From Dictionary to Database: Creating a Global Multi-Language Series

Ilan Kernerman
K Dictionaries Ltd
Nahum 8, Tel Aviv 63503, Israel
E-mail: ilan@kdictionaries.com

Abstract

K Dictionaries is developing a global dictionary series for learners and general users, which includes over twenty languages to date. Each language core consists of a database that is used for creating monolingual dictionaries, and for adding translations and making bilingual dictionaries and eventually multilingual ones. A single macrostructure is used for the entire series, and the same microstructure is applied for the entries of all the dictionaries – while being adapted to suit the characteristics of each language. The cores can be modified in relation to any language pair and different components extracted for particular audiences in print editions and digital applications. The data is formatted in XML and Unicode, and the DTD has undergone numerous amendments and improvements over the time to solidify the structure and enrich it. The work is done by lexicographers worldwide using a dedicated editing software, whether offline or online, aided by editorial styleguides, technical manuals and full support from the K Dictionaries team. Further use will include data reversal, linking translations to their own language cores and enhancing multilingual web linkage.

Keywords: dictionary; database; content; technology; application

1. Introduction

In 2005, K Dictionaries (KD) began working on eight bilingual dictionaries for French learners of foreign languages[1]. We started by developing a monolingual dictionary database for French and for each of the eight languages, and then proceeded to translate the French core to the eight languages and these languages to French. The process took four years and the dictionaries were published in France in 2009, with each print edition accompanied by an electronic version for PC.

During that period more language cores and language pairs were compiled, and the French project has become the cornerstone of a global series consisting at present of over twenty languages[2], including forty bilingual titles, twenty in preparation and twenty more currently in planning. Over the last six years the entry structure has evolved and some of the language cores were expanded. The dictionaries are gradually released worldwide in different forms for all types of media.

To enable the creation of such extensive, complex and multilingual data, as well as its efficient processing, maintenance, updating and application, it was necessary to establish and nourish a fine technological infrastructure, which is extremely robust and solid on the one hand and open and flexible on the other hand. From the start we decided to encode the data in XML format and use Unicode to enable work on practically any language[3]. An XML Editor was configured according to the entry structure and provided to the lexicographers for their compilation. The DTD took several months to devise and continued to be modified all along to improve its microstructure and accommodate additions and changes. The data can thus be used with adaptable XSL documents to prepare for print and various digital versions. A desktop application was developed to produce an electronic version of each dictionary, and eventually online versions were conceived as well.

To our knowledge, this is a first such systematic attempt to develop a worldwide dictionary series for language learners and general users on such a relatively broad scale. Our editorial and technical concepts have evolved in the making, many errors were made and changes introduced. Some of our lessons may prove valuable to others as well.

2. Content

On the outset, the main target users of the initial French bilingual dictionary series were identified as French native speakers learning a foreign language at lower to intermediate levels. The publisher, Assimil, who is a leader in foreign language teaching materials in France, was keen on offering this added value to users of its *Méthod Assimil* coursebooks. At the same time, these dictionaries had to stand on their own and satisfy other users too.

It was decided that the dictionaries would be

[1] Arabic, Chinese, Greek, Japanese, Polish, Portuguese, Russian, Turkish.

[2] The cores currently available include Arabic, Chinese (Simplified, Traditional), Czech, Dutch, English, French, German, Greek, Italian, Japanese, Korean, Latin, Norwegian, Polish, Portuguese (Brazil, Portugal), Russian, Spanish, Thai and Turkish; in addition, Hebrew, Hindi and Swedish are due.

[3] XML (Extensible Markup Language) is a set of rules for encoding documents in machine-readable format. The DTD (Document Type Definition) is a set of markup declarations that define a document type for XML, and the XSL (Extensible Stylesheet Language) is used to transform and render XML documents. Unicode is a standard for consistent text encoding, representation and handling in most writing systems. (Definitions from Wikipedia, http://wikipedia.org/)

medium-size and bi-directional, catering for both encoding and decoding purposes. To satisfy Assimil's needs, they had to cover the vocabulary in their *Méthod* for each language. They were going to have up to 680 pages, but when their data proved to be growing size this limit was first extended to 800 pages and finally the average page number is about 1,000. The print edition was going to be accompanied by an electronic version on CD-ROM.

We began by preparing a sort of monolingual learner's dictionary core for each language, which could function as a base for adding translations and developing bilingual dictionaries. The entries include a brief definition for each sense, and examples often consist of short phrases rather than full sentences. The defining vocabulary consists of all the headwords in the dictionary and their inflections.

Generally-speaking, developing the L1 database puts into play a *deconstruction* of the language, mapping its 'atoms and quarks', and enabling its *reconstruction* in a variety of lexicographic terms. Then, in the translation process, a brand new equilibrium is reset concerning each specific language pair in play.

An important editorial change occurred half-way through the compilation, in the form of replacing the definition in French entries by sense indicators. The reasoning was to facilitate the use for French users and focus solely on their need for disambiguation of the different meanings of polysemous French words. Figure 1 demonstrates the French definitions in the first occurrence substituted by sense indicators in the second (both underlined):

accueil [akœj] *nm*
1 manière de recevoir qqn ou qqch
◊ *faire bon / mauvais accueil à qqn*
2 lieu où l'on reçoit des visiteurs
◊ *Adressez-vous à l'accueil !*
accueil [akœj] *nm*
1 réception ◊ *faire bon / mauvais accueil à qqn*
2 lieu ◊ *Adressez-vous à l'accueil !*

Figure 1: Replacing definitions by sense indicators

To save on space in the print editions, the definitions were later removed altogether from the entries of the other languages, but were kept in the electronic versions. The translation consists chiefly of providing L2 equivalents for each sense, example and phrase, often including pronunciation and grammatical details. The translators may modify the L1 entry to suit it to the L2, particularly by changing examples of usage, and sometimes they also suggest substantial changes in the L1 entry structure, such as concerning the classification of senses, that can help to improve it. Figure 2 demonstrates an alteration of the original example of usage from the Arabic version to the Russian one, to better suit the French-Russian pair:

auxiliaire *adj*
1 qui aide [muˈsaːʔid] **مساعد**
◊ *personnel auxiliaire* شخص مساعد
auxiliaire *adj*
1 qui aide **вспомога́тельный** [fspəmaˈgatʲilʲnɨj], **подсо́бный** [patˈsobnɨj]
◊ *économie auxiliaire* подсо́бное хозя́йство

Figure 2: Modifying the example of usage

Basically, each language core consists of 12,000 main headwords[4], but some languages were doubled or quadrupled[5]. The list of headwords was devised primarily according to frequency and importance. The principle components of the entry are outlined in Table 1:

○ *the headword*:	
L1	headword, pronunciation and alternative script[6], part of speech, grammatical gender and number, irregular forms
○ *the attributes*:	
L1	geographical usage, subject field, register, sense qualifier, range of application, synonyms, antonyms, notes
○ *the sense indicator*:	
L1	(when no attribute was indicated for a sense of a polysemous entry) a hyperonym or 'preposition- filler' (e.g. *of* something) – for disambiguation
○ *the definition*:	
L1	a succinct definition for each sense
L2	translation for each sense of the headword (NOT of its definition), pronunciation, grammatical gender and number as appropriate
○ *the examples of usage*:	
L1	example(s) of usage for each sense of polysemous entries, usually consisting of short phrases rather than full sentences
L2	idiomatic translation of the example(s)
○ *the compositional phrases*:	
L1	collocational phrases (idioms, compounds, etc), with/without definition and/or example – may be part of a given sense or form a sense of its own
L2	idiomatic translation of the phrase (and its examples of usage)
○ *the sub-headwords*:	
L1	run-ons usually consist of part of speech change and derivates, and may include part or all of main headword components
L2	translation of each L1 component, as above

Table 1: The entry's main components

[4] Chinese has 3,000 main headwords consisting of single characters, and their derivations appear in the form of sub-entries

[5] Dutch, French, German, Italian and Portuguese have 25,000 main entries, Spanish has 50,000.

[6] The IPA (International Phonetic Alphabet) is used for most languages. Chinese includes Pinyin, Hebrew includes alternate script with 'vowels' (*nikud*), and Japanese includes Kanji, Katakana, Hiragana and Romaji.

The French titles were published by Assimil in 2009 (DAK), each accompanied by a desktop application that is downloaded from the publisher's server by using an individual code that is inserted in the inside back-cover, and installed on the user's PC.

In the meantime, work on more languages was launched, and the first dictionary to actually appear in the global series was Norwegian/Spanish in Norway, accompanied by the electronic version on CD-ROM (*Spansk ordbok*). The publisher, Vega, published the next title, Italian, in 2010 (*Italiensk ordbok*). German is published this autumn, to be followed by French and Polish in spring 2012.

The next countries in which titles are due to appear in print are Japan (2011) and Brazil (2012). Meanwhile, Dutch bilinguals have been available online since 2010 on http://mijnwoordenboek.nl/, and iPhone applications were released by Abbyy as part of their Lingvo series.

3. Technology

K Dictionaries operates worldwide on the development of lexicographic content, and its final dictionary products and services are released by others. In this spirit, since the turn of the century our cooperation has gradually shifted from traditional publishers to a wide range of technology firms, while our focus has increasingly centred on creating content that can be used and re-used, again and again, whether fully or partially, for any purpose and in any type of media. To attain the goal of offering high-quality content we must rely on high-technology. Moreover, we started to develop our own electronic applications, both to experiment with how the dictionaries might look like and to offer them to 'low-tech' partners from the publishing industry. As a result our company has become to some extent a *technology-based* (and *technology-oriented*) content provider.

As an aside, for a long time we have considered that all contemporary *lexicography* actually is (or should be) *e-lexicography*, so this distinction today is redundant (and not quite justifiable for the last quarter century, since the advent of corpus-based dictionaries[7]).

From the outset of this project we devoted close attention to the technology-related aspects of creating the data and putting it to play. Nevertheless, fresh insights and new issues kept appearing all along, leading to constant amendment of the editing software and the other tools used for our work.

Overall, it has become increasingly difficult and artificial to separate the content from the technology features of our work, and the division made in this paper is not fully precise and is mainly intended to serve its description purposes. The terms used herein are thus not always applied in their conventional sense. The following is a general overview of some of the main technology-derived dimensions of this project.

3.1 Editing Software

To maintain our independence and freedom of creation, as well as to help to reduce costs, we opted to use an XML editor to compile the data and to configure it for our purposes on our own, rather than utilize an existing dictionary writing system. However, it might come as no surprise to those familiar with this topic if we admit that after six years of experience it is still not absolutely clear which of these options is actually preferable and whether there indeed is any clear-cut answer to this question.

The editing software was developed for over half a year, but has continued to undergo endless updates since then. The main reasons for revising this software were:

- correction and improvement;
- adjustment to characteristics of new languages;
- adaptation to modifications in entry components;
- compatibility with changing operating systems.

The editing is done at a distance, usually from the editor's home, and the software can be used for either compiling L1 entries or adding L2 translation equivalents. The use is enabled either online or offline, and to work offline the editors must first install the software on their machine, which often requires personal adjustment to each individual computer and operating system. The software is accompanied by detailed documentation concerning its installation and manipulation, including a meticulous account of the entry microstructure. Whenever necessary, our staff provides full technical support for any need the lexicographer might have. Figure 3 demonstrates an extract of a Dutch entry in the XML Editor:

[7] COBUILD (1987) was the first dictionary to be entirely based on a computerized corpus.

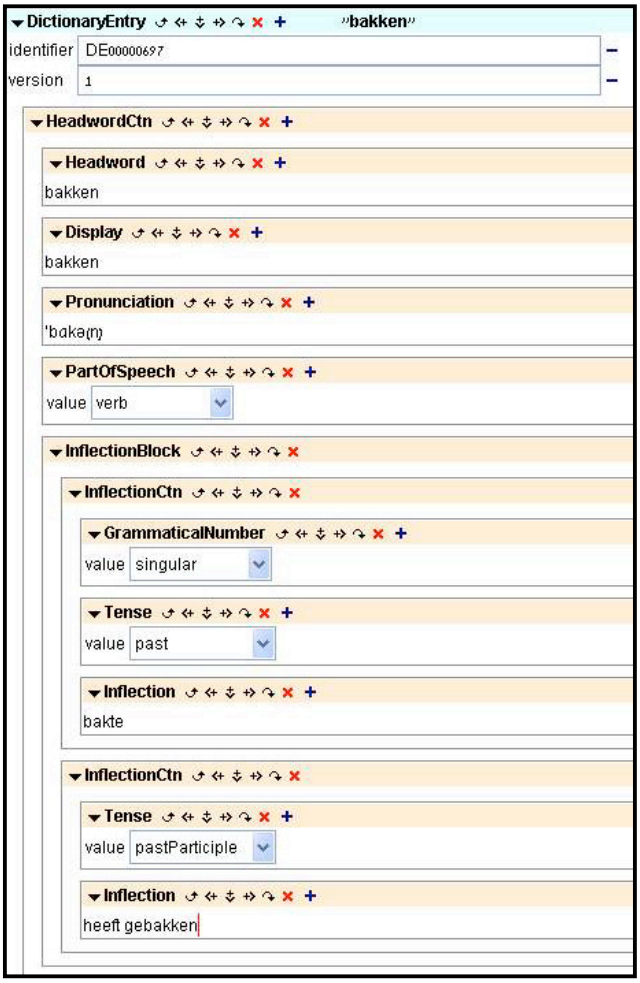

Figure 3: An extract of an entry in the XML Editor

3.2 Headword List

The chief editor of each language begins by preparing an editorial styleguide for compiling the entries and by drafting the list of main and sub-headwords. The headword list includes also the part of speech and cross-references, and for certain languages other elements are also indicated, such as alternative scripts. The headwords are selected according to frequency and importance, and will form the essence of the entries. Table 2 demonstrates an extract from a typical headword list.

The technology team then processes this headword information into 100 XML files, including 120 headwords each (i.e. 12,000 main entries in total), which will form the backbone for compiling the actual entries. In the compilation process it is possible to remove, add and change headwords from the initial list.

MAIN	RUNON	POS	XREF
antever		vt	
antiaéreo		adj	
antialérgico		n	
antialérgico	antialérgico	adj	
antibiótico		n	
anticaspa		adj	
anticoncepcional		adj	
anticoncepcional	anticoncepcional	n	
anticonceptivo			anticoncepcional
antidepressivo		n	

Table 2: An extract from the Portuguese (Portugal) headword list

3.3 Entry Microstructure

The editing software provides in advance for all possible components of the entry, many of which are going to be selected from a pre-defined dropdown menu. Elements that were not pre-defined in the software can be attributed as free values. For example, the word class for many languages will indicate the grammatical gender and grammatical number, as demonstrated in the DTD extracts in Figure 4:

```
<!ELEMENT GrammaticalGender EMPTY>
<!ATTLIST GrammaticalGender
  value notPredefined
  |masculine|feminine|masculine-feminine|
  neuter|masculine-neuter|masculine-feminine-n
  euter|feminine-masculine) #REQUIRED
  freeValue CDATA #IMPLIED
>
<!ELEMENT GrammaticalNumber EMPTY>
<!ATTLIST GrammaticalNumber
  modifier (only|usually) #IMPLIED
  value (notPredefined|singular|dual|plural)
  #REQUIRED
  freeValue CDATA #IMPLIED
>
```

Figure 4: Extracts from the DTD

3.4 Metalanguage & Localization

To facilitate the work on all different languages within a single macrostructure, all the labels used in the editing software appear in English, abbreviated. Their L1 equivalents are recorded in a separate list and will eventually replace the English labels in the final product. Other localization documents include in particular the transcription key for each language. Table 3 presents a brief extract from the key for Arabic:

IPA	(Sampa)	Unicode	EditorSet	Transcription	Translation	Arabic
Consonants						
Plosives						
b	b		Keyboard	ba:b	door	باب
t	t		Keyboard	tis?'	nine	تسع
D	d		Keyboard	da:r	home	دار
t'	t`	t + 02bc 700	Keyboard + IPA Ext	t'a:bi?'	stamp	طابع
d'	d`	d + 02bc 700	keyboard + IPA Ext	d'arab	he hit	ضرب

Table 3: Extract from the Transcription Key for Arabic

3.5 Editor-Friendly

The editing software features a Preview button that enables the lexicographer to instantly view the XML data in a clear reading style of an HTML document, to easily review the entries and introduce changes on the fly. Figure 5 demonstrates an HTML preview of the Spanish-Dutch entry that appeared in Figure 3 above:

estribo [estɾi'βo] *nm* **1** <en equitación> pieza de la silla de montar en que coloca los pies un jinete

{nl} - stijgbeugel *de*

◊ *estribos de plata*

{nl} - *zilveren stijgbeugels*

2 <apoyo para el pie> plataforma a modo de escalón que sirve para ascender a un vehículo

{nl} - voetrust *de*

◊ *los estribos de una motocicleta*

{nl} - *de voetrusten van een motorfiets*

♦ **perder los estribos** perder por enfado el control de uno mismo

{nl} - uit zijn vel springen

◊ *Me haces perder los estribos.*

{nl} - *Ik spring door jou uit mijn vel.*

Figure 5: An extract from a Spanish-Dutch entry in the HTML preview

Recently we devised an alternative for using the XML Editor to enable translators to work directly on an MS Word DOC, where they just need to insert the translation components in ready-made fields that were converted from the XML data and will be re-converted back to XML with the help of the ID tags. The number of possible translation equivalents is limited to three per sense and one for each example and expression. Figure 6 shows a sample Dutch-German translation in DOC:

bakken ['bɑkə(n)] *v* (*sg pt* **bakte**, *pp* **heeft gebakken**) **1** <[voedsel]> eten in een koekenpan op heet vuur of in een hete oven gaar laten worden

{ TC00002151: Translation [backen] }

{ TC00002151: Translation [braten] }

◊ *een ei bakken*

{ TC00002152: Translation [ein Ei braten] }

◊ *een taart bakken*

{ TC00002153: Translation [eine Torte backen] }

♦ **gebakken lucht** alles wat iemand zegt of doet die overdrijft

{ TC00002154: Translation [heiße Luft] }

♦ **er niets van bakken** iets helemaal niet kunnen

{ TC00002156: Translation [nichts gebacken bekommen] }

Figure 6: Translation in an MS Word document

3.6 Corpus

When we first set on the French project we discovered there were no publicly accessible corpora for French and the other languages. It was therefore decided to rely on whatever private corpora held by some of the editors and to refer cautiously to evidence found on the Internet. In addition, each editor-in-chief signalled out existing dictionaries that the lexicographers may turn to for general reference only, with copying strictly forbidden. Recently we began using the Sketch Engine corpus of Lexical Computing for Dutch[8], with satisfactory results.

3.7 Defining Vocabulary

This paper generally forgoes most editorial aspects that are not strongly concerned with the technological features of the series, but one that is most noteworthy and also somewhat underlines the *e* spirit of our venture is concerned with the defining vocabulary. It was decided that the vocabulary for each language will consist of all the headwords and their various inflections.

[8] http://sketchengine.co.uk/.

The reason is our expectation that most of the use of these dictionaries will be electronically, so that hyperlinking any word in the text to its appropriate entry should be possible. To enable this, all that is necessary is to have morphological connections among all the words.

Thus, once the full language core is compiled, we process a list of all the words used therein, then proceed to associate each word that is not a headword to its entry.

Unfortunately, although such an extensive defining vocabulary (based on the 12,000 main headwords) would seem to grant tremendous room for the lexicographers to maneuver, we found that often they did not fully abide by this regulation and did include other words in their definitions. Our solution to this problem will be to incorporate a new feature in the software that alerts the editor to any word that is not part of the list of headwords.

Meantime, those words that cannot be associated to entries become prime candidates for inclusion in any expansion of the dictionary.

3.8 QA & Processing Tools

Once data is received from editors or translators, it undergoes checking by the project manager as well as initial automated analysis of the contents, to basically confirm that all relevant components are in place. For example, that each sense has at least one example of usage, that an attribute or an indicator is included in addition to the definition, that the translation is accompanied by its phonetic transcription, etc. Figure 7 shows the home screen of our Utilities tool (currently being revised):

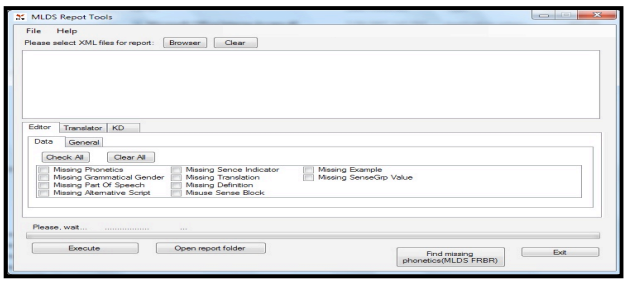

Figure 7: A snapshot of (former) automatic QA tool

Further processing of the data might concern changes made in the L1 entry during the translation, listing words that are not on the headword list, running statistics on the number of senses or examples, etc.

Each language core is organized in its own database, with all the translations available for the L1, as demonstrated in Figure 8. The data construction makes it possible to process specific language combinations or extract any components.

déballer [debale] *vt* <sortir de> sortir qqch de l'endroit où il était
{ar} - فَرَدَ [fa'rada]
{br} - desembalar [dʒɪzemba'lar], desembrulhar [dʒɪzembru'ʎar]
{de} - auspacken
{el} - ξεπακετάρω [ksepace'taro], ξετυλίγω [kseti'liɣo]
{es} - desembalar [ðesemba'lar]
{it} - sballare [zba'l:are], disimballare [dizimba'l:are]
{ja} - 取（と）り出（だ）す、荷（に）ほどきする toridasu, nihodoki suru
{nl} - uitpakken
{no} - å pakke ut
{pl} - rozpakowywać [rɔspakɔvɨvatʂ]
{pt} - desembalar [dəzẽbe'lar], desembrulhar [dəzẽbru'ʎar]
{ru} - распако́вывать [rəspa'kovɨvətʲ], расклáдывать [ras'kladɨvətʲ]
{tr} - boşaltmak
{zh} - 开箱,拆包;取出 kāixiāng，chāibāo ; qǔchū
◊ *déballer ses vêtements*
{ar} - *ملابسه فرد*
{br} - *desembalar as roupas*
{de} - *seine Kleidung auspacken*
{el} - *ξεπακετάρω τα ρούχα*
{es} - *desembalar sus ropas*
{it} - *disfare le valigie*
{ja} - *衣類（いるい）を取り出す* irui o toridasu
{nl} - *zijn kleren uitpakken*
{no} - *å pakke ut klærne sine*
{pl} - *rozpakować swoje ubrania*
{pt} - *desembalar as roupas*
{ru} - *расклáдывать вéщи*
{tr} - *giysilerini boşaltmak*
{zh} - *取出衣物* qǔchū yīwù

Figure 8: An HTML preview of an extract from the French multilingual database

4. Application

Having the raw dictionary data in XML format enables putting it into use in numerous ways, whether for print or digital media, or for applying different components to suit different user groups.

4.1 Same Translation in Polysemous Entries

Usually each sense of the entry has all its relevant information appearing together, including the translation equivalent of the meaning and the example of usage. However, in DAK it was decided that when all senses of a polysemous entry happen to have the same translation equivalent it will be placed at the headword level before the first sense, to facilitate the user's comprehension of the entire entry and to economize on the typographical representation. Figure 9 demonstrates two ways of displaying an entry, when each sense has a different translation and when all the senses have the same translation.

croyance *nf*
1 fait de croire **wiara** [vjara] *f* ◊ *la croyance à la liberté wiara w wolność* ◊ *la croyance en Dieu wiara w Boga*
2 *conviction* **wierzenie** [vjɛʒɛɲɛ] *nt* ◊ *les croyances religieuses wierzenia religijne*
croyance *nf* **crença** [ˈkrẽsɐ] *f*
1 fait de croire ◊ *la croyance à la liberté* crença na liberdade ◊ *la croyance en Dieu crença em Deus*
2 conviction ◊ *les croyances religieuses* as crenças religiosas

Figure 9: Different presentation of an entry when each sense has a different translation (Polish, above) or the same translation (Portuguese, below)

4.2 Same Entry for Different Users

Since the data contains all the relevant entry components, it is possible to make use of different elements to cater specifically for each target group. The next two figures demonstrate different uses of the French/Japanese database in dictionaries targeted for French learners of Japanese and for Japanese learners of French.

Figure 10 shows a sample entry in the French-Japanese section. For French users it provides sense indicators and Romaji, whereas for Japanese users it provides phonetic transcription of the headword, definitions instead of sense indicators, and no Romaji.

destin *nm*
1 fatalité **運命（うんめい）、宿命（しゅくめい） unmee, shukumee** ◊ *accepter son destin 運命を受（う）け入（い）れる* unmee o ukeireru
2 vie **人生（じんせい）、生涯（しょうがい） jinsee, shoogai** ◊ *un destin cruel 残酷（ざんこく）な人生* zankoku na jinsee
destin [dɛstɛ̃] *nm*
1 avenir décidé à l'avance **運命、宿命** ◊ *accepter son destin 運命を受*
2 existence, vie **人生、生涯** ◊ *un destin cruel 残酷*

Figure 10: French-Japanese entry for French speakers (above) and Japanese speakers (below)

Figure 11 shows a sample entry in the Japanese-French section. For French users the entries are arranged in Roman alphabetical order and the Romaji script of the headword appears first, and the entry includes definitions. For Japanese users the entries are arranged according to Kanji and do not include Romaji, the part of speech of the headword is in Japanese, there are sense indicators for the purpose of disambiguation, rather than definitions, and the French translation of each sense is accompanied by its phonetic transcription.

ho¯o1 ほう 方 *n*
1; 方向（ほうこう）, 方面（ほうめん）hookoo, hoomen **direction** ◊*山の方へ行く yama no hoo e iku aller en direction de la montagne*
2 いくつかの中のひとつ ikutsuka no naka no hitotsu **ceci** ◊*こちらの方を選ぶ kochira no hoo o erabu choisir celui-là*
ほう 方 *名詞*
1 =方向 **direction** [dirɛksjɔ̃] *f* ◊ *山の方へ行く aller en direction de la montagne*
2 選択 **ceci** [səsi] ◊*こちらの方を選ぶ choisir celui-là*

Figure 11: Japanese-French entry for French speakers (above) and Japanese speakers (below)

4.3 Desktop (Offline) Application

The print edition of each dictionary is accompanied by an electronic version that the user can install on his/her computer. It can either be downloaded from the publisher's server, by inserting an individual code that appears in the book, or is offered on a CD-ROM that is added to the book (and can, of course, be released also on its own, regardless of the print edition).

In order to produce this application for each title in a semi-automated mode, we developed a generic XML dictionary 'shell' that can absorb any lexicographically structured data in XML format and process it into an electronic dictionary application for PC according to its own configuration. Its main features are:

- dual display of both dictionary sections on the same screen, and full linkage between the two with easy transfer from one to the other;
- various search options in either language, including advanced, wildcard and soundex searches, on different entry components;
- full hyperlink of the words used in the dictionary to their appropriate entries in either section, including headword inflections;
- hyperlinking items in illustrations and words in supplements to their relevant dictionary entries;
- compatibility with other desktop applications, including a hotkey for direct connection;
- back/forward paging that keeps track of all the entries that were previously loaded, which may be erased and restarted at any point;
- audio pronunciation, including self-recording;
- adapting the part of speech and all other labels, as well as the help guides and installation instructions, to the user's native language;
- a skin engine enabling users to transform the visual aspects of the interface to their liking;
- loading an unlimited number of dictionaries concurrently, and selecting which languages to work with – whether for L1 or L2.

Figure 12 reproduces a screenshot from the dictionary application for French learners of Japanese, where the French translation of the second sense of the Japanese entry was doubleclicked and opened in the French-Japanese dictionary section appearing below it.

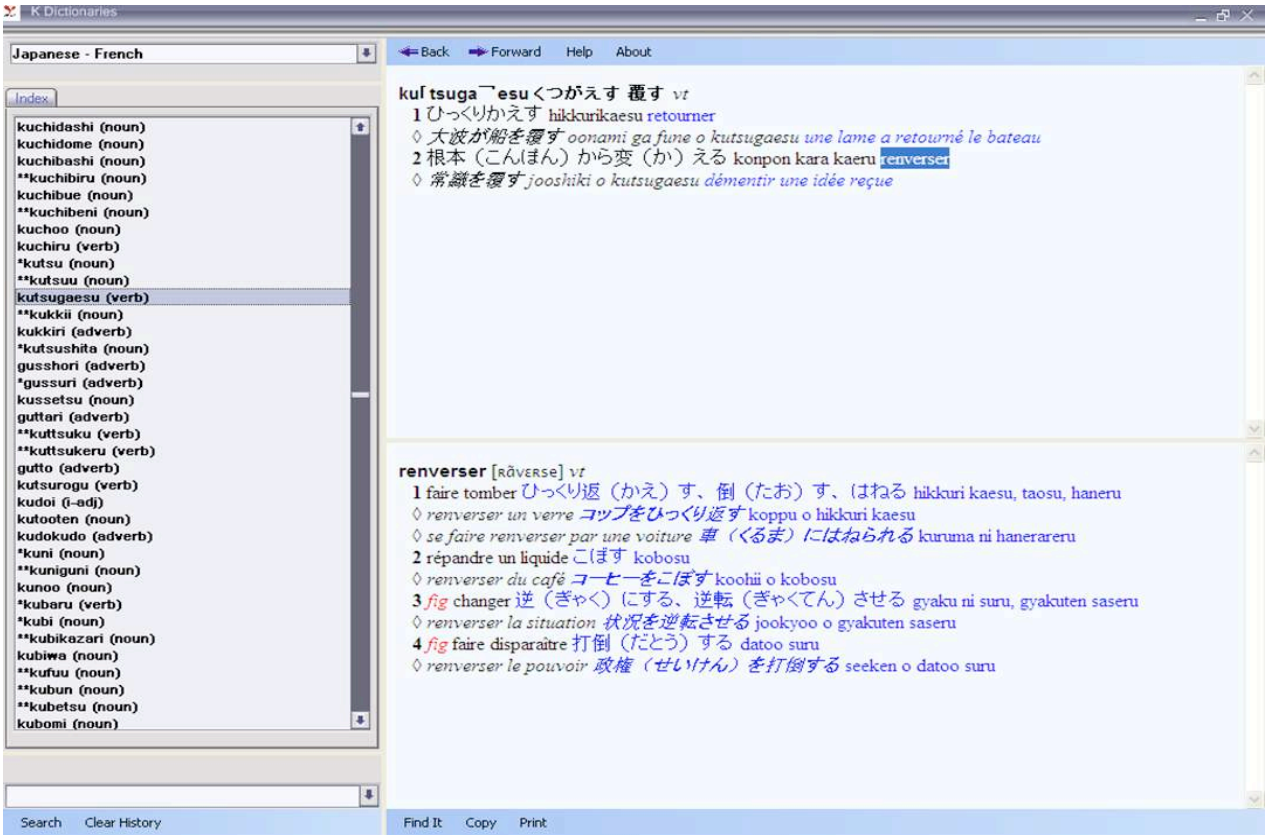

Figure 12: A screenshot from the desktop application of the Japanese/French dictionary

4.4 Online Application

Over the last couple of years we developed a test-site for online dictionary applications. The initial purpose is to investigate how to offer dictionaries online and how features of online dictionaries might affect the actual compilation and use of lexicographic content. The site also serves to demonstrate online apps to our partners. Since KD is basically a B2B company, and does not present its products and services directly to the actual end-users, this site is not targeting the general public.

As with the desktop application, the XML structure of the data enables its fairly straightforward application for online use as well. Figure 13 shows a screenshot from a subset of the main site that is dedicated to a series of Norwegian bilingual dictionaries, where it is possible to easily switch among languages, hyperlink words to their entries, etc.

4.5 Other Applications

KD does not create other electronic applications on its own, but cooperates with a wide range of technology partners for this purpose. The same XML data of our dictionaries is then used just as easily to develop versions for their own desktop and online usages, as well as for smartphones, tablets, handheld devices and any other form of digital media.

Figure 13: A screenshot of a site for Norwegian bilingual dictionaries

5. Conclusion

Developing our dictionaries as an extensive database rather than as specific products demands a considerably higher initial investment over a considerably longer period of time on the one hand, but the consequences include many more potential by-products over a much longer term on the other hand.

In addition to continuing to solidify and enrich the content, we plan to further extend and exploit its database applications particularly as a base for semi-automatic development of new content – such as by data reversal, linking translations to their own language entries or combining several languages together – and to enhance its integration with various corpora and related applications.

6. References

COBUILD. *Collins COBUILD English Language Dictionary*. Glasgow: Collins. 1987.

DAK. *Dictionnaires Assimil Kernerman. arabe, chinois, grec, japonais, polonais, portugais, russe, turc*. Paris: Assimil. 2009.

Italiensk ordbok. Italiensk-norsk / Norsk-Italiensk. Oslo: Vega Forlag. 2010.

Spansk ordbok. Spansk-norsk / Norsk-spansk. Oslo: Vega Forlag. 2008.