A Bilingual Corpus for Lexicographers

Sabine Citron & Thomas Widmann
Sabine.Citron@harpercollins.co.uk
Thomas.Widmann@harpercollins.co.uk
HarperCollins Publishers
Dictionaries Division
GLASGOW
GB-G64 2QT

Abstract
What can a parallel corpus bring to lexicography? How can a parallel corpus be tailored to the needs of lexicographers? This paper addresses both these issues on the basis of a real bilingual corpus compiled at HarperCollins Publishers. Unique in its kind, this corpus comes with a set of tools: an aligner designed to provide perfect matching between L1 and L2, and a concordancer, the GUI enabling lexicographers to access and view the findings of the aligned corpus. Initial results are surprisingly good, and show that pairs of matched bilingual sources are as much of a step forward for bilingual lexicography as corpora were when they were first introduced into the dictionary-making process.

1 Introduction

Much existing research on the subject of parallel, or bilingual, corpora for lexicography has dealt with finding translations automatically.[1] We at Collins already have good lists of translated terminology in the form of bilingual dictionaries. We were therefore looking for a novel way of using high-quality parallel corpora to improve tried-and-tested dictionary translations, especially those for users at the top-of-the-range end of the market. We set out to test this hypothesis by building a corpus of contemporary English and French.

2 Data hunting and data gathering

The first challenge we faced was in obtaining suitable data. Translated texts in electronic format are very common. Many are freely available. There may even be parallel texts to be exploited on the Web. When we started looking at ready-made bilingual corpora, however, we found that they were not what we were looking for.

Many of the existing banks of translated texts are from large international organizations such as the UN and its affiliates, or from the EU. There are also the famous proceedings of the Canadian Parliament, better known as Hansard. The undoubted advantage of these

[1] Véronis, ed. (2000), Mihalcea & Simard (2005)

sources is their vast size. Their weakness is that they contain very much a specific type of language, perfectly correct, but not always entirely representative of real language. What we try to analyse and describe in our dictionaries, especially for the sophisticated users of top-of-the-range dictionaries, especially for their encoding needs, is the language people will encounter and want to reproduce in oral or written form in their daily life and in a business environment, rather than UN-speak, for example.

Moreover, it is also surprisingly difficult to identify the real source language of many of the texts generated by international organisations. These texts are sometimes written by non-native speakers and it is not always easy to clearly identify which was the original in a set of matching texts.

Literary texts are an obviously suitable source for lexicographical needs: along with newspapers and magazines, novels tend to be the written source closest to real language. Here we met our second challenge: translated newspapers are not very common, and permissions were an issue. As for literary texts, copyright laws only allow us to use them without express permission if they are from the earlier part of the 20th century. But language has changed too much since the days of Proust and Jules Verne, and our dictionaries aim to reflect contemporary language. We therefore decided to go through the proper channels, and to seek permission from publishers to use their texts. Being part of a large publishing house ourselves, we only had half a battle on our hands: we looked for HarperCollins titles translated into French, and for HarperCollins translations of French books.

We will skip over the difficulty in getting permissions from French publishers to use their texts in this age of electronic data manipulation. To cut a long story short, we are now the proud owners of a French-English corpus of over 2 million words, potentially 3 million.[2] This is a sizeable bilingual corpus, and one of the largest of its kind. It is still tiny, especially compared to the monolingual corpora we are used to working with at Collins, which total around 640 million words for the Bank of English alone, 940 million with English variants and foreign languages. We are all agreed that size matters where corpora are concerned. But our parallel corpus is a good pilot to test what could be done with a larger body of bilingual text.

3 Alman, the manual aligner

Much work has gone into writing sentence-alignment programs,[3] but when we made some experiments on literary samples, we got poor results. The closest in functionality to our requirements seems to be TRADOS' WinAlign but its aim is different. Most sentence-alignment tools have been designed for non-literary texts, where one can assume that sentences will be a literal translation of the original, and that there will be a rough correspondence in sentence length between L1 and L2. Also, in such texts, there will typically be a large percentage of proper nouns and loan words that can be used to aid the alignment. However, good literary translations are by essence different from the original, and include few words

[2] Bilingual corpora are measured in total number of words of both languages added together.

[3] For example the Vanilla Aligner: see Danielsson & Ridings (1997).

that will uniquely link two sentences. After spending some time researching this area, we concluded that it would be best to write a tool to manually correct automatic alignments. Some experiments showed us that it would be worth focusing on correction rather than alignment, so over time, this tool ended up as a manual aligner which we called Alman.

Alman is a fairly basic tool. It will take two files, display them side-by-side, suggest an alignment and wait for user input. The algorithm for suggesting an alignment is very simple in that it will just search forward for the first sentence-boundary character. The suggested alignment is then underlined, and the user can accept it by pressing the *return* key. The user can also correct the underlined selection by moving backwards or forwards by a character, a word or a sentence. If a section of text is found only on one language, it can be ignored. If an error is found (this is likely if the text was obtained by OCR), it can be corrected: Alman will launch an external editor to let the user edit the underlined text. Alman also provides unlimited undo functionality. All commands are available through single keystrokes, which speeds up the process considerably.

It might sound very time-consuming to align texts manually, but we found that an experienced translator can align 20,000 words in an hour. While this is obviously longer than the time spent running an automatic aligner, it is less time than we would have spent correcting erroneous alignments produced by a non-specialised tool. And of course the quality of the resulting aligned bilingual corpus is much higher – perfect alignment really is essential for a sophisticated and detailed analysis of comparative patterns of behaviour of words or multiword units. Good quality not only improves subsequent results, but it also means we effectively are able to use more of the texts.

Alman is written in Perl and has currently only been tested on Unix. When the alignment is completed, it is exported to an XML-like format.

4 Concordancing

We experimented with several parallel concordancers. The closest to our needs was the Stuttgart Corpus Workbench (CWB). Its display was however not tailored to our needs in the sense that it truncates citations and displays all L1 citations in a block together, followed by all corresponding L2 citations in another block of text, making individual matching difficult. We therefore wrote our own simple concordancer in Perl. This finds matching strings in either source or target in the XML-like format described above and displays the aligned chunks, highlighting the string. The concordancer allows for wildcard searches. It also does negative searches (to exclude obvious translations). It does not yet include a lemmatiser: this is the next obvious improvement.

5 Our corpus

At the end of all this, what do we have? We have a unique parallel corpus of a reasonable size, with excellent quality of content and output, made of contemporary material and fully aligned. As far as we know, no other such corpus is available today. Our corpus is arguably the largest parallel corpus of contemporary fiction, the largest parallel corpus without alignment errors, and one of the few parallel corpora with consistent *L1* identification.

Our hypothesis is that, just like monolingual corpora replaced lexicographers' intuitions

(not lexicographers!) for the compilation of dictionary framework, parallel corpora will usefully supplement their intuitions in the search for good L2 equivalents

6 Results

Of course a corpus is only a means to an end. We will show two representative examples (for want of more space) to demonstrate how our corpus can be used to improve translations in practice.

Let us take a fairly common French word, the adverb *jadis*. Oxford-Hachette translates this as *formerly, in the past*. Harrap's *Shorter* has *in times past, formerly*. Collins-Robert has *in times past, formerly, long ago*. These are all fine translations. However, if we look at the parallel corpus citations for this word, the English word *once* appears as the dominant translation for the French *jadis*. A few examples:

"I think," he pronounced, gloomily, "that our kind, we like the cigarettes so much because they remind us of the offerings that **once** they burned for us, the smoke rising up as they sought our approval or our favor."

• «Je crois que si on aime tellement les clopes, nous autres, c'est parce qu'elles nous rappellent les offrandes qu'on brûlait pour nous, **jadis**, dit-il. La fumée qui s'élevait vers nous quand on cherchait notre approbation ou notre faveur.

Saint Bride, who was **once** Bridget of the two sisters (each of the three was a Brigid, each was the same woman).

• sainte Bride, qui était **jadis** Bridget aux deux soeurs (toutes les trois s'appelaient Brigid, et toutes les trois n'en formaient qu'une) ...

A **once**-famous comedian, believed to have died in the 1920s, climbed out of his rusting car and proceeded to remove his clothing: his legs were goat legs, and his tail was short and goatish.

• Un comique **jadis** célèbre, qu'on croyait disparu dans les années 20, descendit de sa voiture rouillée et entreprit d'ôter ses vêtements: il avait des pieds de bouc et une courte queue caprine.

The tip of the point had collapsed over the years and a tree which had **once** stood upright there now grew outwards at an angle of forty-five degrees, its trunk stripped bare by the elements and only a small fuzz of green left at its tip.

• L'extrémité de la corniche s'était éboulée au fil des ans et un arbre qui s'était **jadis** dressé poussait désormais à un angle de quarante-cinq degrés, son tronc dénudé par les éléments, avec seulement un petit duvet vert autour de sa pointe.

Once, billions of years ago, they were bacteria; free-living and vicious they entered other cells and multiplied like viruses until they split their host asunder, so releasing multiple offspring into the world.

• **Jadis**, il y a des milliards d'années, c'étaient des bactéries; autonomes et vicieuses, elles

pénétrèrent d'autres cellules et se multiplièrent comme des virus jusqu'à ce qu'elles mettent leur hôte en pièces, libérant ainsi de nombreux rejetons de par le monde.

Of the 18 occurrences of *jadis* in our parallel corpus, 11 point to *once* as being the translation. (the other citations are *a long time (ago)*, *old days*, *of yore*, *ancient* and *formerly*). The English word *once* now seems to be an obvious and uncontroversial translation solution. It may seem obvious in retrospect, but this translation was not in the trusted dictionaries we consulted, including our own.

We also tested the parallel corpus on multiword units. Let us take the example of *faire la fête*, again a very common expression (and practice). Collins-Robert and Harrap's *Shorter* give the perfectly acceptable translations *to live it up, to have a wild time*. Oxford-Hachette opt for the first of these, *to live it up*. Translations appearing in our parallel corpus for *faire la fête* are *to celebrate, to have a party* and *to party* – much better and more likely than *to live it up* and *to have a wild time*.

One might have noticed that in our first example, *jadis* is actually in the source, not the translation. The same applies to *faire la fête*. In both cases the lexeme we wanted to translate was found in the L2 part of the corpus, not in L1. While this might appear counter-intuitive, it is our experience that this gives us better translations. Even the most experienced of translators are influenced by word choice (and sentence structure) of the text they are translating. When translators translate words and sentences, they rely on their own intuitions and on dictionaries (in turn based on lexicographers' intuitions). By turning the L1/L2 combination around, catching writers unawares, and using L1 data to improve translations, we are replacing intuitions with real language. Our technique makes it possible to find translations within naturally occurring language.

7 Conclusion

Our 2-million-word parallel corpus already allows improvements even to well-polished dictionary entries, and even to common words. The larger the parallel corpus grows, the more accurate its results will become in terms of frequency and range of coverage.

References

A. Dictionaries

(2005) *Collins Robert French Dictionary* aka *Le Robert & Collins Senior*. (Seventh edition.) Glasgow & Paris, HarperCollins Publishers & Dictionnaires Le Robert.

(2001) *The Oxford-Hachette French Dictionary* aka *Le Grand Dictionnaire Hachette-Oxford*. (Third edition) Oxford & Paris, Oxford University Press & Hachette.

(2004) *Harrap's Shorter Dictionary*. (Seventh edition). Edinburgh, Chambers Harrap Publishers Ltd.

B. Other literature

Danielsson, P., Ridings, D. (1997), *Practical Presentation of a 'Vanilla' aligner*, in Reyle, U., Rohrer, C. (eds), *The TELRI Workshop on Alignment and Exploitation of Texts*. Ljubljana, Institute Jozef Stefan.

Mihalcea, R., Simard, M. (2005), *Parallel texts*. In *Natural Language Engineering*. 11 (3). Cambridge, Cambridge University Press, p. 239-246.

Véronis, J. (ed.) (2000), *Parallel Text Processing. Alignment and Use of Translation Corpora*. Dordrecht/Boston/London, Kluwer Academic Publishers.