Roda P. Roberts & Catherine Montgomery, University of Ottawa

The Use of Corpora in Bilingual Lexicography

Abstract

The publication in 1994 of the first bilingual dictionary to use corpus material, the Oxford-Hachette English-French, French-English Dictionary (OXHA), marked the start of a new age of corpus-based bilingual lexicography. OXHA's innovation is being taken a few steps further in the Bilingual Canadian Dictionary (BCD) project. In addition to using two unilingual corpora – one in English, the other in French, as does the OXHA team, the BCD also uses a bilingual, translated corpus. While the OXHA team only used corpora at the editing stage, the BCD team uses corpora throughout the dictionary-making process. This paper will present the corpora used by the BCD, the various stages at which they are used, and the way they are used.

1. Introduction

Although the use of corpus material has become relatively well-established in unilingual lexicography since the publication of the *Collins Cobuild English Language Dictionary* in 1987, it was not until 1994 that the first bilingual corpus-based dictionary made its appearance. The Oxford-Hachette English-French, French-English Dictionary (OXHA) is indeed innovative as B.T.S. Atkins claims (1994:xix), in that it uses two unilingual corpora, one in English, the other in French, to ensure that the dictionary represents English and French as they are used not only by the lexicographers, but by hundreds of native speakers. Atkins (1994:xxiv) rightly predicted that OXHA would "mark the start of a new age of corpus-based bilingual lexicography". OXHA'S innovative use of corpus material is being taken a few steps further in the Bilingual Canadian Dictionary (BCD) project, which has been using electronic text corpora since 1990. This paper will present the corpora used by the BCD, the various stages at which they are used, and the way they are used, highlighting differences with the OXHA project where possible. Concrete examples will be given to illustrate the contribution that corpus use makes in the preparation of a bilingual dictionary.

2. The BCD Corpora

The advantages of using electronic text corpora in lexicography had already been well-documented when the BCD project was launched in 1988 (e.g. Renouf 1986, Sinclair 1985, Sinclair 1987). However, the BCD has an additional reason to use corpus material: given its stated goal to present English and French as used in Canada and given the fact that there are few, if any, reasonably exhaustive reference works dealing with either Canadian English or Canadian French, corpus evidence becomes crucial in the dictionary-making process. Hence, our two textual databases contain primarily Canadian texts.

The first textual database, called TEXTUM, contains unilingual corpora in English and French. Presented below is a breakdown of what is presently available.

TEXTUM

Legend: **N** = newspaper; **P** = magazine, journal; **F** = Fiction;
GD = government documents
G = general; **ST** = scientific/technical
CD = Canadian; **US** = American; FR = France

ENGLISH	SIZE (in millions of words)	FRENCH	SIZE (in millions of words)
Gazette **(N, G, CD)**	6,7	Presse canadienne - française **(N+P, G, CD)**	77,0
Canadian Press **(N+P, G, CD)**	129,0	Leméac **(F, G, CD)**	0,9
Canadian Geographic **(P, G, CD)**	0,3	ACFAS **(P, ST, CD)**	0,13
Queen's **(N+P+F, G, CD)**	5,0	Le Monde **(N, G, FR)**	17,1
Department of Energy **(GD, ST, US)**	27,2	Ouest France **(N, G, FR)**	4,9
Wall Street Journal **(N, G+ST, US)**	41,8		

TEXTUM is queried using the concordance-generating program PAT, which is well-suited for lexicographic purposes. It enables users to get KWIC contexts 64 characters long, or larger contexts. It allows users to find significant patterns that occur after the queried word (but not before). Although PAT only finds exact matches, you can use wild cards or truncated words to find all contexts containing a given string. You can also narrow the search by specifying that you want a word to be near or followed by another word or by eliminating patterns that you do not want. Finally, PAT allows you to intersect the results of two queries.

Our second database, called TransBase, contains translated bilingual texts which are aligned. At the moment, the content is limited to three years of the Hansard, the journal of debates in the House of Commons. The speeches have been translated (mainly from English into French, but also from French into English) by very seasoned translators.[1]

TransBase is queried using TransSearch. The result is pairs of aligned sentences that answer the conditions imposed by the researcher. For example, you could ask for a given English word that is translated by a given French word, or conversely for a given English word that is not translated by a given French word. You can use the settings options to specify the source language (either French or English or both) or to specify that capital letters be pertinent. (Langlois 1995.)

Our two databases are supplemented by English and French texts on CD-ROM or on our LAN server (our reserve corpora), which include some British texts. We exploit them using either the search engines provided with the disks or the concordancer MicroConcord.

All in all therefore, we have a wide range of textual material from which to draw lexicographic evidence.

3. Stages at which different corpora are used

The corpora are being used at three different stages of the dictionary-making process: establishment of the nomenclature, entry preparation and revision (editing).

The nomenclature is established using not only existing dictionaries, but also an alphabetical list of words in TransBase, accompanied by their frequency. In addition, lexical items that are not found in at least three of our base dictionaries[2] are verified for the frequency of their use in TEXTUM.

All lexicographers, who have access to TEXTUM from their workstations, use both the English and French corpora during entry preparation. They use the source language corpora (e.g. English, when preparing an English-French entry) for finalizing the framework of the entry, which has been drafted on the basis of dictionaries and other reference works, and for selecting free combinations, collocations, compounds and fixed expressions for inclusion in the entry. They use the target language corpora (e.g. French, when preparing an English-French entry) at the translation stage to verify equivalents proposed by bilingual dictionaries or by themselves. TransBase, our translated corpus, is used at the end of the translation stage to ensure that no good equivalents have been missed. Since, during entry preparation, corpus use is limited to verification and additions, lexicographers are restricted in the number of corpus examples they are allowed to use.

At the revision stage, revisors look over all the material used by lexicographers in the preparation of their entries, including the corpus examples analyzed. If they deem it necessary, they do a further corpus search (both in the two databases and in the reserve corpora) to confirm certain items.

4. How corpora are used in entry preparation

Since the use of corpora for entry preparation is more innovative than their use for editing purposes, the former will be the focus of attention below.

4.1 Preparing the source language framework for the entry

After assembling basic documentation from unilingual and bilingual dictionaries for a given headword and preparing an initial sense chart based on senses given in dictionaries, the lexicographer consults TEXTUM in the source language and prints out a selection of one-line contexts. Preliminary consultation is limited to a total of 100 concordance lines per word. Following the analysis of one-line contexts, longer contexts can be obtained and more specific searches can be undertaken as needed. For source language research, TEXTUM is used for the following elements of the entry.

4.1.1 Determining the spelling of the headword

For words with spelling variants, TEXTUM serves as the guide to which of the spelling variants is the primary form used in Canada; the most frequently used form is placed first and other variants follow. For instance, TEXTUM was used to choose between *waistband*, *waist-band* and *waist band* in English, and *bibitte* and *bébitte* in French.

4.1.2 Determining and ordering senses of the headword

The BCD lists major senses of a word in separate sense divisions. While the starting point for establishing senses is unilingual dictionaries, TEXTUM is used to determine whether senses given in dictionaries are actually used in Canada; for example, *bluff* in the sense of a person who bluffs given by several unilingual dictionaries had no occurrences in TEXTUM. Similarly, when a lexicographer is not familiar with a given sense of a word, its appearance in TEXTUM serves to confirm its use. We obviously have to bear in mind that less common senses or words may not always be found in TEXTUM.

An analysis of one-line contexts from TEXTUM may also reveal that the word is used in a sense or nuance not accounted for in dictionaries. If TEXTUM provides enough evidence to justify it, another sense division will be added. This was the case for *aîné* in the sense of elder of a tribe, and *scrum* in the sense of an impromptu press conference.

If it appears that a word or sense is a Canadianism, lexicographers can consult the *Le Monde* and *Ouest-France* corpora in TEXTUM to verify whether the word is restricted to Canadian French or is also used in France. In the former case, the headword is labelled *(CD)*. For English, lexicographers can search the American *Wall Street Journal* in TEXTUM or the British *MicroConcord Corpus Collections* in our reserve corpora to determine whether a word designated a Canadianism is also used in the U.S. and/or Great Britain.

Finally, TEXTUM is a guide to ordering sense divisions of a polysemous headword. The BCD lists more common usage before less common usage and TEXTUM serves as a source of information on frequency of usage of the various senses. Corpus data reveals, for example, that *snowbird* is more frequently used in the sense of a person who spends the winter in a warmer climate than as the common name of the bird called a junco.

4.1.3 Grammatical analysis of the headword

The grammatical category of a word form is identified on the basis of both information found in dictionaries and examples of usage. Occasionally, there are "contradictions" between different dictionaries and between dictionary information and corpus examples regarding grammatical category. Actual usage as revealed in TEXTUM is analysed to determine the grammatical categories of the headword to be included in the BCD. For instance, *unifamilial(e)* is listed as an adjective in two French dictionaries, as a noun in another and as both a noun and adjective in two other dictionaries; TEXTUM shows that it is used as both a noun and an adjective, with noun usage predominating.

4.1.4 Structural analysis of the headword

Corpus analysis reveals how the headword functions in sentences and the common structures in which it is found. For each sense division of the headword, the BCD normally includes a free combination where the headword is used without any special syntactic or semantic restraints. TEXTUM serves as the preferred source of free combinations which may add grammatical, structural or stylistic information.

The BCD gives particular attention to collocations, compounds and fixed expressions which are essential to idiomatic translation. TEXTUM serves to verify the currency in the Canadian context of collocations, compounds and fixed expressions listed in general and specialized dictionaries. Repeated patterns found in TEXTUM can also suggest collocations not included in existing unilingual or bilingual dictionaries. While PAT enables the lexicographer to search only for right-hand collocates (e.g. ***doughnut*** *effect* is aligned under *doughnut* and ***babillard*** *électronique* is aligned under *babillard*), MicroConcord allows a search for left-hand collocates as well.

As the majority of texts featured in TEXTUM are not specialized, TEXTUM is less useful for identifying or confirming technical compounds, for which the term banks and dictionaries remain the primary sources.

4.2 Researching target language equivalents

4.2.1 Finding equivalents

Existing bilingual dictionaries and term banks serve as the usual starting point for identifying equivalents for a given headword. However, Canadian words (e.g. *niaisage*) or senses of a word (e.g. *patenter*) are often not included in existing bilingual dictionaries. In this case, lexicographers must rely on the definitions provided by unilingual dictionaries, their own knowledge of the source and target languages and mental translation of source language contexts taken from TEXTUM to decide on suitable equivalents. Although more limited in scope, TransBase is also a source of equivalents, particularly for Canadian words or senses of a word (e.g. *achalandage*) not found in bilingual dictionaries.

4.2.2 Confirming equivalents

If a lexicographer or revisor has doubts about the use in the Canadian context of an equivalent found in existing bilingual dictionaries, TEXTUM is used to verify whether it is a suitable equivalent. For instance, for the French adjective *patent*, two bilingual dictionaries list *patent* along with *obvious* and *evident* as one of the English equivalents. However, TEXTUM revealed that, in Canadian English, the adjective *patent* is rarely used in the sense of obvious/evident and would therefore not be suitable as a general equivalent. TransBase can also be consulted to confirm whether an equivalent proposed by a lexicographer or revisor for a given headword has been used by translators.

TEXTUM is also used to ascertain the most frequent spelling for an equivalent with spelling variants. This is particularly important since the BCD gives only the most common spelling for an equivalent.

5. Conclusion

Corpora are a reflection of language in use on which a dictionary should be based. There is little doubt that corpus use in the BCD project has enriched the entries prepared so far. Moreover, this project has demonstrated the importance of integrating corpus use into all stages of bilingual dictionary-making.

However, despite the many advantages that corpora present, they must be used with some caution. The lexicographic evidence they provide must be subjected to the sound judgement of lexicographers. And lexicographers must ensure that they do not become overwhelmed by corpus evidence! That is why the BCD project limits the number of corpus examples to be extracted and is developing guidelines for their analysis and use. For corpora, like all lexicographic tools, are a lexicographer's best friend only if they are appropriately used.

Notes

1. While translations are avoided whenever possible in bilingual terminology, their usefulness in bilingual lexicography, which involves a translation phase, must not be ignored.
2. The base dictionaries for the English nomenclature are the *Random House Webster's College Dictionary*, the *Gage Canadian Dictionary*, the *Collins Robert Senior French-English, English-French Dictionary*, and the *Oxford-Hachette French Dictionary*. Base dictionaries for French are the *Lexis*, the *Dictionnaire québécois d'aujourd'hui*, the *Collins Robert Senior French-English, English-French Dictionary*, and the *Oxford-Hachette French Dictionary*.

References

Atkins, B.T.S. 1994. "A Corpus-Based Dictionary", in: *The Oxford-Hachette French Dictionary*. Oxford/Paris: Oxford University Press/ Hachette Livre, pp. xix–xxvi.

Langlois, L. 1995. "Computerized Tools Used in the BCD Project". Unpublished paper.

Renouf, A. J. 1986. "The Exploitation of a Computerized Corpus of English Text", in: M. Rivas (ed.), *Actes du VIIIème Colloque G.E.R.A.S.* Paris: Université de Paris-Dauphine.

Sinclair, J.M. 1985. "Lexicographic Evidence", in: R. Ilson (ed.), *Dictionaries, Lexicography and Language Learning*. Oxford: Pergamon.

Sinclair, J.M. ed. 1987. Looking Up: *An Account of the COBUILD Project in Lexical Computing*. London/Glasgow: Collins.