Towards more and better phrasal entries in bilingual dictionaries

Sylviane Granger & Marie-Aude Lefer

Abstract

Although the phraseological coverage of dictionaries has improved considerably in recent years, bilingual dictionaries are still lagging behind. The objective of our paper is to show that including a range of multi-word units (MWUs) extracted via the n-gram method can considerably enhance the quality of English<>French bilingual dictionaries. We show how multiword units extracted from monolingual corpora can enhance the phraseological coverage of bilingual dictionaries and suggest ways in which the presentation of these units can be improved. We also focus on the role of translation corpora to enhance the accuracy and diversity of MWU translations in bilingual dictionaries.

1. Introduction

The impetus from the corpus-driven *Collins Cobuild Dictionary of English* (1987) and the growing influence of Sinclair's (1991) idiom principle have led to a gradual improvement in the phraseological coverage of dictionaries. Collocations in particular have been the subject of considerable attention. Separate collocation boxes are now included in some monolingual learners' dictionaries (MLDs) (e.g. *Macmillan English Dictionary for Advanced Learners* and *Longman Dictionary of Contemporary English*). These boxes highlight the preferred lexical relation between two lexical items in specific syntactic patterns, such as 'adjective + noun' (*heavy rain*) or 'verb + adverb' (*apologize profusely*). A more recent development in this area is the design of collocation dictionaries (e.g. *Oxford Collocations Dictionary for Students of English*, *Macmillan Collocations Dictionary* and *Le Robert Dictionnaire des combinaisons de mots*). Bilingual dictionaries, however, are lagging behind in this field: "[t]he extraordinary range of lexical and grammatical information they [MLDs] include is rarely even approached by the best bilingual dictionaries available" (Rundell 1999). Although the last decade has seen some progress, as testified by the work reported in several studies (cf. Ferraresi et al. 2010, Murano 2010, Todirascu et al. 2007), bilingual dictionaries still have a long way to go before they can match the phraseological coverage of MLDs. More generally, though, Sinclair's (2008: 408) observation that dictionaries are still "almost exclusively focused on the word, so much so that it is difficult to treat a phrase adequately in a dictionary and difficult to find it once it is there" remains true today, whatever the type of dictionary.

Corpus-based approaches to phraseology have uncovered a wide range of multi-word units and methods to extract them (for a survey, see Granger & Paquot 2008). Apart from collocations, a wide range of highly relevant units are largely disregarded. Among these we find recurrent contiguous strings of various lengths that are referred to as n-grams or 'lexical bundles' (Biber et al. 1999). While many such strings present no interest for lexicographic purposes (e.g. bigrams such as *of the* or trigrams such as *is that the*), many lexical bundles fulfil important functions in language. Biber et al. (2004) classifies them into three main categories: discourse organizers (e.g. *on the other hand, if we look at*), stance expressions (e.g. *it is important/necessary to*) and referential expressions (e.g. *in the United States*).

The objective of this article is to show that including a range of multi-word units (MWUs) extracted via the n-gram method can considerably enhance the quality of English<>French bilingual dictionaries. In section 2 we present the dictionaries we have compared, the corpus we have relied on and the method used to extract the MWUs. In section

3 we show how MWUs extracted from monolingual corpora can enhance the phraseological coverage of bilingual dictionaries and suggest ways in which the presentation of these units can be improved. In section 4 we focus on the role of translation corpora to enhance the accuracy and diversity of MWU translations. In section 5 we sum up some of the main findings of our study and point to some priorities for the future.

2. Data and methodology

The two bilingual French<>English dictionaries used here are online subscription-based dictionaries: *Le Grand Robert & Collins* version 2.0 (RC) and *Hachette Oxford* (HO). Both are corpus-informed: English and French monolingual data were used to devise and/or revise the bilingual entries.

The bilingual corpus used to extract the MWUs is a translation corpus containing articles on politics, economics, culture, technology and tourism from *Label France*, a quarterly magazine published by the French Ministry of Foreign Affairs. Although the corpus is quite small (c. 900,000 words), it presents a number of advantages, among others its homogeneity as regards time of publication (1998-2008) and genre and the clear identification of French as the source language and English as the target language.

The methodology involves two consecutive stages. The first stage relies exclusively on the French corpus data from which n-grams are extracted. For this study, 2- to 5-grams with a minimum frequency of 20 in the Label France corpus were extracted. After weeding out irrelevant strings, candidates for dictionary inclusion were checked against the two dictionaries with a view to improving both coverage and presentation of MWUs. The second stage, which relies on the English translations of the French bundles, aims to assess to what extent the corpus data can help improve the translations currently provided in dictionaries.

3. Improving the phraseological coverage

N-gram extraction can be carried out on the basis of monolingual corpora. At this stage, translation corpora are not needed. This is excellent news as monolingual corpora exist in abundance even for lesser-resourced languages, while bilingual corpora are still very scarce. For example, in Granger & Lefer (forthcoming) we make use of the British National Corpus (BNC) to identify the lexical bundles based on the adverb *yet*. For the current study, we have used the French part of the Label France translation corpus for n-gram extraction as French does not have a reference corpus like the BNC. Using the settings set out above, we extracted c. 6,000 n-gram types and each of the authors subsequently weeded out the n-grams which were unlikely to be of lexicographic interest (e.g. *dans le* 'in the', *de la France* 'of France', see Table 1).[1] Only complete chunks were kept and the chunks were lemmatized manually (e.g. the sequences *dans le cadre de/du/d'/des* were grouped under the phrasal lemma *dans le cadre de*). We were left with a list of 425 n-grams which were worthy of inclusion (e.g. *y compris* 'including' or *à l'instar de* 'following the example of, like', see Table 2). The n-grams were then categorised into frequency bands, ranging from fewer than 50 occurrences per million words (least frequent n-grams) to 800-900 occurrences per million words (most frequent n-grams).

Table 1. Sample of rejected n-grams.

2-grams: *dans le, et à, une véritable, la pauvreté, que vous*
3-grams: *de la France, que l'on, ne sont pas, l'ensemble des, à travers le*
4-grams: *il n'y a, le président de la, n'est pas un, n'a cessé de, qu'est-ce que*

5-grams: *l'un des plus grands, de la société de l', ils ont choisi la France, par le ministère de la*

Table 2. Sample of selected MWUs.

2-grams: *par ailleurs, en effet, y compris, à terme, bien sûr, bien entendu, droits fondamentaux*
3-grams: *en matière de, par rapport à, à cet égard, autant dire que, à l'étranger, chiffre d'affaires*
4-grams: *dans le cadre de, dans la mesure où, à l'instar de, il est vrai que, droits de l'homme*
5-grams: *de ce point de vue, qu'il s'agisse de, il en est de même, sur l'ensemble du territoire*

For the comparison of the corpus-driven list with the two bilingual dictionaries, we used the full text search option offered by both dictionaries, which enables users "to search for all occurrences of a word or sequence of words wherever they appear on both sides of the dictionary" (RC introduction). The 425 MWUs were classified into the following four categories:

- MWU as headword entry [headword];
- MWU as sub-entry or highlighted decontextualised MWU [sub-entry];
- MWU in example sentences [example];
- MWU not included in the French-to-English side of the dictionary [absent].

The overall results are included in Table 3.

Table 3. Coverage of MWU list in HO and RC.

	HO		RC	
Category	Nr of MWUs	%	Nr of MWUs	%
Headword	58	13.6	20	4.7
Sub-entry	194	45.6	226	53.2
Example	121	28.5	113	26.6
Absent	52	12.2	66	15.5
Total	425	100	425	100

As can be seen in Table 3, very similar trends emerge from the meta-analysis of the two dictionaries: (1) the two favourite strategies consist in listing the lexical bundles as sub-entries and in example sentences (together, the two categories account for c. 75% of the MWUs); (2) granting headword status to lexical bundles is the exception rather than the rule (even though the two dictionaries differ in this respect, see the discussion below). It is also noteworthy that c. 15% of the MWUs are not listed in the HO and RC French-to-English parts.

A small minority of the MWUs are included as **headword entries** (cf. full list in Table 4; the shared items are in bold). The two dictionaries slightly differ in their treatment of MWUs, with the HO dictionary granting headword status to three times more MWUs than RC (20 MWUs in RC and 58 MWUs in HO). However, the number of headword MWUs in both dictionaries remains extremely low. There is clearly a long way to go before fulfilling Sinclair's (2010 [2007]: 37) wish to see multi-word units given "the same status as the usual headwords". The fact that only 5 to 14% of the MWUs have been considered worthy of headword status is quite unfortunate as phrases are much easier to find if they are listed as headwords rather than hidden in a bilingual entry (cf. Tono 2000).

Table 4. Full list of MWUs included as headwords in HO and/or RC.

HO				RC	
à l'instar de	développement durable	**Moyen-Orient**	sciences humaines	**à l'instar de**	**rendez-vous**
aide au développement	directeur général	**parce que**	sciences sociales	**Beaux-arts**	**Royaume-Uni**
art contemporain	droits de l'homme	pays en développement	secrétaire général	**c'est-à-dire**	**savoir-faire**

684

art de vivre	**Education nationale**	pays industrialisés	septième art	chanson française	**tandis que**
arts plastiques	effet de serre	**peut-être**	société civile	**chef d'œuvre**	**vis-à-vis**
assemblée générale	état de droit	point de vue	**tandis que**	d'abord	
bande dessinée	**Extrême-Orient**	pouvoirs publics	tout à coup	**Éducation nationale**	
Beaux-arts	grand public	premier ministre	tout à fait	**Extrême-Orient**	
c'est-à-dire	**Grandes Ecoles**	prise de conscience	tout à l'heure	**Grandes Écoles**	
chef d'œuvre	**Ile-de-France**	prise en compte	tout de même	**Île-de-France**	
chiffre d'affaires	jeux olympiques	**Proche-Orient**	tout de suite	**Moyen-Orient**	
collectivité locale	langue française	quant à	vie quotidienne	n'importe	
commission européenne	long métrage	**rendez-vous**	**vis-à-vis**	**parce que**	
conseil européen	mètre carré	**Royaume-Uni**		**peut-être**	
déclaration universelle	monuments historiques	**savoir-faire**		**Proche-Orient**	

Table 4 shows that the units included as headword entries are mainly hyphenated compounds (e.g. *rendez-vous, savoir-faire, Extrême-Orient*), 'noun + adjective' compounds (especially in HO) (*art contemporain*, *directeur général*) and 'noun + preposition + noun' compounds (*chef d'oeuvre*, *effet de serre*, *point de vue*, *prise en compte*). A few complex conjunctions (*parce que, tandis que*) and cultural terms (e.g. *Éducation nationale, Grandes Écoles*) also feature in the lists. The two dictionaries mainly rely on orthographic criteria to grant headword status. In the RC French-to-English dictionary part, for example, hyphenated compounds are treated as headwords (e.g. *lave-glace* 'windscreen washer') while open compounds (e.g. *gaz naturel* 'natural gas') are given sub-entry status. The situation is different in HO, with some 'noun + adjective' and 'noun + preposition + noun' compounds listed as headwords. These differences across dictionaries can be quite confusing for dictionary users.

About 50% of the units are included as **sub-entries**. Grammatical criteria seem to be at play, with complex prepositions and adverbs standing out as favourites for sub-entry status (they are mainly labelled as 'locution' (phrase), 'locution prépositive' (prepositional phrase), 'locution adverbiale' (adverbial phrase)). Open compounds are also frequently granted sub-entry status. A sample of the MWUs which are listed as sub-entries in both dictionaries is provided in Table 5.

Table 5. Sample of MWUs with sub-entry status in HO and RC.

phrase	*à la fois, à la suite de, à part, à vrai dire, au monde, autant de, de la part de, de taille, en vigueur, mettre en œuvre, par la suite, peu à peu, sans doute*
adverbial phrase	*de près, en effet, en général, en moyenne*
prepositional phrase	*à côté de, à partir de, au delà de, auprès de, hors de, près de*
compound	*communauté internationale, construction européenne, enseignement supérieur, être humain, jeune femme, première guerre mondiale, révolution française*

About a quarter of the units are simply included in an example with no particular highlighting. A sample of MWUs appearing in example sentences in both HO and RC is given in Table 6.

Table 6. Sample of MWUs in example sentences in HO and RC.

à l'âge de	*deux fois plus*	*l'un des plus*
à l'œuvre	*d'ici à*	*mais aussi*
art moderne	*en âge de + INF*	*on assiste à*
au carrefour de	*en ce qui concerne*	*par jour*
au quotidien	*en forme de*	*plutôt que*
au sommet de	*en la matière*	*pour des raisons*
au travail	*en matière de*	*pour la première fois*
bien plus	*encore plus*	*pour qui*
c'est ainsi que	*et autres*	*sous la direction de*
c'est aussi	*et non*	*sur le marché*
c'est pour + INF	*et surtout*	*sur le plan*
dans le domaine de	*grandes villes*	*toujours plus*
dans les années	*il ne s'agit pas de*	*tout au long de*
de ... à nos jours	*le nombre de*	*tout aussi*
de plus de	*le plus grand nombre*	*un peu plus*
depuis sa création	*loin de + INF*	*un peu plus de*
deux fois par	*loin d'être*	*vie politique*

Many of the chunks included in this category are quite complex and would deserve a less cursory treatment. Including MWUs in an example sentence does not only make them less visible (cf. Bentivogli & Pianta 2002), it also has two other main drawbacks: (1) the units are highly contextualised; and (2) in most cases, only one of the possible translation equivalents is mentioned. This is the case, for example, for *dans le domaine de* which RC includes in an example sentence in the entry for *effort* (cf. Figure 1).
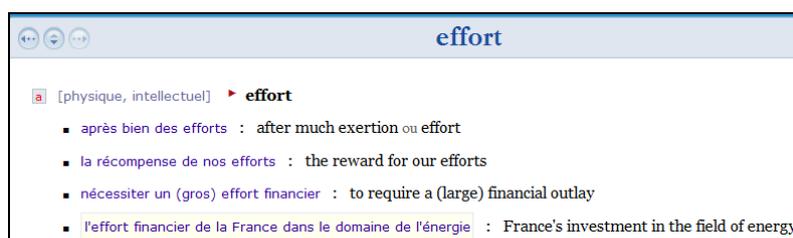

Figure 1. MWU in example sentence in RC.

A non-negligible proportion of the MWUs (c. 12% in HO and c. 16% in RC) is **absent** from the French-to-English dictionary sides. In our typology, the 'absent' category includes two sub-categories: (1) cases where the MWU is not recorded at all; and (2) cases where the French MWU only appears as a translation in the English-to-French side of the dictionary.[2] The reason for including the latter instances in the 'absent' category is that the French MWU appears 'accidentally' as a translation of an English word and the information provided is therefore often not particularly helpful. As illustrated in Figure 2, MWUs are often found in highly context-bound example translations (e.g. the chunk *un plus grand nombre* in the example sentence *many came but more stayed awa*y translated as *beaucoup de gens sont venus mais davantage* or **un plus grand nombre** *se sont abstenus*). In addition, a full text search does not always give direct access to the two sides of the dictionary and therefore requires extra effort which many users will not be ready to expend (see section 4 below).

Figure 2. MWU in example translation in RC.

As appears from Table 7, many of the chunks in the 'absent' category are longer units (3- to 5-grams) which function as discourse organizers or complex adverbs of time/place, as well as some compounds and terms. Most of them fall into the <50 occurrences per million words frequency band.

Table 7. Sample of MWUs absent from the French-to-English HO and RC.

à partir des années	*droits de l'enfant*
au début des années	*droits fondamentaux*
au milieu des années	*en partenariat avec*
avant l'âge de	*et notamment*
avant même que	*états membres*
c'est pour cela que	*étudiant étranger*
c'est pourquoi	*gaz à effet de serre*
ce n'est sans doute pas	*grandes entreprises*
de ce point de vue	*qu'il s'agisse de*
de plus en plus de	*sans oublier*
de près de	*sur le devant de la scène*
depuis l'âge de	*un grand nombre de*
depuis une dizaine d'années	*un plus grand nombre*
d'ici et d'ailleurs	*visant à*

It is extremely difficult to identify the principles that guide the decision to include a MWU in one or the other category (e.g. headword vs. sub-entry, sub-entry vs. example). For example, why do the conjunctions of subordination *parce que* and *tandis que* have headword entry status in both dictionaries while *tant que* and *afin que* only feature in examples in the RC and as sub-entries in HO? Why are *en faveur de* and *en fonction de* sub-entries, but not *en forme de* or *en matière de*? Similarly, why are some 'noun + adjective' compounds listed as headwords (e.g. *langue française*, *art contemporain*, *monument historique*) while others appear as sub-entries (e.g. *chanson française*, *enseignement supérieur*, *recherche scientifique*)?[3]

Our study brings out the power of our corpus-driven methodology which has helped us identify a relatively high number of MWUs that are not included in the dictionary but would clearly deserve to be. Our results are all the more encouraging as the corpus used was very

687

small. A more representative monolingual corpus would undoubtedly yield a much larger number of potentially relevant units. Finally, our study is a prompt to reconsider the very notion of phrasal entry. Rather than rely on form-based criteria (such as the presence of a hyphen in the string), decisions should be made on solid phraseological criteria, i.e. the degree of syntactic fixedness and/or semantic compositionality, as well as other criteria such as corpus-attested frequency.

4. Enhancing the translation range

For this part of the analysis, in addition to the two dictionaries examined so far, we add a third online dictionary, the Larousse French-English dictionary (LA).[4] Once relevant multi-word units have been extracted from corpora, translation corpora can be used to identify frequent translation equivalents. In this section we illustrate the benefits that can be gained from using translation corpora by zooming in on two lexical bundles: *de plus en plus (de)* and *sur le plan (de)*. These chunks do not appear as headword entries: *de plus en plus (de)* is listed as a sub-entry of *plus* in the three dictionaries while *sur le plan (de)* is listed as a sub-entry of *plan* in LA and in example sentences in the *plan* entries in RC and HO.

For this second stage of our analysis, we decided not to rely on the full text search facility for a number of reasons. First, the full text search options of the three dictionaries require very different search strategies. While RC gives at-a-click access to the two dictionary sides, the French-to-English and English-to-French sides of HO and LA cannot be searched simultaneously. It is unrealistic to assume that dictionary users will systematically switch from French-to-English to English-to-French when looking up French MWUs. Moreover, the HO full text search is only available in the advanced options. Looking up *sur le plan* using the default settings in HO, for example, will not generate any result. Second, it is quite time-consuming to peruse the numerous hits generated by full text searches (e.g. 52 hits for *sur le plan* in RC). In view of this, we restricted our analysis to the translations of *de plus en plus (de)* and *sur le plan (de)* recorded in the *plus* and *plan* entries. Tables 8 and 9 give the range of translation equivalents for *de plus en plus (de)* found in the three dictionaries and in Label France respectively.

Table 8. Translation of *de plus en plus (de)* in the three bilingual dictionaries.

MWU	HO	RC	LA
de plus en plus + **verb**	**more and more** *il fume de plus en plus : he smokes more and more*		*les prix augmentent de plus en plus : prices are increasing all the time*
de plus en plus + **adj**	*de plus en plus difficile : more and more difficult* *de plus en plus chaud : hotter and hotter*	**more and more** *il fait de plus en plus beau : the weather's getting better and better* *aller de plus en plus vite : to go faster and faster* *de plus en plus drôle : funnier and funnier ; more and more funny*	**more and more, increasingly** *de plus en plus dangereux : more and more OU increasingly dangerous* *cela devient de plus en plus facile/compliqué : it's getting easier and easier/more and more complicated* *le ciel devenait de plus en plus sombre : the sky was growing darker and darker*
de plus en plus + **adv**	*de plus en plus loin : further and further*		**more and more** *de plus en plus souvent : more and more often*
de plus en plus de + **noun**			(+ countable) **more and more, a growing number of**

688

				de plus en plus de gens : more and more people, an increasing number of people
				(+ uncountable) **more and more**
				il y a de plus en plus de demande pour ce produit : demand for this product is increasing, there is more and more demand for this product
				Elle a de plus en plus de fièvre : her temperature is rising

Table 9. Translations of *de plus en plus (de)* in the Label France corpus.

increasingly	90 (54%)
more and more	33 (20%)
increasing/growing + N	19 (11%)
ever + comparative	10
comparative + comparative	6
with increasing frequency	3
Other	6
TOTAL	167 (100%)

Two translations of *de plus en plus* stand out in the bilingual dictionaries: *more and more* and 'comparative + comparative' (e.g. *hotter and hotter*). Surprisingly, one common structure, '*de plus en plus de* + noun', is only included in one dictionary (LA). For this structure too, *more and more* is suggested together with another equivalent, viz. *a growing number of*. As shown in Table 9, *more and more* is clearly not the dominant translation in the corpus: it only accounts for 20% of the translations. The addition of the 'comparative + comparative' category, which is also promoted in dictionaries, only brings this percentage up to 23%. The dominant translation in the corpus is *increasingly*, which is only mentioned in one dictionary (LA). In fact, an examination of the examples shows that *increasingly* is the dominant translation when *de plus en plus* modifies an adjective or a verb as in Examples 1 and 2. *More and more* is hardly ever used in those structures: the respective proportions of use of *increasingly* vs. *more and more* are 92% vs. 2% for adjectives and 88.5% vs. 11.5% for verbs. In fact, the majority of the occurrences of *more and more* in the corpus (73%) are found as translations of '*de plus en plus de* + noun', a structure that is under-represented in dictionaries (see Example 3). The corpus also brings out an interesting translation equivalent which is not mentioned at all in the dictionary entries, viz. '*ever* + comparative' (cf. Examples 4 and 5).

(1)   *En France, consommateurs et actionnaires semblent **de plus en plus** sensibles à la notion d'éthique.*
        *Consumers and shareholders seem **increasingly** aware of the notion of ethics.*
(2)   *En France, le cheval est **de plus en plus** employé par la police.*
        *In France, horses are **increasingly** being used by the police.*
(3)   ***De plus en plus de** gens s'essayent aux carnets de voyage.*
        ***More and more** people are trying their hand at travel diaries.*
(4)   *Des normes **de plus en plus** sévères limitent les émissions sonores des véhicules.*
        ***Ever harsher** standards are limiting the noise emitted by vehicles.*
(5)   *...les outils qui les aideront à déchiffrer le monde qui les entoure et à s'adapter à ses mutations **de plus en plus** rapides.*

*...the tools that will help them make sense of the world and adapt to its **ever more** rapid changes.*

The analysis of *sur le plan (de)* presents a similar picture. Tables 10 and 11 show that the phrase has a large number of translation equivalents and that there is relatively little convergence between the description given in dictionaries and the equivalents extracted from the corpus. One striking difference is that bilingual dictionaries favour phrasal equivalents, many of which are rare in the corpus. For example, *as far as x is concerned* and *as regards*, which are presented as the dominant translations in LA, do not occur at all in the corpus and other suggested equivalents, such as *from the point of view of* and Adv-*ly speaking,* only occur once. Although phrasal equivalents are also in evidence in the corpus, in particular those featuring the nouns *level* and *terms*, a third of the occurrences feature a single word equivalent in the form of an adverb (mostly in –*ly*), adjective or noun (see Examples 6-9). It is interesting to note that it is more common to translate *sur le plan de* by a single adverb in –*ly* than by the longer phrase - Adv–*ly speaking* - suggested in RC and LA.

Table 10. Translation of *sur le plan (de)* in the three bilingual dictionaries.

MWU	HO	RC	LA
***sur le plan de* + noun or adj**	*sur le plan de l'efficacité = from the **point of view** of efficiency, in **terms** of efficiency* *sur le plan politique/économique/personnel = from a political/an economic/a personal **point of view**, in political/economic/personal **terms***	*sur le plan du confort :* **as far as** comfort **is concerned** / **as regards** comfort *sur le plan moral/intellectuel :* mora**lly**/intellectua**lly speaking** /on the moral/intellectual **plane**	**as regards, as far as... is concerned** *sur le plan intellectuel: intellectual**ly speaking** sur le plan personnel : at a personal **level***

Table 11. Translations of *sur le plan (de)* in the Label France corpus.

Adv(-*ly*)/ Adj/N	20 (32%)
at/on Adj *level/at the level of* N	18 (29%)
in Adj *terms*	13 (21%)
with regard to N	2
Adv-*ly speaking*	1
from the point of view of	1
Other	7
TOTAL	62 (100%)

(6)   *Paris est un centre irremplaçable **sur le plan culturel.***
        *Paris is an irreplaceable centre **culturally**.*

(7)   *L'embryon ainsi obtenu devient alors une source de cellules souches neutres **sur le plan immunologique.***
        *The embryo thus obtained then becomes a source of **immunologically** neutral original cells.*

(8)   *L'Etat et les autorités locales proposent un ensemble de mesures incitatives tant **sur le plan fiscal que financier.***
        *The Government and the local authorities offer both **fiscal and financial** incentives.*

(9)   ***Sur le plan diplomatique**, les interventions de l'Europe dans le conflit yougoslave furent longtemps désordonnées.*

*In **diplomacy**, Europe's interventions in the Yugoslav conflict were for a long time uncoordinated.*

5. Conclusion

Our study raises the important issue of the place of phrasal entries in dictionaries. There is converging evidence that a large number of current web-based lexical searches concern multi-word strings (Lannoy 2010, Verlinde & Peeters forthcoming). It is therefore imperative to facilitate the access to longer strings and consider them not as second-rate citizens, but as fully-fledged entries on a par with single words. Our analysis shows that the addition of lexical bundles to bilingual dictionaries could contribute to improving their phraseological coverage. Coupled with efforts made by other scholars in the area of collocations (cf. Ferraresi et al. 2010), our results show that major advances are within reach as regards both the actual number of phrasal units included in bilingual dictionaries and their translation. However, a caveat is in order. The methodology we suggest is promising but the results will only be as good – or bad – as the corpus used. The Label France corpus is a good quality corpus but taken on its own, it is too small and too limited stylistically to be used as a basis for extensive revision of current bilingual dictionaries. Large, representative bidirectional translation corpora are an urgent desideratum for the future of bilingual lexicography.

Notes

[1] We have not computed precise inter-rater reliability scores but the degree of convergence between the two analysts was very high.

[2] The exact breakdown of the two sub-categories is as follows: HO: 34 MWUs not recorded at all and 18 MWUs exclusively found in the English-to-French side; RC: 26 MWUs not recorded at all and 40 MWUs exclusively found in the English-to-French side.

[3] The examples are taken from the HO dictionary.

[4] http://www.larousse.fr/dictionnaires/francais-anglais

References

Bentivogli, L. and E. Pianta 2002. 'Detecting Hidden Multiwords in Bilingual Dictionaries.' In A. Braasch and C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002. Copenhagen: CST*, 785–93.

Biber D., S. Conrad and V. Cortes 2004. 'If you look at…: Lexical bundles in university teaching and textbooks.' *Applied Linguistics* 25: 371–05.

Biber D., S. Johansson, G. Leech, S. Conrad and E. Finegan 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Ferraresi A., S. Bernardini, G. Picci and M. Baroni 2010. 'Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation.' In R. Xiao (ed.), *Using Corpora in Contrastive and Translation Studies*. Newcastle upon Tyne: Cambridge Scholars Publishing, 337–362.

Granger, S. and M.-A. Lefer (forthcoming). 'Enriching the phraseological coverage of high- frequency adverbs in bilingual dictionaries.' In K. Aijmer and B. Altenberg (eds.), *Advances in corpus-based contrastive linguistics*. Amsterdam & Philadelphia: Benjamins.

Granger, S. and M. Paquot 2008. 'Disentangling the phraseological web.' In S. Granger and M. Paquot (eds.), *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: Benjamins, 27–49.

Lannoy, V. 2010. 'Free online dictionaries: Why and how?' In S. Granger and M. Paquot (eds.), *eLexicography in the 21st century: New challenges, new applications*. Louvain-la-Neuve: Presses universitaires de Louvain, 173–181.

Murano, M. 2010. *Le traitement des séquences figées dans les dictionnaires bilingues français-italien, italien-français*. Milano: Polimetrica.

Rundell, M. 1999. 'Dictionary use in production.' *International Journal of Lexicography* 12.1: 35–53.

Sinclair, J. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, J. 2008. 'The phrase, the whole phrase, and nothing but the phrase.' In S. Granger and F. Meunier (eds.), *Phraseology. An Interdisciplinary Perspective*. Amsterdam: Benjamins, 407–410.

Sinclair, J. 2010 [2007]. 'Defining the Definiendum.' In G.-M. de Schryver, (ed.), *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Kampala: Menha Publishers, 37–47.

Todiraşcu A., U. Heid, D. Ştefănescu, D. Tufiş, C. Gledhill, M. Weller and F. Rousselot 2007. 'Vers un dictionnaire de collocations multilingue.' *Cahiers de linguistique* 33.1: 161–186.

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 the Ninth EURALEX International Congress, EURALEX 2000*, Stuttgart, Germany, August 8th–12th, 2000. Stuttgart: IMS, 855–861.

Verlinde, S. and G. Peeters (forthcoming). 'Data access revisited: the Interactive Language Toolbox.' In S. Granger and M. Paquot (eds.), *Electronic Lexicography*. Oxford: Oxford University Press.