of an EMD that assembles a number of semi-bilingual dictionaries for

learners of English was initiated in 2000 by Kielikone, a language technology company from Finland with experience in electronic dictionaries since the late 1980's (Herpiö, 2001). They used 20 language versions of Password dictionary, 18 of them sharing one common English core (based on *Chambers Concise Usage Dictionary*, CCUD) and two based on another (*Harrap's Easy English Dictionary*, HEED) [1], to publish GlobalDix as part of their MOT Dictionary Shelf and as a stand-alone product on CD-ROM and online, including platforms for Windows, Mac, Unix and Linux, intranet and mobile phone[2].

The semi-bilingual dictionary was launched by Kernerman Publishing in Israel in the mid-1980s for non-native learners of English and was later also known as the *bilingualized* dictionary (cf. Reif, 1987; Kernerman, L., 1994; Nakamoto, 1994). Its main innovation was to use the core of an English monolingual learner's dictionary with the addition of brief translation equivalents in the learner's native language for each sense of the entry. The first edition published for speakers of Hebrew (*Oxford Student's Dictionary for Hebrew Speakers*, 1986) was based on *Oxford Student's Dictionary of Current English* (1978), and the second for speakers of Arabic was based on HEED (*Harrap's English Dictionary for Speakers of Arabic*, 1987), which also served as a base for a few more languages. However, most semi-bilingual versions that followed in cooperation with local publishers worldwide were based on CCUD.

The beauty of GlobalDix was to present side by side translation equivalents for each specific sense of an English word or phrase (including definition and example) from semi-bilingual dictionaries for different languages, enabling the user to compare languages indirectly through the English intermediary. It thus served as a hybrid link for bilingual and multilingual matching, yet lacked full harmony among all the languages because of its reliance on two separate English layers. Another drawback was that while users could look up words in any of the languages, this search was restricted to the list of translations rather than to having a decent headword list for any of the languages.

Over the years KD proceeded to add new language versions to the EMD dataset, unified the English core around a single (CCUD-updated) base for all the language translations, introduced word-to-word reverse indexes for many of the languages to English and combined morphological links for English and certain languages (thus enhancing their searchability), and also upgraded the XML structure overall. The data has since been used in multiple forms and formats by different publishing partners worldwide, such as online dictionaries offering multi-language translations to English

[1] Chinese Simplified, Dutch, Finnish (WSOY), French, German, Hungarian, Icelandic (EDDA), Italian*, Japanese, Latvian (Zvaigzne), Lithuanian (Alma Littera), Norwegian (Aschehoug), Polish, Portuguese Brazil (Martins Fontes), Portuguese Portugal, Russian (Russky Yazik), Slovak (SPN), Spanish*, Swedish (Studentlitteratur), Turkish; language versions marked * are based on HEED, and all others on CCUD.

[2] cf. The 21-Language GlobalDix. *Kernerman Dictionary News*, 10, p. 3. (2002).

native speakers and foreign users (Dictionary.com, TheFreeDictionary) or semi-trilingual mobile apps including Korean and one more language equivalent to the English lemma for Korean speakers and foreign users (Daol), etc. Figure 1 presents an extract of an entry from a draft online 42-language version.


away [əˈwei]

♦ adverb

1 to or at a distance from the person speaking or the person or thing spoken about: *He lives three miles away (from the town); Go away!; Take it away!*

Afrikaans weg
Arabic بَعِيداً، في مَكان بَعِيد
Bulgarian разстояние
Catalan a; lluny
Chinese Simplified 离
Chinese Traditional 離
Croatian dalje
Czech daleko; pryč
Danish væk; bort(e)
Dutch weg
Estonian eemal(e), ära
Farsi دور؛ دور از
Finnish poissa
French (au) loin
German weg
Greek μακριά, σε απόσταση
Hebrew -מִמֶּרְחָק מ
Hindi दूर
Hungarian el, messzire
Icelandic burt
Indonesian jauh
Italian di distanza; via
Japanese 離れて
Korean 떨어져
Latvian projām
Lithuanian toli, šalin
Malay jauh
Norwegian bort(e)
Polish stąd
Portuguese Brazil longe
Portuguese Portugal longe
Romanian departe
Russian на таком-то расстоянии; прочь
Serbian odavde
Slovak odtiaľ; preč
Slovenian stran
Spanish a; lejos
Swedish från, ort[ifrån], i väg, undan
Thai ในที่ทางเดิน
Turkish uzakta, uzağa, başka yere
Ukrainian відділік; геть
Urdu دور ، دوری
Vietnamese xa

2 in the opposite direction: *She turned away so that he would not see her tears.*

Afrikaans weg
Arabic إلى الجهة المُعاكِسَه
Bulgarian в другата посока
Catalan cap a l'altre costat, en una altra direcció
Chinese Simplified 掉转
Chinese Traditional 掉轉
Croatian od
Czech stranou
Danish væk; den anden vej
Dutch weg
Estonian kõrvale
Farsi به سوی دیگر
Finnish vastakkaiseen suuntaan
French de l'autre côté
German weg
Greek προς την αντίθετη κατεύθυνση
Hebrew לְכִיוּוּן אַחֵר
Hindi विपरीत दिशा
Hungarian el(fordul)
Icelandic í burtu, undan
Indonesian pada arah yang berlawanan
Italian dall'altra parte
Japanese あちらへ
Korean 다른 방향으로
Latvian prom; projām
Lithuanian į šalį
Malay ke arah lain
Norwegian bort, vekk, unna
Polish w drugą stronę
Portuguese Brazil de costas
Portuguese Portugal de costas
Romanian în altă parte
Russian в сторону
Serbian u suprotnom pravcu
Slovak stranou, nabok
Slovenian vstran
Spanish hacia el otro lado, para otra parte
Swedish bort, om
Thai ในที่ทางตรงข้าม
Turkish başka yöne, başka tarafa
Ukrainian геть
Urdu مخالف سمت پر
Vietnamese tránh xa

Figure 1: Extract of an entry from a draft online 42-language version of the EMD

In 2013–2014 KD has undertaken a new round of thorough editorial revision and update of the (CCUD-based) English dictionary core, pursued by the translation of over 2,000 new entries in most of the language versions available then. The ensuing new EMD dataset currently contains a total of approximately 1.7 million translations in 43 languages, referring to 30,000 English entries (i.e. words and phrases) that include 39,000 senses with 38,000 examples of usage.

3. Multilingual glossaries

The EMD revision was succeeded since the end of 2014 by the development of newly refined reverse L2-English indexes that became the base for multilingual glossaries[3]. In the past, such indexes consisted simply of word-to-word lists, some including the part of speech of the L2 headword. The headwords were derived from the list of translations in the original semi-bilingual English dictionary for the particular L2, and were manually revised to keep, adjust or remove any item and to edit its matching English headword-turned-into-translation. The new indexes, however, were conceived to link the L2 headword precisely to each specific corresponding sense in polysemous entries of the original English dictionary core, rather than to the English headword, and finally list these English equivalents according to frequency and importance rather than in alphabetical order. Consequently, once a new L2-English index is ready it can be automatically turned into a multilingual glossary by associating the translations in all other languages for each sense of the English entry (now a translation). In this way, if N reverse indexes are made then N*N−1 new connections can be obtained. The following three simple steps can serve to portray the general process:

1. Have EN>DE, EN>ES, EN>FR, EN-RU (etc.)
2. Add FR>EN
3. Obtain FR>EN>DE, FR>EN>ES, FR>EN>RU (etc.)

The raw index is produced by automatic processing of the original English-L2 data, a process that incorporates some basic rules meant to help manipulate more complex data, for example pertaining to headwords and translations that happen to have variations (particularly regarding punctuation marks e.g. slash, brackets, comma).

Technically, the program first parses the EMD's XML files and creates basic tables. It searches all the Translation containers and compounds and associates each one with its Sense. The Sense set includes the following components:

- Translations for all the languages
- Definition
- Example(s) of usage

[3] The languages indexed and multilingualized so far include Catalan, Chinese Simplified, Danish, Dutch, Estonian, French, German, Hungarian, Indonesian, Italian, Japanese, Polish, Portuguese Brazil, Portuguese Portugal, Russian, Slovene, Spanish and Swedish.

- Headword and part of speech

The outcome of the initial parsing is illustrated in Figure 2:

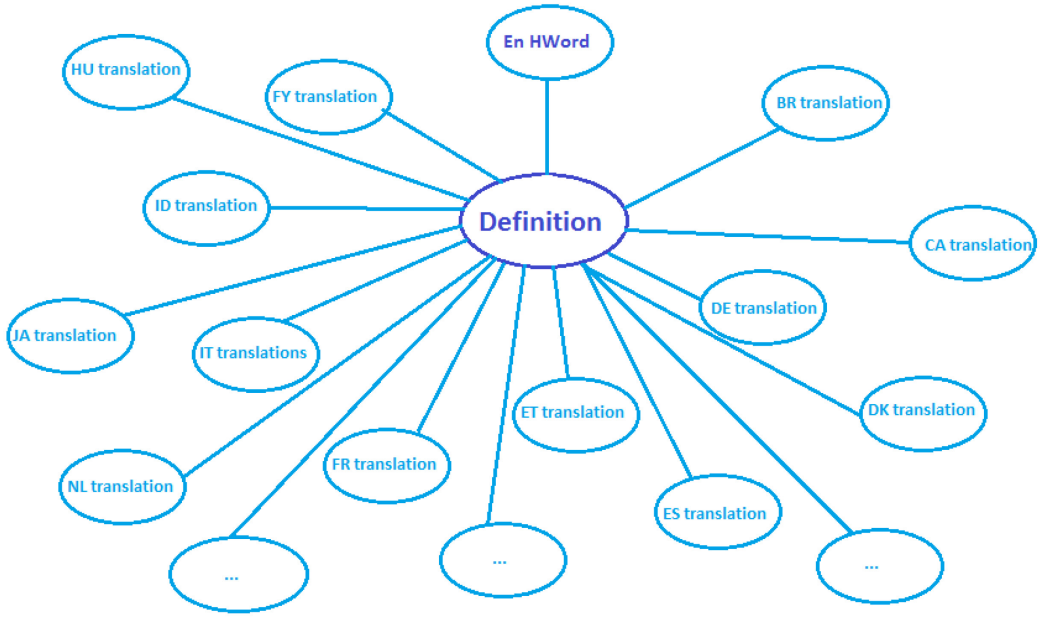

Figure 2: Parsing the XML data and preparing translations in different languages

The main characteristics of the Sense set consist of the Definition and the associated L2 Translation. Each Sense has an identifier, which will serve 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 the 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
- (English) Definition
- (English) Example of usage (if appropriate)

As a result, the L2 Translation (from EMD) becomes an L2 Headword. Now the program brings together all the Senses in the EMD that were associated with it as a Translation and lists them alphabetically (according to the original English Headword and Sense number). Subsequently, the L2 Headword is composed as follows:

- 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 ensuing raw index then undergoes thorough manual editing, using an especially dedicated software tool. In general, the editor reviews the L2 translations-turned-into-headwords to decide which items to keep intact, change into appropriate headwords or remove if not relevant, and adjusts their automatically allocated parts of speech. As for the English translation equivalents, the editor removes inappropriate ones and adds others, as well as rearranging them according to frequency and importance[4]. Figures 3 and 4 present sample screenshots of editing the index using this special tool.

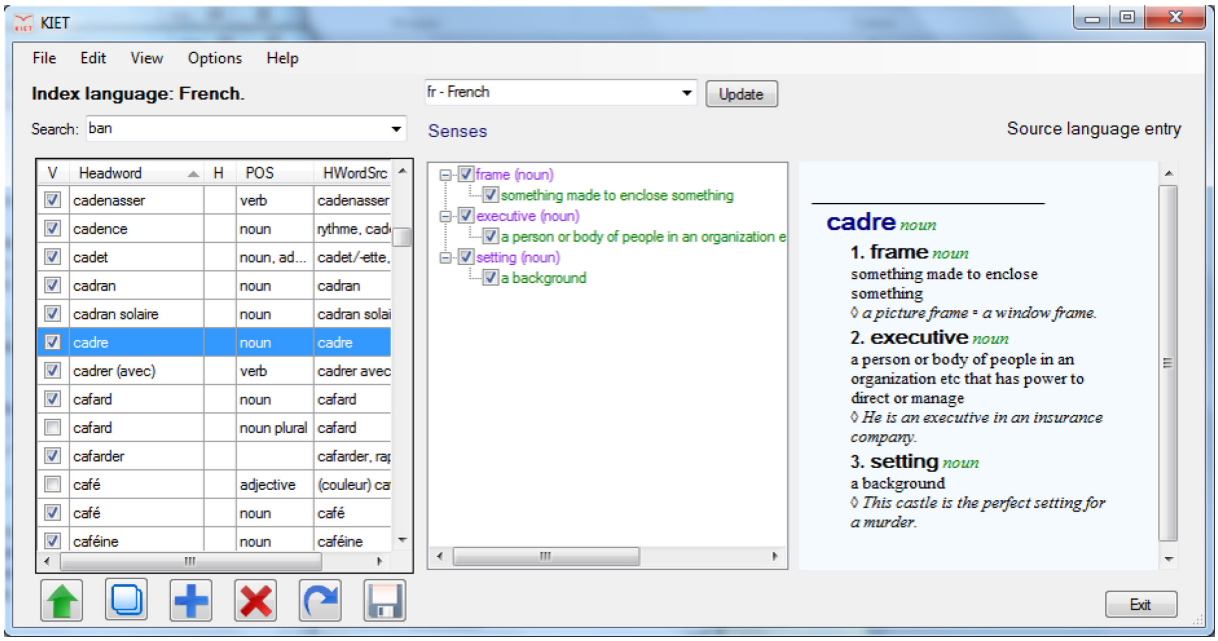

Figure 3: Editing the French Headwords in the Index Editorial Tool

[4] A detailed account of this editorial process is available in Egorova (2015) in this volume.

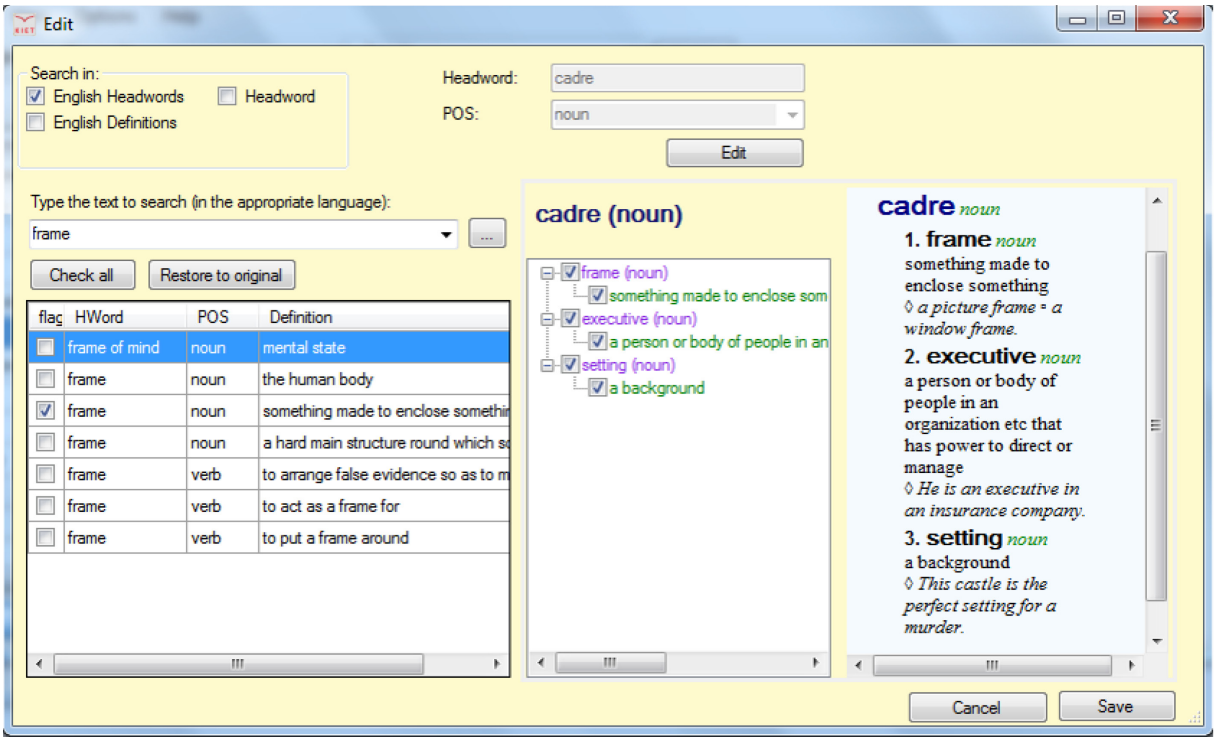

Figure 4: Editing the English Sense equivalents in the Index Editorial Tool

The detailed editing of the English translations according to each specifically matching sense, rather than just suiting the corresponding headword, offers a reasonable base to automatically produce fair-quality multilingual glossaries by adding the translations into all other languages from the EMD. Figures 5 and 6 present two samples of the results, the first featuring the English translation/sense with the other language translations derived from it, and the second integrating the English equivalent together with all other languages without exposing its fundamental linking role.

messen *verb*

1. gauge to measure (something) very accurately
2. measure to find the size, amount etc of (sth)
3. measure to show the size, amount etc of
4. measure (with **against**, **besides** etc) to judge in comparison with
5. measure to be a certain size
6. meter to measure (*especially* electricity etc) by using a meter
7. take to make a note, record etc

Figure 5: German multilingual entry exposing its primary English equivalent link

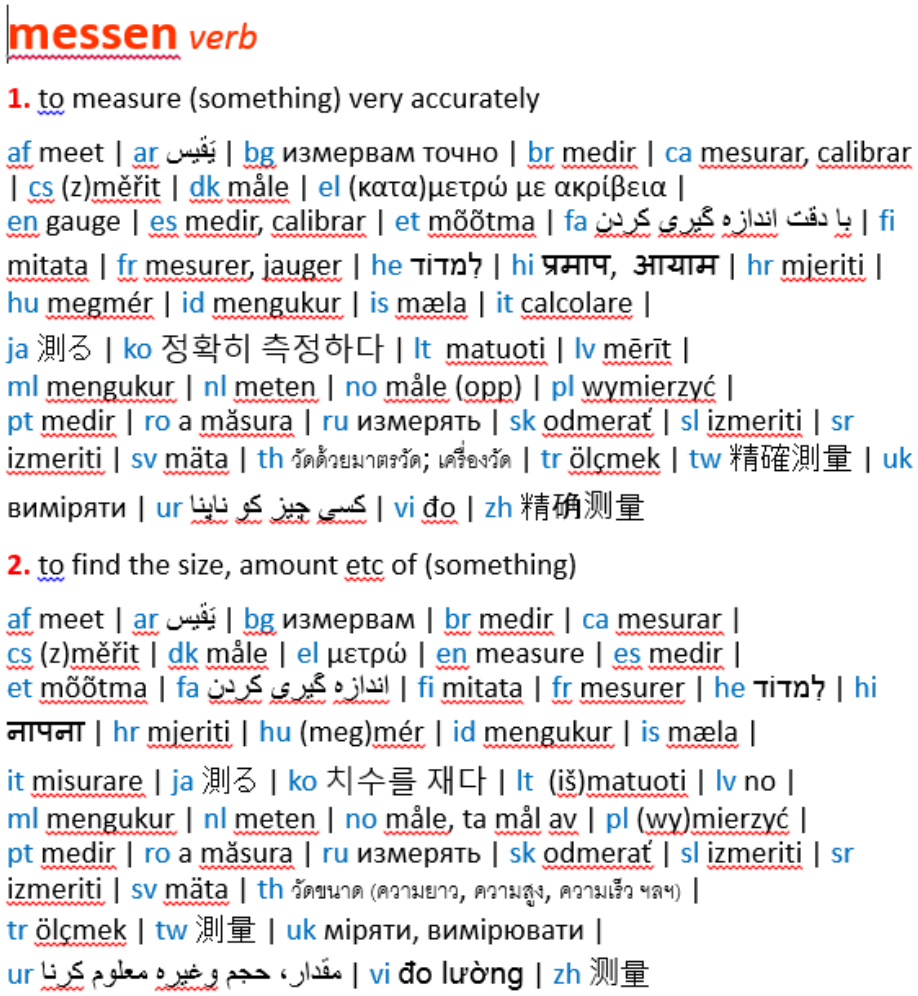

Figure 6: German multilingual entry combining its primary English equivalent link with all the other language equivalents

Unfortunately, these automatically-generated multilingual glossaries are bound to contain inaccuracies due to the indirect nature of juxtaposing different languages via the English common ground. Nevertheless, they offer some merit for basic translation purposes and serve as an advanced base for amending higher quality matching, useful in particular for less-common language pairs. At this stage, there is no information about the precise rates of the "inaccuracies" in the L2–L3 automatic matching, and this remains to be further investigated.

4. Fully multilingual dictionaries

In 2005 KD began to create the Global series, with the first multilingual combinations becoming available since 2009[5]. The Global series has its foundation in monolingual

[5] KD's BLDS: A brief introduction. *Kernerman Dictionary News*, 17, pp. 1–2 (2009).

lexicographic datasets for different languages (Kernerman, I., 2011)[6], each serving as a base for adding translations and developing bilingual dictionaries. Thus, whenever one of the core languages has several bilingual versions, putting their data together produces a multilingual dictionary. This process is similar in principle to that of composing the EMD. However, the Global entry microstructure is much more elaborate and allows for more than one translation equivalent per sense, as compared to usually just a single translation per language in the EMD. In addition, the examples of usage are translated as well, unlike the EMD's semi-bilingual base that has translations only for the meanings of the word or phrase. These differences lead to significantly richer results. Moreover, since the languages that consist of translations exist also as L1 cores in the Global series, many of the translations can be associated to their full entries and the information provided can be (re-)expanded again and again. Figures 7, 8 and 9 display French monolingual, bilingual and multilingual entries, respectively.

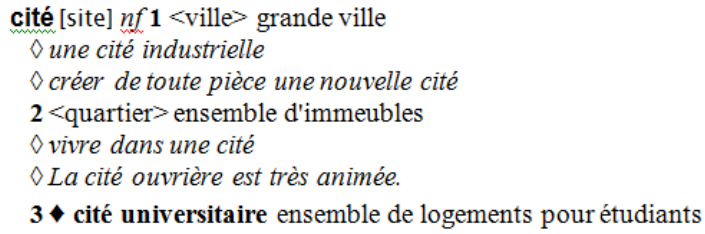

Figure 7: Global French monolingual entry

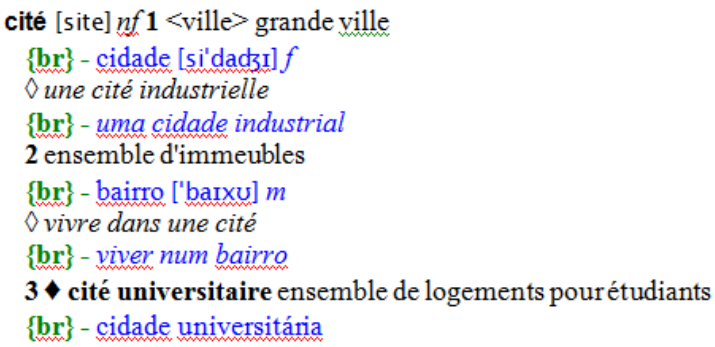

Figure 8: Global French bilingual entry (French–Portuguese)

[6] Global series language cores available so far include Arabic, Chinese Simplified, Chinese Traditional, Czech, Danish, Dutch, English, French, German, Greek, Hebrew, Italian, Japanese, Korean, Latin, Norwegian, Polish, Portuguese Brazil, Portuguese Portugal, Russian, Spanish, Swedish, Thai and Turkish.

Figure 9: Global French multilingual entry

5. Further developments

In 2014 KD had a first taste of converting its data from XML (Extended Markup Language) to RDF (Resource Description Framework)[7] format, based on the Lexicon Model for Ontologies (*lemon*)[8], through academic cooperation at Madrid Polytechnic University and Leipzig University (Klimek & Brümmer, 2015). RDF is a data model developed by the World Wide Web Consortium (W3C), serving as the basic mechanism to formally describe any type of *resource* – whether words, documents, people, physical objects or abstract concepts – along a *subject-object-predicate* pattern and thus making it more easily sharable and interconnectable (Gracia, 2015). The RDF transformation is a vital step in uniformizing our lexicographic datasets into a common structure in order to facilitate cross-linking content from different dictionaries, enriching it by exterior multi-language lexical and other resources, and having it published as Linked Data on the Web.

The processes described in sections 2 and 3 already constitute attempts to link our internal resources to each other, and thereby expand them exponentially, and the same can be said about the fairly simple and straightforward process described in section 4. Next challenges consist of linking the various Global language core resources to each other – such as by linking an L2 translation to the information it has as an (L1) entry in its own monolingual set and on to translations in L3, L4, etc. – and to other internal resources such as the EMD and multilingual glossaries. For example, the Portuguese translation in Figure 7 could be linked to that lemma which exists as a headword in

[7] Resource Description Framework, cf. http://www.w3.org/rdf
[8] http://www.lemon-model.net

the Portuguese core with its translations to another language, and so on and so forth. Likewise, the same item could be linked (also) to the Portuguese translation in the EMD and to the multilingual information it has as part of the Portuguese glossary. This development can be defined as moving on from *multilingual* to *multilayer*, in the sense that each language part in any of the lexicographic datasets constitutes one layer of information and that these different layers are interconnected, as part of further expansion of these multi-language opportunities.

Whereas the internal process described above could suffice with keeping the data in XML format and is just enhanced by its RDFication, linking with other resources on the Web relies exclusively on the RDF format. For example, the data could then be enriched by open resources such as WordNet, Wiktionary and Babelnet, to name just a few well-known open source lexical websites. KD is starting to develop a new API that will enable such exterior linking, both for extracting new data from other resources and for disseminating its own data more efficiently to others. The data manipulation described in this paper may seem in parts as a revolution with respect to traditional lexicography, but it still only scratches the surface of a new threshold to future prospects.

6. Acknowledgements

I would like to thank the reviewers of this paper for their comments, in particular Reviewer 3 for the detailed remarks.

An earlier version of this paper was presented at the 20th Biennial Dictionary Society of North America Meeting (DSNA-20), held at the University of British Columbia in Vancouver (BC), on 6 June 2015.

7. References

Egorova, K. (2015). Editing an automatically-generated Russian-English index. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Linking lexical data in the digital age. Proceedings of eLex 2015, Herstmonceux Castle, 11-13 August 2015. Herstmonceux Castle, UK.* Available at: https://elex.link/elex2015/.

Gracia, J. (2015). Multilingual dictionaries and the Web of Data. *Kernerman Dictionary News*, 23, pp. 1–4.

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

Kernerman, I. (2011). From dictionaries to databases: Developing a global series for language learners. In I. Kosem & K. Kosem (eds.) *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011, Bled, 10-12 November 2011.* Bled, Slovenia. Available at: http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-0.pdf.

Kernerman, L. (1994). The advent of the semi-bilingual dictionary. *Password News*, 1, p. 1.

Klimek, B. & Brümmer, M. (2015). Enhancing lexicography with semantic language databases. *Kernerman Dictionary News*, 23, pp. 5–10.

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

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

Websites:

Babelnet. Accessed at http://www.babelnet.org. (12 June 2015)

Dictionary.com. Accessed at: http://www.dictionary.com. (1 August 2007)

TheFreeDictionary. Accessed at: http://www.thefreedictionary.com. (12 March 2009)

W3C. World Wide Web Consortium. Accessed at: http://www.w3.org. (12 June 2015)

Wiktionary. Accessed at: http://www.wiktionary.org. (12 June 2015)

Wordnet. Accessed at: http://www.wordnet.princeton.edu. (12 June 2015)

Dictionaries:

Daol. *Kernerman Semi-Trilingual English Korean L3 Dictionaries*. (2013). Seoul: DaolSoft.

CCUD. *Chambers Concise Usage Dictionary*. (1986). Edinburgh: W&R Chambers.

GlobalDix. *GlobalDix*. (2001). Helsinki: Kielikone.

Harrap's English Dictionary for Speakers of Arabic. (1987). Toronto: Kernerman Publishing.

HEED. *Harrap's Easy English Dictionary*. (1980). London: Harrap.

Oxford Student's Dictionary for Hebrew Speakers. (1986). Tel Aviv: Kernerman Publishing.

Oxford Student's Dictionary of Current English. (1978). Oxford: Oxford University Press.