Alison Dickens & Raphael Salkie
The Language Centre, University of Brighton, England

Comparing Bilingual Dictionaries with a Parallel Corpus[1]

Abstract

This paper investigates how translation equivalents in French-English bilingual dictionaries compare with equivalents found in a parallel corpus of French and English texts. A method for measuring the degree of mismatch is proposed, based on the distinction between BASIC TRANSLATION EQUIVALENCE and RICH TRANSLATION EQUIVALENCE. The method is potentially automatable and is intended to help lexicographers focus their application of parallel corpora in the most useful way. We make some suggestions about how to incorporate rich translation equivalence into bilingual dictionaries.

1. The INTERSECT corpus

Parallel corpora (consisting of texts in one language and their translation into another) are widely seen as important resources in a number of fields (cf. Hartmann 1995). Several projects are now under way to develop and exploit these resources: these range from work on two languages, such as the Norwegian-English corpus at Bergen (cf. Johansson and Hofland 1994) to large scale multilingual projects like EAGLES (cf. Llisterri 1994).

The INTERSECT project is based round a French-English parallel corpus, which we have compiled over the last two years and which is continually being extended. The aims and scope of INTERSECT are set out in Salkie (1995a). We currently have over a million words in each language, covering a variety of written text types and with approximate parity in source texts in the two languages. Our corpus does not claim to be balanced: in particular, scientific and technical texts are under-represented in relation to their importance in the work of translators. Searches and concordances are effected using ParaConc, a Macintosh parallel concordancer developed by Michael Barlow at Houston (cf. Barlow 1995).

Two million words is small by today's standards, and there are also disadvantages in not including spoken language material and in the lack of balance. With a small corpus the obvious thing to do is to focus on

common words. In comparing the corpus with dictionaries this is a logical approach in any case: if the corpus gives some clues about which words occur fairly often, this in itself is useful information for lexicographers. The next question is: which of these frequently occurring words might be worth special attention in a dictionary? The method presented in this paper is intended to answer this question.

2. Rich Translation Equivalence

A central concern in working with parallel corpora is the notion of **rich translation equivalence** (RTE) (cf. Salkie 1995b). This notion is best explained by starting with an example. Ask a bilingual speaker for the French equivalent of *help* and the short answer will probably be *aide* or *aider*. In bilingual dictionaries these are typically the most prominent equivalents of *help*. Dictionaries which are large enough will include some other equivalents as well. Skilled translators, however, have a much richer array of options to translate *help* than the dictionaries suggest. We found 121 instances of *help* in the INTERSECT corpus, of which 60 corresponded to a form of *aide* or *aider* (the figures are skewed by the inclusion in the corpus of a software manual which invariably rendered the "help" facility as "aide". Leaving aside these examples, we had 81 instances of *help* corresponding to 20 instances of *aide* or *aider*). Examples of other translations are:

1) ...the CSTC and its affiliated organisations do not co-operate with the authorities to help investigations into complaints lodged ...
1) ... la CSTC et ses organisations affiliées ne collaborent pas avec les autorités pour faciliter l'enquête sur les faits dénoncés ...

---

2) The government hopes that the easing of external tensions and the working of democratic institutions will help create the circumstances in which bans on trade union activities in certain organisations will serve no useful purpose.
2) Le gouvernement civil espère que l'allégement des tensions externes et que le fonctionnement des institutions démocratiques contribueront à créer les conditions qui rendront inutile l'interdiction des activités syndicales dans certaines organisations.

---

3) As well as a generational effect, perhaps due to the fact that young people today are less integrated into Roman Catholicism or atheism which help to keep the parasciences at bay, there is also a sociopolitical dimension.
3) A cet effet de génération, dû peut-être à une moindre intégration des jeunes au catholicisme ou à l'athéisme qui éloigne les parasciences, s'ajoute une dimension socio-politique.

Similarly, out of 201 instances of AIDE/AIDER (161 ignoring the sofware manual again) in the French texts, only 60 used a form of HELP. Other translations included:

4) L'objectif consiste à procéder à des recherches sur les systèmes d'information qui aideront à l'exécution de la multitude des travaux non routiniers réalisés par l'homme dans l'environnement du bureau.
4) The objective is to carry out systems research on the information systems that will support the wide range of non-routine tasks performed by humans in the office environment.

---

5) ...[Le gouvernement] leur a fourni toutes informations utiles sur la réalité de la situation syndicale en Tunisie afin d'aider à trouver une solution aux problèmes en suspens.
5) ... the Government accepted the missions of the International Labour Office and the ICFTU, furnishing all pertinent information on the trade union situation in Tunisia in the hope of reaching a solution to these problems.

---

6) J'ai travaillé avec l'association Valentin-Hauy, qui consacre ses efforts à aider les aveugles, pour mettre au point des circuits, rédiger des étiquettes et des catalogues en braille.
6) Then, in collaboration with the Valentin-Hauy Association, which works for the blind, I devised routes through the museum and prepared labels and catalogues in Braille.

Dictionaries tend to give only a limited picture of the full range of strategies that skilled translators use. We refer to this limited picture as **basic translation equivalence** (BTE). The wider range of strategies that the corpus shows up is what we call **rich translation equivalence**.

The implications of this richness are many, and we cannot explore them all here (cf. Salkie 1995b for a broader discussion). For lexicography, three key questions emerge:

1. What is the best method for measuring richness ?

2. Which words are most amenable to this method (so that lexicographers can focus their corpus work on these words)?

3. To what extent, and in what manner, should rich translation equivalence be incorporated into dictionaries?

We turn now to a possible method of responding to the first two questions, and then we make some suggestions about the third.

3. A measure of richness in translation equivalence

A lexeme in one language can have more than one equivalent in another for a variety of reasons. The source language lexeme may have several senses, only some of which correspond to the most common equivalent in the target language. The problem is particularly acute when the source language (SL) lexeme has a broad sense and a range of uses which do not match those of any single basic equivalent. If the lexeme is also frequent enough to show up considerable RTE in a small corpus, then it will be a candidate for special treatment in a bilingual dictionary. Let us call lexemes which meet these two criteria *pivot words*.

It would be useful for a bilingual lexicographer to be able to identify pivot words in advance of writing their entries in a dictionary. This would result in more efficient use of the corpus, since items that do not require special treatment could be handled without the need for extensive corpus analysis. Using corpora is time-consuming, and it makes sense to use them only when there is a reasonable likelihood of gaining information without too much effort.

A good place to start looking for pivot words is the literature on translation. Delisle (1993:166) gives a list of thirty words which he describes as "mots à haute fréquence et très polyvalents qui foisonnent dans les textes pragmatiques anglais". Table 1 shows the result of searches for twelve of these words in the INTERSECT corpus.

The "Total" column in table 1 gives the total number of occurrences in the corpus. The "TEs 2+" column reveals how many of the occurrences are translated in the same way more than once, while the "tail" column shows how many translation equivalents occur once only. The final column indicates those cases where no equivalent is discernible, because the translation uses a completely different structure, or avoids using a TE for some other reason (cf. example (3) above).

The results for the last five words in the table rules them out as candidates for special treatment. For *with* there are simply too many occurrences for further investigation to be cost-effective. For *regular* and *facility* there are too few occurrences. For *control* and *type* there is considerable regularity in the way these words are translated, so these words do not look like candidates for special treatment. (A look at table 3 shows, however, that dictionaries cope with the regularity in different ways for these two words).

This leaves seven words which meet the criteria for pivot items: they are reasonably frequent and have a range of translation equivalents, as indicated by the size of the tail in relation to the total number of

occurrences. If we look more closely at *involve*, we see the array of equivalents in table 2.

Table 2 also shows the equivalents given in the entries for INVOLVE in three general-purpose bilingual dictionaries which are in wide use. The equivalents are set out in the order in which they occur in the entry. Using this data for each of the twelve words we constructed a table to compare the corpus findings directly with each dictionary (table 3).

Column 4 of table 3 shows the percentage of corpus examples which use the first translation equivalent given in the dictionary entry. A low figure in this column suggests that unskilled dictionary users who do not look beyond the first translation equivalent may need special help with that word. Column 5 gives the percentage of corpus examples which match any other translation equivalents given in the entry after the first one. If the sum of columns 4 and 5 is low, then the user will probably appreciate a good range of examples of usage, to suggest more RTE. The added value that the examples in fact give is indicated in column 6, and the total match – TE's and examples – is shown in column 7. A low figure in this column suggests that the word is a strong contender for special treatment. The threshold for special treatment should presumably come somewhere between 55.2%, the highest figure for *major*, which probably does need special treatment, and 65%, the lowest for *authority*, which probably does not. Determining a precise workable threshold figure is a matter for further investigation with a larger sample of words.

As large parallel corpora become available in the coming years, a small corpus like ours may still be a useful first tool for lexicographers. The method proposed here can give an indication of those items which can profitably be investigated further using a large corpus.

4. What kind of special treatment?

Now that we have earmarked the pivot items for which some special treatment is likely to be beneficial, what can lexicographers, constrained as they are by space and cost considerations, usefully do with this information?

The least expensive solution is simply to flag the items in question with an indication that the user might wish to consider a richer range of translation equivalents than the ones given in the dictionary. A slightly more generous approach would be to refer the user to a monolingual dictionary or some other reference work which gives a fuller picture of the range of options available. (For the three dictionaries we looked at,

the same publisher issues a monolingual dictionary of at least one of the two languages).

More ambitiously, a box with information about the pivot item could be included near the entry for the item. This might explain what a pivot item is, suggest a variety of strategies for translating the item, and give a fuller range of examples than in a normal entry, with discussion of how users can regard the examples as a springboard to finding a translation in the context in which they need to translate the word. This strategy would take up space, of course, but the point of our method is to isolate the relatively small number of words for which a box would be cost-effective. We are dealing with a few hundred words at most, and if there is not room for all of them in a printed dictionary then the most frequent can be selected. Alternatively, a separate work could be published with the explicit aim of helping advanced users and translators.

Finally in an electronic dictionary the obvious step is to give the user direct access to the parallel concordances that the lexicographer has used. On a CD the amount of space that 500 concordances would take up is small. A dictionary which supplied this access would be easy to market: it could claim to have "distilled the rich array of strategies used by skilled translators and brought them to your desktop".

Note

1. We wish to thank Adam Kilgarriff for his help in writing this paper. He is not responsible for any remaining errors.

References

Barlow, M. 1995. *A guide to ParaConc.* Houston: Athelstan.

Delisle, J. 1993. *La traduction raisonnée.* Ottawa, Les Presses de L'Université d'Ottawa.

Hartmann, R. 1995. Contrastive textology and corpus linguistics: on the value of parallel texts. Paper read at the First International Conference on Contrastive Pragmatics and Semantics, University of Brighton, April 1995.

Johansson, S. and K. Hofland. 1994. Towards an English-Norwegian parallel corpus. In U. Fries, G. Tottie and P. Schneider (eds.), *Creating and using English language corpora*, Amsterdam: Rodopi, pp. 25–37.

Llisterri, J. 1994. EAGLES (Expert Advisory Group on Language Engineering Standards) text corpora working group: introduction. EAGLES document EAG-CWG-IR2. (Available by ftp from

nicolet.ilc.pi.cnr.it [username: eagles; password: eagles] where the file is /corpora/corpintr.ps).

Salkie, R. 1995a. The INTERSECT project at Brighton University. *Computers & Texts* 9 (May 1995), pp. 4–5.

Salkie, R. 1995b. Parallel corpora, translation equivalence and contrastive linguistics. Paper read at the Association for Literary and Linguistic Computing / Association for Computing in the Humanities Joint International Conference, University of California at Santa Barbara, July 1995.

Dictionaries

Collins-Robert French Dictionary: French-English & English-French. (3rd Edition). London: HarperCollins and Paris: Dictionnaires Le Robert, 1993.

Larousse Grand Dictionnaire: français-anglais & anglais-français. Paris: Larousse, 1993.

Oxford-Hachette French Dictionary: French-English & English-French. Oxford: Oxford University Press and Paris: Hachette, 1994.

Tables

item	Total	TEs 2+	Tail	Omissions
approach (n)	50	32	15	3
identify	69	53	11	5
major (adj)	105	87	11	7
authority	140	119	16	5
involve	88	51	24	13
affect (v)	73	43	27	3
issue	164	122	18	24
with	671	538	25	108
regular	20	16	4	0
control (n)	83	79	4	0
facility	16	9	4	3
type	211	191	8	12

Table 1. Corpus Occurrences

Corpus (88 examples)	Collins-Robert	Larousse	Oxford-Hachette
participer (13)	impliquer	impliquer	impliquer
Ø (13)	mêler	comporter	nécessiter
impliquer (11)	entraîner	concerner	entraîner
comporter (3)	nécessiter	toucher	faire
intervenir (3)		absorber	participer
engager (3)			mêler
concerner (2)			concerner
entraîner (2)			prendre
comprendre (2)			s'engager
s'agir de (2)			prendre part à
se mêler à (2)			
Others (47) (e.g. avoir trait à, s'occuper de, nécessiter, etc)			

Table 2: Translations of *involve*

Word	Total in Corpus	Which diction-ary	First TE %	Other TEs %	Examples %	Total match %	Total not matched %
approach (n)	50	CR	52	0	4	56	44
		L	52	0	0	52	48
		OHD	0	58	2	60	40
identify	69	CR	52.2	4.3	0	56.5	43.5
		L	52.2	2.9	0	55.1	44.9
		OHD	52.2	1.4	0	53.6	46.4
major (adj)	105	CR	16.2	0	34.3	50.5	49.5
		L	2.9	37.1	14.3	54.3	45.7
		OHD	25.7	29.5	0	55.2	44.8
authority (n)	140	CR	25	52.1	2.1	79.2	20.8
		L	25	52.1	3.6	80.7	19.3
		OHD	25	39.3	0.7	65	35
involve	88	CR	12.5	4.5	14.8	31.8	78.2
		L	12.5	5.7	20.5	35.7	74.3
		OHD	12.5	22.7	2.3	37.5	72.5
affect (v)	73	CR	2.7	54.8	0	57.5	42.5
		L	2.7	53.4	6.9	63	37
		OHD	4	43.8	0	47.8	52.2
issue	164	CR	22.6	42.7	2.4	67.7	32.3
		L	22.6	38.4	1.2	62.2	37.8
		OHD	6.7	55.5	0	62.2	37.8
with	671	CR	37.6	27	7.7	72.3	27.7
		L	37.6	27	5.7	70.3	29.7
		OHD	13.9	53.1	24.1	91.1	8.9
regular (adj)	20	CR	65	0	0	65	35
		L	65	5	15	85	15
		OHD	65	5	15	85	15
control (n)	83	CR	0	68.7	6	74.7	25.3
		L	0	67.5	7.2	74.7	25.3
		OHD	24.1	49.4	1.2	74.7	25.3
facility	16	CR	6.3	62.5	0	68.8	31.2
		L	6.3	62.5	0	68.8	31.2
		OHD	0	62.5	0	62.5	37.5
type	211	CR	76.3	14.2	5.7	96.2	3.8
		L	76.3	14.2	0	90.5	9.5
		OHD	76.3	14.2	0	90.5	9.5
1	2	3	4	5	6	7	8

Table 3: Percentage of matches between dictionaries and corpus