Reverso Context: Redefining dictionaries and language tools

Théo Hoffenberg

Théo Hoffenberg is CEO of Softissimo Inc. He is an engineer by training (graduate of École Polytechnique in Paris), an entrepreneur by profession (founder of Reverso among others), a designer and architect of NLP tools and content by passion, and an amateur linguist.
theo@reverso.com

The last issue of this publication had a brief article by Ilan Kernerman entitled Dictionary *n. Obsolete*? (KDN21, 2013). Although the title is certainly provocative, it's quite clear that today's dictionaries are not exactly what they used to be. In the same issue Colin McIntosh wrote about the new definition for 'book' in *Cambridge Advanced Learner's Dictionary*, which focused more on the content than the physical form we used to associate it with (ibid.).

Combining the two approaches, we could say the dictionary as we knew it (i.e. a book consisting of a list of entries indexed in alphabetical order, containing definitions or translations, examples of usage, compounds, etc) is probably already obsolete. Nevertheless, we are likely to still use the word 'dictionary' to refer to a very different concept, just as we still use 'telephone' for something that no longer resembles the large contraption with a rotary dial it represented 30 or 40 years ago.

Soon the word 'dictionary' will likely refer to a tool that helps us find the most appropriate choice in a certain context, offering users easy access to meanings and translations of words and phrases, along with relevant examples of usage, etc.

Within this framework, Reverso presents a new approach to translation aids that in a few years might be understood as a dictionary, but for the time being can have different names: a new type of example-based dictionary; a bilingual concordancer; a search engine for large bilingual texts (*bitexts* in NLP jargon) aligned at word and phrase level; a bilingual aligner providing translation for relatively frequent sequences of words; a provider of frequent wording suggestions and their translations; an analyzing tool that applies linguistic principles to big data; a terminology checker based on balanced corpora.

These descriptions may seem intimidating at first, and may even bring to mind a Rube Goldberg machine, or a white elephant. However, in practice, novices and experts alike find this approach efficient and easy to use. Therefore we call it simply Reverso Context (RC). While the text below offers some insight on how our idea of a 'dictionary' works, readers are invited to also experience it firsthand in order to fully appreciate the innovative features of this linguistic tool.

Changes

Dictionaries of previous generations had various limitations. Firstly, the total number of characters must fit an acceptable volume (say below 2,000 pages for the large ones). Thus, reducing the content size and eliminating redundancy and inconsistencies soon became a huge task that required enormous work from authors and editors looking to fit comprehensive data within restricted space. Secondly, the print was in black and white, occasionally including only one additional color.

The use of such dictionaries demanded the reader's active participation to interpret signs or symbols (e.g. ~ to replace the headword, -> for cross-reference), as well as to cross-reference the indicators suggested (subject, object type, preposition use, etc) with the actual context in order to choose the most appropriate meaning. Cross-checking in the opposite direction was also in the hands of the user. Moreover, looking up phrases such as *not at all*, *je m'en vais* or *pas tout de suite* often proved to be a difficult task.

At that time, there were no search engines, and intensive users of foreign languages were fewer (though perhaps more motivated). Nowadays, many people are used to search engines and machine translation, hence some laziness or higher expectations on their part. When searching for appropriate vocabulary, modern users expect answers to be instant, precise and varied.

Users are also increasingly used to ask questions in whichever way they come to mind, without rephrasing or adapting to query syntax, and still be able to obtain relevant answers.

In 2000, based on comprehensive dictionaries from Collins, we made a big step by putting computer power into use to enhance the user experience for dictionary look-up: no more ~ to replace the headword or other such abbreviations; use of color to identify components of an entry (blue for source language, black for target language, green for domain indicators, red for grammar, etc); direct access to compounds; and full-text search to find examples in both directions (for example, *faire miroiter* appears as the translation for *dangle* even though *dangle* is not given as a translation for *faire miroiter*). This feature was initially implemented in our dedicated software environment called *Lexibase*, which included Collins bilingual and monolingual dictionaries. The software has been updated and is still in use today, available for Internet, intranet and PC, and the same environment has been applied to many more dictionaries since then.

Despite its extensiveness, the content itself was originally designed with the intention of producing a book. This means that the variety of the examples and the coverage of derivations, among other components, were limited, and focus was put more on avoiding redundancy rather than on expanding coverage.

Dictionaries of this type were not only limited, but also extremely costly to develop, because of the 100% human factor and the strict editorial rules and compactness. As a result, most dictionaries for non-major language pairs, such as French-Arabic or French-Japanese, never reached the comprehensiveness of those for English-Spanish, for example.

But even the largest dictionary content originally designed for a print edition cannot provide the full coverage expected today in terms of examples, derivatives or context, let alone up-to-date vocabulary and technical terms.

Examples

Let's take some concrete examples of the benefits of this advanced language tool that allows users to communicate in languages that are not their mother tongue.

Suppose a French-speaking person wants to translate *je m'en vais* into English. Wouldn't it be nice to type this text in an entry box, and get translation suggestions including examples that use both the searched item and the suggested results in context? What if the same possibilities existed for a Spanish speaker looking to translate *me voy*? This is precisely what RC is about.

Users these days are accustomed to getting relevant results in a blink of an eye and effortlessly. In this sense, RC caters to "pampered" users that no longer wish to lemmatize such phrases. After all, knowing that *je m'en vais* stems from *s'en aller* isn't obvious, and searching through the sub-entries of *aller* for the verb's pronominal or reflexive forms can be tiresome. Additionally, if you happen to be a linguist, you know that important information is often lost through lemmatization, as not all verb forms relate to the original meaning of the root.

This same example can be observed from another viewpoint. Although native speakers' intuition allows them to know that *je m'en vais* can have very different meanings according to context, new learners or even proficient non-native speakers may find it difficult to grasp the different nuances of this phrase. In fact, the tone of this expression can range from neutral to aggressive and threatening, and its meaning varies when it precedes a verb, in which case it expresses a will to take action.

Looking up *me voy* in the large Collins English-Spanish dictionary, for example, one might automatically switch to full-text search to display relevant examples containing this text, but still not find the direct translation of the phrase itself. In addition, RC offers more than 8,000 short texts containing the item, of which over 1,000 are aligned to *I'm going* and 400 to *I'm leaving* and *I'm off*.

RC is also particularly useful for finding examples of usage and translation of phrases that cannot be translated independently. Take for example the phrase *shy of* + number or quantity, which can be translated as *un peu moins de*, meaning that a certain amount is less than expected. The examples enable users to find the most suitable expression for each particular situation. The same applies for other words, such as *sorted*, *get sorted*, *get things sorted*, *get myself sorted*, etc.

For the linguist, RC offers more interesting features, allowing to identify trends or validate theories and lexicons. It responds to questions such as: What are the most frequent translations for this word or phrase? Which frequently used phrases contain this word? What does this word translate into when not in this phrase?

Taking an example, a quick look-up of *upside* in RC will show that most examples of usage are related to *upside down*. A more advanced search provides translations excluding the phrase *upside down*. Then, if a certain phrase (e.g. *upside risks*) is too widely represented, it can be excluded from the search. Alternatively, simply looking up *an upside* will provide translations of *upside* as a noun.

When translating *siège* from French to English, words such as *seat*, *siege* and *headquarters* may come to mind. Although one may think that *seat* is the most generic translation and that *headquarters* is used mainly as part of the compound *siège social*, a search with RC would show that *headquarters* is by far the most commonly used equivalent. Moreover, the "-" option can be applied in the RC (*siège* -{*siège social*}) to check if *siège* is translated into *headquarters* even when it is not part of *siège social*.

Non-natives who are proficient in a foreign language often need assistance

Reverso is a leading multiple-language portal offering online translation, definition, spelling, grammar, conjugation and more to over 10 million regular users worldwide. Reverso–Softissimo provides customized solutions including products and services for corporate clients.
http://reverso.net/
http://context.reverso.net/
http://reverso.softissimo.com/

to validate their choice of words. This process of producing a coherent translation is what linguists describe as *encoding*, for *production* purposes.

One way to do it is to use online dictionaries, starting with bilingual dictionaries from the source language into the target language and then searching the definition or synonyms in the target language to find the most appropriate one in context. Another option would be to look up the definition or synonym of the assumed translation and see whether it is appropriate, also using the reverse translation (translating back into the source language).

For example, to translate *acharné* a French-English dictionary would provide equivalents such as *fierce*, *bitter*, *relentless* and *unremitting* as first proposals, with only a few examples. To search further, one could look for synonyms for *relentless* and find *ruthless*, *unrelenting* and *uncompromising*. However, in order to find the best translation, one should have a near-native level of English, or at least read the definitions or the "back translations".

Moreover, if one were to use this adjective as part of *travail acharné*, results could be surprising. The RC search shows that it is widely used, and that its translation is *hard work*, although *hard* is not among the proposed translations or synonyms for *acharné*.

If the *relentless* translation is chosen, one could check sentences containing the first one in the source language, and the second one in the target language, finding more than ten relevant examples.

Conclusion

Bilingual concordancers have already proven to answer certain requirements that dictionaries could not meet. This explains why their use has spread in recent years and why RC could profit from being easy-to-handle by average users though providing more advanced features. RC will continue to innovate and push towards the dictionary of the future. First, by improving and diversifying content resources, adding new and varied corpora that encompass diverse fields and language levels, and expanding coverage to both written and spoken language. Second, by introducing new ways to customize the user's search experience. For example, large organizations that have voluminous corpora are already able to prioritize their content with more pertinent features, and subject domain, regional variants and the language level will also be possible to filter in the future. Last but not least, RC will strive to maintain high quality when dealing with large data volumes thanks to automatic cleaning scripts as well as processing user feedback. With this, we hope to be to the dictionary what a smartphone is to the old telephone today.

Reverso is adding the following KD titles to its services in 2014:

- monolingual dictionaries for English, French | German | Spanish
- bilingual, bi-directional French dictionaries for Arabic | Dutch | Hebrew | Portuguese Brazil and Portugal | Russian
- phrasal French/Hebrew dictionaries

Published in collaboration with K DICTIONARIES

An introduction to iFinger and Clarify Language Service

Knut Haga

iFinger is a provider of digital dictionaries integrated in Microsoft Windows environment. The first version of the iFinger software was released in 2000, with the main goal of offering convenient look-up solutions in high-quality dictionaries. The portfolio varies from glossaries to unabridged monolingual and bilingual dictionaries from HarperCollins, Merriam-Webster, K Dictionaries, Pons and Cappelan Damm, as well as iFinger's own terminology for the medical, technical and legal domains.

Since its inception, over 3.5 million users have accessed this dictionary software from CNET's Download.com. It enables tailored application for multiple users in the corporate, educational and governmental sectors. At present iFinger has more than 200,000 users in the educational sector in Norway and more than 50,000 users in corporate and governmental sector.

Overall, the demand for language services is growing constantly. iFinger aims to meet this demand by developing new cloud-based language tools that will be available as native solutions for all common operating systems. This service is branded as Clarify and is offered for free to the general public, with premium content available through annual subscription to corporate and government users.

The initial launch in June 2014 offers a free dictionary service for 30 languages, including 670 dictionaries covering the languages of 3.4 billon people. There are mobile apps for iOS, to be followed soon by apps for Android and Windows Mobile, as well as by new services for Machine Translation and Text To Speech.

This entire activity is being transferred from iFinger to the new Clarify service, which is designed for unlimited global growth.

http://clarifylanguage.com
http://ifinger.com