

Enriching the phraseological coverage of high-frequency adverbs in English-French bilingual dictionaries

Sylviane Granger and Marie-Aude Lefer
Université Catholique de Louvain, Belgium

As pointed out by Stig Johansson (2007), one of the most obvious applications of cross-linguistic corpus research is in bilingual lexicography. The treatment of phraseological units in bilingual dictionaries, however, still lags far behind that in monolingual dictionaries. We compare the coverage of one category of phraseological units, namely 'lexical bundles' (Biber *et al.* 1999), in English-French bilingual dictionaries with their patterns of usage in various corpora. A detailed study of two high-frequency adverbs, *encore* in French and *yet* in English, shows that the use of monolingual and translation corpus data can help improve the number of lexical bundles included in the bilingual dictionary entry, the number and accuracy of translation equivalents, and the prototypicality and authenticity of the examples. Using corpora also makes it possible to avoid categorial bias in translations.

Keywords: phraseology, bilingual lexicography, adverbs, lexical bundles, translation

1. Introduction

In many of his publications, from the early descriptions of the English-Norwegian parallel corpus project (1994) to his magisterial volume published in 2007, Stig Johansson has explicitly identified bilingual (and multilingual) lexicography as one of the most obvious applications of cross-linguistic corpus research. Probably due to the lack of good quality bilingual corpora (especially translation corpora), the impact of corpora on bilingual dictionaries has however been fairly limited. As pointed out by Salkie (2008), this situation contrasts sharply with that for monolingual lexicography: "[t]ranslation (parallel) corpora are standard tools in several

fields, such as translator training, machine translation, contrastive linguistics, and various language engineering applications. One area where one might expect such corpora to be widely used is bilingual lexicography, but in fact such corpora have not been exploited significantly in dictionary compilation – unlike monolingual lexicography, where it would be unthinkable today not to use single-language corpora”.

One of the aspects of dictionaries that has benefited most from corpus analysis is phraseological coverage. However, advances in this area too have been much slower in bilingual dictionaries than in monolingual ones, a point noted by Rundell in 1999, and which is still largely true today: “[t]he extraordinary range of lexical and grammatical information they [monolingual learners’ dictionaries] include is rarely even approached by the best bilingual dictionaries available”. This is not to say that phraseology is absent from bilingual dictionaries. Major improvements have been made in recent years, notably in the treatment of collocations (Atkins 1996; Lubensky & McShane 2007).

The “huge area of syntagmatic prospection” (Sinclair 2004: 19) opened up by corpus linguistics covers a wide range of multiword units. In addition to traditional categories such as idioms, proverbs and collocations, automatic techniques have uncovered some phraseological patterns that do not fit into any of the usual classifications. Among them are lexical bundles, which Biber *et al.* (1999: 990ff) defined as “simple sequences of word forms that commonly go together in natural discourse”. These routinised sequences, or “‘preferred’ ways of saying things” to use Altenberg’s (1998: 122) words, include verbal phrases (e.g. *suffice it to say*), nominal phrases (*the extent to which*), prepositional phrases (*in the case of*) and adverbial phrases (*as yet*). Unlike the units that were focused on in pre-corpus phraseology, these units tend to be semantically compositional. As a result, they are less salient and even very advanced users of the language – among them, translators and lexicographers – may fall into the trap of literal translation, an option which leads to clumsy, if not downright incorrect, formulation. Lexical bundles have been largely neglected in bilingual dictionaries. They are sometimes included as middle- or back-matter sections designed to help users write essays or letters, but as pointed out by Granger and Paquot (2008), these sections feature some questionable and/or infrequent phrases and would greatly benefit from the incorporation of more corpus data.

It is therefore urgent to find ways of ‘phrasing up’ the bilingual dictionary. This is all the more important as several studies have shown that users have a preference for bilingual dictionaries (see Lew 2004: 18–20). Some recent studies have convincingly demonstrated the key role that bilingual corpus data can play in improving the collocational coverage of bilingual dictionaries (e.g. Ferraresi *et al.* 2010). In this article our focus is on a different category of multiword units, that

of lexical bundles, identified by the n-gram extraction method. We focus on high-frequency words, a category which displays a high rate of phraseological uses. As adverbs have received scant attention so far, we selected two high-frequency adverbs: the French adverb *encore* and the English adverb *yet*, which are frequent translation equivalents.¹ In Section 2 we describe the corpora and the bilingual dictionaries used for the investigation. In Section 3 we examine the coverage of corpus-derived bundles in bilingual dictionaries both quantitatively (the proportion of bundles included) and qualitatively (their place in the dictionary micro-structure). Section 4 describes the contribution of corpus data to the translation of lexical bundles in bilingual dictionaries, and the final section presents some concluding remarks and suggestions for future research.

2. Data and methodology

Our study relies on both lexicographic and corpus data. The two types of data are described in this section, together with the methodology used to investigate lexical bundles and their translation equivalents. For the dictionary analysis, we made use of the *yet* and *encore* entries in three English-French electronic dictionaries: *Le Grand Robert & Collins* (2008) (henceforth referred to as RC), *Grand Dictionnaire Hachette-Oxford* (2003) (henceforth HO) and *Harrap's Unabridged Pro* (2004) (henceforth HU). HO and RC are corpus-informed: English and French monolingual data were used to devise and/or revise the bilingual entries. The situation for HU is less clear: its introduction specifies that the dictionary is based on "texts in searchable databases" but no further details are provided.

Our approach relies on a two-stage methodology. The first stage consists of extracting lexical bundles including *encore* or *yet* from original French and English texts respectively. This stage requires the use of monolingual reference corpora, which should ideally be as large and representative as possible. Translation corpora are then used in the second stage to identify the translation equivalents of the chunks uncovered in the first part of the analysis.

The case study of the French adverb *encore* is a first attempt at implementing this methodology. French currently suffers from the lack of a representative corpus along the lines of the *British National Corpus* (BNC) or the *Corpus of Contemporary American English*, and so the study was exclusively based on the *Label France* (LF) unidirectional translation corpus. This corpus contains 1 million words and is

1. We selected *yet* as it is an equivalent of *encore* that has a high rate of phraseological uses (43% of phraseological uses in the *British National corpus*), which is not the case for other frequent equivalents such as *still* (8%) or *again* (15%).

made up of French magazine articles translated into English.² The *LF* corpus was used for both stages of the analysis: extraction of the bundles from the original French texts and identification of their translations in the English target texts. The wider availability of corpus resources for English allowed us to refine this methodology for the analysis of the adverb *yet*. We used the 100-million-word *BNC* as a reference corpus to extract bundles including *yet*. We then relied on two large bidirectional translation corpora to zoom in on the French translations of some of these bundles: *PLECI*³, which contains news and fiction, and the *Europarl5* corpus (Koehn 2005), which consists of the proceedings of the European Parliament. In the *yet* study, to exploit the full potential of the methodology, both translation directions were investigated (*yet* in English source and target texts).

We used the n-gram method to extract lexical bundles. This method is employed in a wide range of research fields, notably in English for Academic Purposes research (see Biber *et al.* 2004 and Ellis *et al.* 2008), but remains largely under-exploited in bilingual lexicography. We extracted 2- to 5-grams with *WordSmith Tools 5* (Scott 2008) and imposed a frequency cut-off of 5 occurrences per million words for *encore* in *Label France*, and 5 occurrences per 10 million words for *yet* in *BNC*. These are relatively low frequency thresholds: Biber and his colleagues (1999, 2004) used much higher frequencies (between 10 and 40 occurrences per million words). We then manually edited the n-gram lists to weed out strings which were unlikely to be of lexicographic interest (see Tables 1 and 2 for examples of rejected and selected lexical bundles respectively).

Table 1. Examples of rejected n-grams with *yet* and *encore*

English: <i>yet the, yet it, yet at the, yet he had, yet they were, have yet been, is not yet known</i>
French: <i>est encore, encore le, encore dans, reste encore, encore à la, on peut encore, il y a encore</i>

Table 2. Examples of selected lexical bundles with *yet* and *encore*

English: <i>not yet, and yet, yet another, as yet, yet to be (+ past participle)</i>
French: <i>ou encore, pas encore, encore plus, encore aujourd'hui/aujourd'hui encore, là encore</i>

2. The *Label France* and *PLECI* corpora used in this study were compiled at the *Centre for English Corpus Linguistics* (University of Louvain). See <<http://www.uclouvain.be/en-258636.html>> for more information.

3. *PLECI* (Poitiers-Louvain Échange de Corpus Informatisés) is the result of a collaboration between the University of Louvain and the University of Poitiers.

Label France and *PLECI*, which are both sentence-aligned, were analysed with the *ParaConc* multilingual concordancer (Barlow 2002), while the *Europarl5* sub-corpora were investigated via the *Sketch Engine* (Kilgariff *et al.* 2004). Note that for *Europarl5* we only used texts which had English or French explicitly identified as the source language. This was done by creating sub-corpora in the *Sketch Engine* on the basis of the 'speaker.language' criterion (here 'fr' and 'en').

3. Lexical bundles with *encore* and *yet*: Dictionaries vs. corpus data

This section analyses lexical bundles including the French adverb *encore* and the English adverb *yet*, and examines how they are treated in the three bilingual dictionaries reviewed. Section 3.1 looks at the proportion of corpus-extracted lexical bundles included in the dictionary entries and Section 3.2 takes a closer look at the lexical bundles that are included and describes their place in the microstructure of the dictionaries (sub-entry, example, etc.).

3.1 Coverage

In this section, we aim to find out whether lexical bundles are well covered in current bilingual entries. We will first look at *encore* (French-to-English half of the dictionary), before turning to *yet* (English-to-French half).

3.1.1 ENCORE

The corpus analysis of the original French texts of *Label France* shows that lexical bundles including *encore* are very frequent: they make up 47% of all *encore* uses in the corpus (352 out of 745 occurrences), which is a considerable proportion. Combining the different chunks found in the corpus and those listed in the bilingual entries gives us a total of 34 different multiword uses of the adverb. In terms of coverage, as can be seen in Table 3, the number of different lexical bundles found in each of the three dictionaries and in the *LF* corpus is similar (between 18 and 21 chunks out of 34).

Table 3. Number of different multiword uses of *encore* in the bilingual dictionaries and the *Label France* corpus

HU	RC	HO	LF corpus	Total
18	19	21	21	34
53%	56%	62%	62%	100%

This first impression of similarity is quickly dispelled when we look at which lexical bundles are present in each resource. In fact, there is relatively little overlap between the dictionaries and the *LF* corpus: (1) 41% of the chunks that are found in the corpus are not well represented in the bilingual dictionaries (they occur in none, or only one, of the three dictionaries); (2) 26.5% of the lexical bundles that are well covered by the dictionaries (i.e. present in two or three dictionaries) do not occur at all in the *LF* corpus; (3) a meagre 5 chunks out of 34 (14.7%) are included in all three dictionaries and the corpus. To understand the reasons for these striking differences, we need to look at the lexical bundles themselves. The two sets of chunks (those that are only found in dictionaries, and those that are not well-represented in dictionaries but are found in the corpus), are presented in Table 4. It is clear that the lexical bundles that are found only in dictionaries are characteristic of speech, whilst those that are found in the corpus but neglected in dictionaries are typical of writing. Dictionaries seem to favour interactional and attitudinal markers typical of speech, as illustrated in Examples (1) to (6). In the *LF* corpus, by contrast, we find many cohesive markers typical of writing that fulfil a range of functions, such as addition, disjunction, enumeration, concession and emphasis (see Examples 7 to 10).

Table 4. Lexical bundles including *encore* in bilingual dictionaries and in the *LF* corpus

Not found in the <i>LF</i> corpus, but listed in 2 or 3 dictionaries	Found in the <i>LF</i> corpus, but listed in at most 1 dictionary
encore toi ! / encore vous !	ou (bien) encore
encore une chance que	mais encore
encore heureux	là encore
et encore !	une fois encore
mais encore ?	(et) plus encore
quoi encore ? / (et puis) quoi encore ?	et bien/beaucoup d'autres encore
si encore	plus encore ADJ/ADV que
encore autant	encore davantage
encore pire / pire encore	encore et encore
encore rien	encore et toujours / toujours et encore
	encore aujourd'hui / aujourd'hui encore
	mieux encore
	encore à (+ <i>inf.</i>)

- (1) il a dit qu'il avait bien aimé, **mais encore** ? (HU)
he said he liked it – **but what exactly did he say?**
- (2) **encore une chance** qu'il n'ait pas été là ! (HU)
thank goodness or it's lucky he wasn't there!

- (3) **encore vous !** (RC)
(not) you again!
- (4) **si encore** je savais où ça se trouve, j'irais bien (RC)
if only I knew where it was, I would willingly go
- (5) c'est tout au plus mangeable, **et encore !** (HO)
it's only just edible, if that!
- (6) **et puis quoi encore !** (HO)
what next!
- (7) Ce sont quelques-uns de ces artistes que Label France a choisi de vous présenter dans les domaines renommés et bien établis que sont le vitrail, la céramique, le verre **ou encore** la tapisserie. [LF]
For this issue «Label France» has selected a small number of these artists in the well-known and well established fields of stained glass window-making, ceramics, glass-blowing **and** tapestry.
- (8) Or, **là encore**, l'être humain doit coopérer à sa venue, lui permettre d'obtenir une place. [LF]
Yet **here too**, the human being has to co-operate with its coming, to allow it to obtain a place.
- (9) Sur scène, Marianne Sergent, Zouc, Sylvie Joly, les Jeanne, pour la première génération, puis, dans les années 1980–1990, Muriel Robin, les Vamps, Anne Roumanoff, Valérie Lemerrier, Michèle Laroque, **et bien d'autres encore**, renoncent à leurs atouts millénaires mais aliénants (...). [LF]
On stage, Marianne Sergent, Zouc, Sylvie Joly, the Jeannes, for the first generation, then in the 1980–1990s, Muriel Robin, the Vamps, Anne Roumanoff, Valérie Lemerrier, Michèle Laroque, **and many more besides**, are abandoning their age-old but maddening assets (...).
- (10) **Une fois encore**, Sautet transcende le polar en le poussant vers l'étude de mœurs, dans une ambiance très noire qui surprend. [LF]
Once again, Sautet transcends the whodunit by pushing it towards a study of moral standards, in a very black and surprising atmosphere.

It is unsurprising to find uses typical of writing in a written corpus like *LF*; the preference for speech displayed by the bilingual dictionaries is more remarkable. The reason for this preference might be that spoken lexical bundles are more cognitively salient, i.e. they stand out in our minds when we think about language (Hanks 2000). Our study therefore suggests that lexicographers need corpus data and corpus-driven methods for the automatic extraction of recurrent sequences that are less cognitively salient. In the absence of corpus data, they run the risk of overlooking these chunks when devising or revising bilingual entries.

3.1.2 YET

The English adverb *yet*, like *encore*, is frequently found in lexical bundles: 43% of the occurrences of *yet* in the *BNC* are instances of multiword uses (14,724 out of 33,980 occurrences). However, the coverage in bilingual dictionaries is considerably richer than that for *encore*. As shown in Table 5, the most frequent chunks with *yet*, such as *not yet*, *and yet*, *yet another* and *as yet*, are included in the bilingual entries in at least two of the three dictionaries reviewed.

Table 5. Most frequent lexical bundles with *yet* in the *BNC* and their coverage in bilingual dictionaries

Lexical bundles	Frequency in the <i>BNC</i>	Number of dictionaries that include the lexical bundles
<i>not yet</i>	4,357	3
<i>and yet</i>	3,442	2
<i>yet another</i>	1,500	3
<i>as yet</i>	1,419	3
<i>yet to</i> + INFINITIVE	1,115	3
<i>yet to be</i> + PAST PART.	648	2
<i>yet again</i>	642	3
<i>yet more</i>	372	3

The *encore* and *yet* case studies point to an imbalance between the two halves of the dictionaries: lexical bundles seem to be relatively well covered in the English-to-French half, while the coverage of the French-to-English half seems to be much patchier. One tentative explanation for this difference could be that English and French are part of radically different 'corpus cultures'. There are many more corpora of English than of French. Most English dictionaries are corpus-informed, while the 'corpus revolution' appears to have largely passed French lexicography by.

3.2 Presentation

As regards presentation, we found that most lexical bundles including *encore* and *yet* are buried in bilingual entries, i.e. very few chunks are listed as sub-entries in their own right. The boxed bundles in Figure 1 (*and yet*, *not yet*, *yet to*, *as yet*, *yet more*, *yet another* and *yet again*) are either only found in contextualised example sentences (e.g. *the campaign has yet to begin*) or listed as decontextualised items within sub-entries (e.g. *not yet*, *yet again*). The situation is similar in the other two dictionaries examined: no lexical bundle with *yet* is granted sub-entry status (let alone headword status).

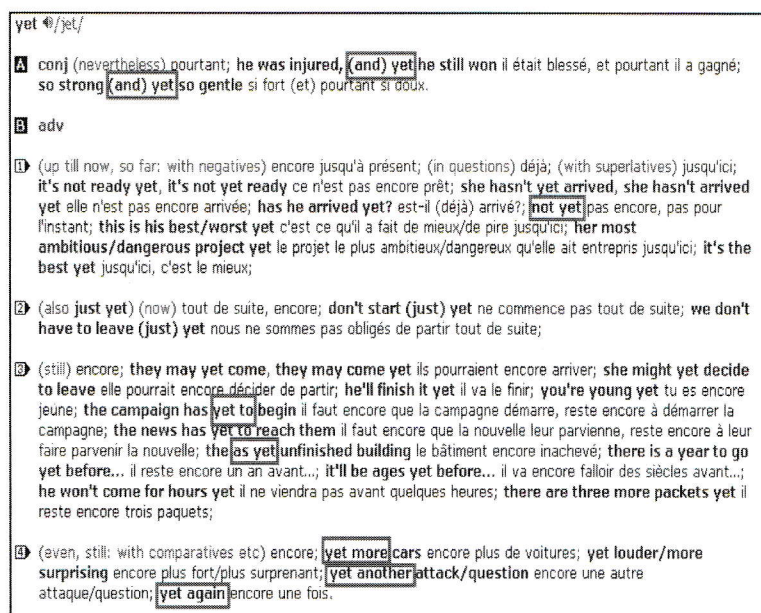


Figure 1. Yet entry in Hachette-Oxford (2003)

In the French-to-English half of the dictionaries, only five chunks (out of 34 if we consider the lexicographic and corpus data together) are listed as sub-entries: *encore que* ('even though') [HO, HU, RC], *et encore* ('if that') [HO, RC], *pas encore* ('not yet') [RC], *encore et encore* ('again and again') [RC] and *si encore* ('if only') [RC]. As in the English-to-French half, the other chunks – if included at all – are buried in the entries, for example in the form of an example.

Clearly, the presentation of lexical bundles could be improved in both parts of the dictionaries, notably by granting headword status to lexical bundles. As shown by Tono (2000), phrases are much easier to find if they are listed as headwords rather than incorporated into the entry itself. Our analysis reveals another weakness of current bilingual dictionaries, viz. the poor quality of the examples. The selected examples are often atypical, and this is particularly problematic in view of the fact that a large proportion of lexical bundles are only introduced via examples. This is best illustrated with *as yet* in the entry presented in Figure 1. The only example mentioned in HO is *the as yet unfinished building*. However, the *the as yet...* structure accounts for only 1.5% of the occurrences of *as yet* in the BNC. A corpus-based phraseological analysis can ensure that the examples included in the dictionary meet the two major requirements of good examples, i.e. authenticity and prototypicality (Cappeau 2010: 129).

4. Enriching the translations with bilingual corpus data

This section is devoted to the translation equivalents of (some of) the bundles including *encore* (Section 4.1) and *yet* (Section 4.2). Our starting-point assumption is that the most frequent equivalents of the adverbs and the lexical bundles of which they form part should be included in the bilingual entries.

4.1 The translation of lexical bundles including *encore*

Table 6 presents the most frequent translations of the *encore* occurrences found in *LF*, in decreasing order of frequency, and their inclusion (or lack thereof) in the three dictionaries. No distinction is made in the table between the phraseological and non-phraseological uses of *encore*.

Table 6. English translations of the most common occurrences (minimum frequency 5) of *encore* in the *Label France* corpus and their appearance in three bilingual dictionaries

English translations	Freq. in <i>LF</i> corpus	HU	RC	HO
<i>still</i>	234	✓	✓	✓
<i>and</i>	48	×	×	×
<i>not yet</i>	43	✓	✓	✓
<i>or</i>	32	×	×	×
<i>even</i> ADJ-er / <i>even more</i> ADJ	30	✓	✓	✓
<i>further</i>	18	×	×	×
<i>even today</i>	15	×	×	×
<i>or even</i>	15	×	×	×
<i>and even</i>	14	×	×	×
<i>again</i>	13	✓	✓	✓
<i>once again</i>	11	×	✓	✓
<i>yet</i>	11	×	×	×
<i>here too</i>	10	×	×	×
<i>as yet</i>	9	✓	×	×
<i>even</i>	9	×	✓	×
<i>or again</i>	8	×	×	×
<i>as well as</i>	7	×	×	×
<i>yet to be</i> + PAST PARTICIPLE	7	×	×	×
<i>another</i>	6	✓	✓	✓
<i>even more</i>	6	✓	×	×
<i>still today</i>	6	×	×	×
<i>more</i> ADJ <i>still</i>	5	×	×	×
<i>too</i>	5	✓	✓	×
Total no. of translations	23	8	8	6

The coverage of the three dictionaries is very similar. But between them the dictionaries only cover between a quarter and a third of the top 23 most frequent translation patterns found in the corpus. Regrettably, these are not necessarily even the most frequent ones. We see, for example, that two frequent translations, *and* and *or*, are not listed. Several translation equivalents extracted from *Label France* could be used to enrich the bilingual entry for both phraseological and non-phraseological uses of *encore*. We will illustrate this by means of two French-English pairs: *encore* – *further*, and *ou encore* – *and/or*.

Encore is a highly polysemous adverb (Mosegaard Hansen 2002). It can, among other things, mean *davantage* in French, and a number of corresponding translations are recorded in the dictionaries (*more*, *even more*, *another*; see Examples 11 to 13). In the *LF* corpus *further* is quite a frequent translation of *encore*, especially when *encore* is used in combination with a verb (see Examples 14 and 15). Even though *further* clearly deserves inclusion in the bilingual entry, it is not mentioned in any of the dictionaries we reviewed.⁴

- (11) *encore* un mot, avant de terminer (RC)
(just) one *more* word before I finish

- (12) réduisez-le *encore* (HU)
Reduce it *even more*

- (13) *encore* une tasse de café (HU)
another cup of coffee

- (14) Or tous les pays n'ont pas ratifié les deux pactes, certains pays privilégiant l'un ou l'autre, ce qui conduit déjà à une dissociation de fait, *encore accentuée* par une autre dissociation qui tient aux mécanismes de contrôle, beaucoup plus développés en matière de droits civils et politiques qu'en matière de droits économiques, culturels et sociaux. [LF]

Yet not all countries have ratified both pacts, some countries favouring one or the other, which in itself leads to a de facto dissociation, *further emphasized* by another dissociation to do with the mechanisms of control, which are far more developed in the area of civil and political rights than in that of economic, cultural and social rights.

- (15) Nous devons aussi *améliorer encore* la sécurité des transports et renforcer la lutte contre le financement du terrorisme. [LF]

We also have to *further improve* transport security and step up the fight against the financing of terrorism.

4. It should be added, however, that *even further* is mentioned in HU. This equivalent is not present in LF.

The usefulness of translation corpus data can also be illustrated with the lexical bundle *ou encore* (and its variant *ou bien encore*), which is used in French to convey enumeration. *Ou encore* is the most frequent chunk with *encore* in *LF*. It accounts for 20% of all *encore* uses in the corpus but is only recorded in one dictionary (HO). Table 7 contains the English translations of *ou encore* found in *LF*, in decreasing order of frequency. The most frequent translations are *and* and *or* (58%), as illustrated in Examples (16) and (17), while the only translation found in HO is *or else* (Example 18). *Or else* is not present in the *LF* corpus. It can be used to enumerate verbs (see Example 18: *swim, go scuba diving or else learn to sail*) but it would be an incorrect translation in the majority of the corpus examples examined here, where *ou encore* is used to enumerate nouns or prepositional phrases (introduced by *à, contre, en, par* or *pour*) rather than verbs. In other words, the example listed in HO is atypical.

Table 7. English translations of *ou encore* in the *Label France* corpus

English translation	Frequency in <i>LF</i> corpus	%
<i>and</i>	47	34.6
<i>or</i>	32	23.5
<i>or even</i>	15	11.0
<i>and even</i>	14	10.3
<i>or again</i>	8	5.9
<i>as well as</i>	7	5.1
<i>and ... too</i>	4	2.9
<i>and also</i>	2	1.5
<i>and again</i>	1	0.7
<i>both</i>	1	0.7
<i>or then again</i>	1	0.7
<i>then</i>	1	0.7
no translation ⁵	3	2.2
Total	136	100

5. The category 'no translation' includes cases of zero translation and cases where a whole sentence or paragraph containing *ou encore* in the source text has not been translated.

- (16) Des groupes financiers comme Dexia Asset Management, BNP Paribas ou encore la Macif Gestion sont engagés dans ce processus. [LF]

Some financial groups such as Dexia Asset Management, BNP Paribas and Macif Gestion are committed to this process.

- (17) Car si l'humour anglais – le terme est même de leur invention –, la fantaisie slave ou encore l'autodérision des Italiens sont, par exemple, réputés, la France, malgré des succès parfois même internationaux dans ce domaine, n'est pas automatiquement associée à l'humour. [LF]

Whereas English humour – they even invented the term – Slavonic whimsy or the self-mockery of Italians, for instance, are well-known, France, in spite of successes in this field sometimes even at international level, is not automatically associated with humour.

- (18) vous pouvez pratiquer la natation, la plongée sous-marine ou encore vous initier à la voile [HO]

you can swim, go scuba diving, or else learn to sail

A caveat is in order, however. In spite of the undeniable usefulness of translation corpus data, such data should not be used indiscriminately in lexicographic projects. The corpus contains a non-negligible number of incorrect translations of *ou encore*, most of which are due to source text interference. In several cases the translator seems to have been unaware of the meaning of *ou encore* and mistranslated it into *or even*, *and even* and *or again* (see Examples 19 to 21). These cases account for 27% of the translation corpus data, which is a substantial proportion.

- (19) Lille renferme l'une des plus prestigieuses collections françaises de peintures, dessins, sculptures, faïences et objets d'art: celle-ci n'a-t-elle d'ailleurs pas fait l'objet d'expositions à New York, Londres ou encore au Japon au cours de ces cinq dernières années? [LF]

The Lille museum houses one of the most prestigious of France's collections of paintings, drawings, sculptures, ceramics and objets d'art, all of which have been exhibited in New York, London and even Japan in the course of the last five years.

- (20) Un apprentissage que les candidats pourront poursuivre, en France, dans l'un des nombreux centres d'enseignement du FLE, les universités, à l'Alliance française ou encore dans les chambres de commerce et d'industrie. [LF]

Applicants will be able to continue the learning process in France at one of the many FLE teaching centres, universities, Alliances Françaises or even at chambers of commerce and industry.

- (21) Milan Kundera, après avoir fui sa Tchécoslovaquie natale, a lui aussi fini par adopter notre langue, comme Cioran, l'écrivain roumain, Vassilis Alexakis, le Grec, ou encore François Cheng, le plus francophone des écrivains chinois. [LF]
 Milan Kundera, having fled his native Czechoslovakia, ended up adopting our tongue, as did Cioran, the Romanian writer, Vassilis Alexakis, the Greek writer, or again François Cheng, the most accomplished of the Chinese writers writing in French.

Some researchers might be tempted to use the imperfection of translation corpora as an argument against using translation corpus data in bilingual lexicography. We believe, however, that lexicographers, who are highly-skilled bilinguals, are perfectly capable of separating the wheat from the chaff. One added bonus of these mistranslations is that they draw lexicographers' attention to frequent errors and provide them with useful material to design corpus-informed warning boxes such as 'do not literally translate *ou encore* into *or even/and even* or *or again*' which could be included in the bilingual entries.

4.2 The translation of lexical bundles with *yet*

This section focuses on the French translation of two frequent lexical bundles including *yet: as yet* (fourth most frequent chunk) and *yet another* (third most frequent chunk). In this part of the study, we have relied on data extracted from *PLECI* and *Europarl5* in both translation directions (*as yet* and *yet another* in English source and target texts), which places us in a good position to assess the added value of bidirectional corpus data.

We examined 135 bilingual concordances of *as yet* and found more than a dozen French equivalents. These are listed in decreasing order of frequency in Table 8 and illustrated in Examples (22) to (24). Only two of these equivalents (*encore* and *déjà*) are recorded in bilingual entries. This once again demonstrates the need for translation corpus data to improve the translations included in bilingual dictionary entries. Interestingly, the results show that it would be advisable to look at both translation directions as some of the French equivalents can only be uncovered by examining the French-to-English direction (i.e. *as yet* in English target texts rather than in source texts). For example, examining cases where *as yet* is found in English target texts makes it possible to unearth equivalents such as *aujourd'hui* (see Example 25) and *jusqu'alors/jusque là*, which would be overlooked if only the English-to-French translation direction were investigated.

Table 8. French equivalents of *as yet* in PLECI and Europarl5

Corpus equivalents	Original English, translated French	Original French, translated English	Total
<i>encore</i>	28	41	69
<i>à ce jour/à ce stade/à la date d'aujourd'hui</i>	7	4	11
<i>Pour l'heure/pour l'instant/pour le moment</i>	8	2	10
<i>jusqu'à présent/jusqu'ici</i>	8	1	9
<i>aujourd'hui</i>	0	8	8
<i>toujours</i>	5	2	7
<i>encore ... aujourd'hui/ aujourd'hui ... encore</i>	0	2	2
<i>jusqu'alors/jusque-là</i>	0	2	2
<i>déjà</i>	1	0	1
<i>dorénavant</i>	1	0	1
<i>en l'état</i>	1	0	1
<i>mais encore</i>	0	1	1
Other ⁶	8	5	13
Total	67	68	135

- (22) In my view, only a very small number of somewhat equivocal issues remain
as yet unresolved. [Europarl5]

À mon avis, seul un nombre très restreint de questions quelque peu équivoques
restent à ce jour non résolu.

- (23) We note that as yet there is no evidence to link depleted uranium and ill health
suffered either by troops or civilians. [Europarl5]

Nous notons qu'il n'y a jusqu'à présent aucune preuve du lien entre l'uranium
appauvri et les problèmes de santé dont ont souffert soit des soldats, soit des
civils.

- (24) As yet, western broadcasters are under little political pressure to cut back.
[PLECI news]

Pour l'instant, peu de pressions politiques poussent les stations occidentales
à réduire leur budget.

6. The 'other' category includes cases of modulation, zero translation (in the English-to-French translation direction, when *as yet* is not translated into French in the target text) and addition (in the French-to-English translation direction, when *as yet* does not correspond to any French item in the source text).

- (25) Nous savons très bien qu'il n'existe **aujourd'hui** aucune étude de toxicologie pour évaluer les conséquences de la dissémination d'OGM dans l'environnement. [Europarl5]
We know full well that **as yet** there is no toxicology study that evaluates the consequences of releasing GMOs into the environment.

The examination of *yet another* and its French equivalents brings to light another interesting feature of bilingual entries. *Yet another* is listed in all three dictionaries and is systematically translated into an adverb or an adverbial phrase (*encore, de plus*), as shown in Examples (26) to (28). However, the corpus data reveal many excellent translations with other parts of speech. These include translations into adjectives such as *nouveau* (new), *énième* (nth, umpteenth), *supplémentaire* (supplementary) (see Examples 29 to 31). Langlois (1996) provides similar examples: while bilingual dictionaries translate the adverb *automatically* into the French adverb *automatiquement*, the translation corpus shows that in 25% of cases *automatically* is translated into something else, such as a phrase including the adjective *automatique* (e.g. *to pay dues automatically: payer les primes par versement automatique*). It can therefore be concluded that another benefit of translation corpora is that they can help lexicographers free themselves from the **categorical bias**, i.e. the tendency to translate a given source item exclusively into a word of the same grammatical category in the target language (translate an adjective into an adjective, an adverb into an adverb, etc.).

- (26) she was **yet another** victim of racism [RC]
c'était une victime **de plus** du racisme
- (27) **yet another** attack/question [HO]
encore une autre attaque/question
- (28) **yet another** bomb [HU]
encore une bombe
- (29) Nous n'avons pas besoin d'un **énième** rapport dirigiste. [Europarl5]
We don't need **yet another** interventionist report.
- (30) I heard the other day of **yet another** attack on the British farmer, namely that the Government is about to impose a tax on fertilizer – that is just as an aside. [Europarl5]
J'ai entendu parler l'autre jour d'une **nouvelle** attaque à l'encontre des agriculteurs britanniques, à savoir que le gouvernement est sur le point de prélever une taxe sur les engrais.
- (31) Finally, I would make a plea on behalf of my group: let us not set up **yet another** agency. [Europarl5]
Pour conclure, je voudrais introduire une demande au nom de mon groupe : abstenons-nous de créer une agence **supplémentaire**.

Table 9 presents the complete inventory of the French equivalents of *yet another* found in *PLECI* and *Europarl5*. As in the *as yet* dataset, we can see that some equivalents, such as *énième* (and its variant *nième*), are more frequent in the French-to-English translation direction. In fact, looking at the French components of the *Europarl5* sub-corpora used in this study (original French and translated French), we find that *énième* is five times more frequent in French source texts than in French target texts translated from English. Many of the studies described in Stig Johansson's 2007 volume point to similar differences between original and translated texts. Clearly, looking at both translation directions holds great potential for lexicographers: it can help them uncover all possible cross-linguistic equivalents and avoid overlooking potentially interesting contrasts.

Table 9. French equivalents of *yet another* in English found in *PLECI* and *Europarl5*

Corpus equivalents	Original English, translated French	Original French, translated English	Total
<i>nouveau</i>	35	8	43
<i>de plus</i>	13	18	31
<i>autre</i>	23	6	29
<i>énième</i>	6	12	18
<i>encore</i>	4	14	18
<i>supplémentaire</i>	7	4	11
Equivalents with less than 5 occurrences	11	15	26
Other ⁷	9	6	15
Total	108	83	191

5. Conclusion

Our study lends support to Stig Johansson's observation that "dictionaries fall short in the light of the evidence from bilingual corpora" (2007: 308). The use of corpus data can help (1) improve the number of lexical bundles included in the bilingual entry; (2) improve the number and accuracy of translation equivalents, especially when the two translation directions are examined; (3) improve the prototypicality and authenticity of the examples; and (4) avoid categorial bias in translations. Our study has also brought to light an interesting difference between the two dictionary halves: the phraseological coverage of the French-to-English

7. The 'other' category includes cases of modulation, zero translation (in the English-to-French translation direction, when *yet another* is not translated into French in the target text) and addition (in the French-to-English translation direction, when *yet another* does not correspond to any French item in the source text).

part of the dictionaries is quite limited as regards the number of bundles with *encore* listed in the entries, while the most frequent bundles with *yet* are included in the entries in the English-to-French part of the dictionary. In both directions, however, the translation of the phraseological units would benefit from corpus-derived insights.

One of the encouraging results of our study is that, in the absence of large balanced translation corpora, the systematic use of monolingual corpora can go a long way towards 'phrasing up' bilingual dictionaries and even small translation corpora bring their share of lexicographically relevant translation equivalents. However, there is ample scope for improvement. As noted by Moon (2008: 333), phraseological patterning varies according to genre. A corpus such as *Label France* will only produce lexical bundles typical of the genre represented in the corpus (in the same way as the *Hansard Corpus* was shown to be biased towards words and lexical patterns that are typical of parliamentary debates (Langlois 1996)). In addition, corpora like *Label France* or the *BNC* are static and cannot capture new patterns of use. Those two caveats argue in favour of collecting and analysing automatically generated web-based bilingual corpora, whose usefulness has been convincingly demonstrated by Ferraresi *et al.* (2010). It is also important to remember that corpus evidence is not everything. Moon (2008: 334) reminds us that "lexicographers still have to use intuition and judgement in selecting, interpreting, and setting out the evidence, rather than simply relaying it to the user as quasi-scientific truth." The combined use of bilingual corpus resources and lexicographers' bilingual expertise should ensure that phraseology, which has only just begun to show its promise in bilingual lexicography, is at the root of major lexicographic developments in the coming years. These advances would be a fitting tribute to Stig Johansson's inspiring legacy.

References

Corpora

- The British National Corpus*, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. <<http://www.natcorp.ox.ac.uk/>>
- Europarl corpus*: <<http://www.statmt.org/europarl/>>
- Label France corpus*: <<http://www.uclouvain.be/en-cecl-labelfrance.html>>
- Poitiers-Louvain Échange de Corpus Informatisés*: <<http://www.uclouvain.be/en-cecl-pleci.html>>

Dictionaries

- Grand Dictionnaire Hachette-Oxford Français-Anglais English-French on CD-ROM*. 2003. Oxford University Press/Hachette Multimédia.
- Grand Robert & Collins électronique français-anglais/anglais-français*. DVD-Rom Version 2.0. 2008. Le Robert/HarperCollins.
- Harrap's English-French Unabridged PRO Dictionary on CD-ROM*. 2004. Chambers Harrap Publishers.

Secondary sources

- Altenberg, B. 1998. On the phraseology of spoken English: The evidence of recurrent word combinations. In *Phraseology: Theory, Analysis and Applications*, A. P. Cowie (ed.), 101–122. Oxford: OUP.
- Atkins, B. T. S. 1996. Bilingual dictionaries: Past, present, future. In *EURALEX 1996 Proceedings*. <<http://www.euralex.org/>>
- Barlow, M. 2002. ParaConc: Concordance software for multilingual parallel corpora. In *LREC 2002 Proceedings*, Las Palmas, Gran Canaria, Spain. <<http://www.mt-archive.info/LREC-2002-Barlow.pdf>>
- Biber, D., Conrad, S. & Cortes, V. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371–405.
- Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Cappeau, P. 2010. Qu'est-ce qu'un bon exemple (oral)? In *L'exemple et le corpus. Quel statut?* [Travaux linguistiques du CerLiCO 23], P. Cappeau, H. Chuquet & F. Valetopoulos (eds), 119–132. Rennes: Presses universitaires de Rennes.
- Ellis, N. C., Simpson-Vlach, R. & Maynard, C. 2008. Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42(3): 375–396.
- Ferraresi A., Bernardini, S., Picci, G. & Baroni, M. 2010. Web corpora for bilingual lexicography: A pilot study of English/French collocation extraction and translation. In *Using Corpora in Contrastive and Translation Studies*, R. Xiao (ed.), 337–359. Newcastle upon Tyne: Cambridge Scholars.
- Granger, S. & Paquot, M. 2008. From dictionary to phrasebook? In *Proceedings of the XIII Euralex International Congress* Barcelona, Spain, 15–19 July, E. Bernal & J. DeCesaris (eds), 1345–1355.
- Hanks, P. 2000. Contributions of lexicography and corpus linguistics to a theory of language performance. In *Proceedings of Ninth Euralex International Congress*, U. Heid, S. Evert, E. Lehmann & C. Rohrer (eds), I: 3–13. Stuttgart: IMS Stuttgart University.
- Hansen, M. M.-B. 2002. La polysémie de l'adverbe encore. *Travaux de Linguistique* 44: 143–166.
- Johansson, S. 1994. Towards an English-Norwegian parallel corpus. In *Creating and Using English Language Corpora*, U. Fries, G. Tottie & P. Schneider (eds), 25–37. Amsterdam: Rodopi.
- Johansson, S. 2007. *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies* [Studies in Corpus Linguistics 26]. Amsterdam: John Benjamins.

- Kilgariff, A., Rychly, P., Smrz, P. & Tugwell, D. 2004. The sketch engine. In *Proceedings of Euralex 2004*, Lorient (France), 105–116. <<http://www.sketchengine.co.uk/>>
- Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit X, Phuket, Thailand, September 13–15, 2005, Conference Proceedings: the tenth Machine Translation Summit*, 79–86.
- Langlois, L. 1996. Bilingual concordancers: A new tool for bilingual lexicographers. *Expanding MT horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, 2–5 October 1996, Montreal, Quebec, Canada (Washington, DC: AMTA), 34–42.
- Lew, R. 2004. *Which Dictionary for Whom? Receptive Use of Bilingual, Monolingual and Semi-Bilingual Dictionaries by Polish Learners of English*. Poznań: Motivex.
- Lubensky, S. & McShane, M. 2007. Bilingual phraseological dictionaries. In *Phraseology. An International Handbook of Contemporary Research*, H. Burger, D. Dobrovolskij, P. Kühn & N. R. Norrick (eds), 919–928. Berlin: de Gruyter.
- Moon, R. 2008. Dictionaries and collocations. In *Phraseology: An Interdisciplinary Perspective*, S. Granger & F. Meunier (eds), 313–336. Amsterdam: John Benjamins.
- Rundell, M. 1999. Dictionary use in production. *International Journal of Lexicography* 12(1): 35–53.
- Salkie, R. 2008. How can lexicographers use a translation corpus? In *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies*. Zhejiang University, Hangzhou, 25–27 September, R. Xiao, L. He & M. Yue (eds). <<http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Salkie.pdf>>
- Scott, M. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software. <<http://www.lexically.net/wordsmith/>>
- Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.
- Tono, Y. 2000. On the effects of different types of electronic dictionary interfaces on L2 learners' reference behaviour in productive/receptive tasks. In *Proceedings of EURALEX 2000*.