A Methodology for Locating Translations of Specialized Collocations

Marie-Claude L'Homme, Nathalie Prévil and Benoît Robichaud
Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec) H3C 3J7 CANADA
{mc.lhomme,nathalie.previl,benoit.robichaud}@umontreal.ca

Abstract: This paper presents a method for locating translations of specialized collocations for the purpose of balancing lists of collocations in specialized resources. The main steps of the method are: 1. Identifying collocations in a source language for which translations are missing in a target language using an encoding based on lexical functions (Mek'čuk 1996); 2. Locating possible translations of the collocates in the source language in a bilingual resource; 3. Validating equivalents of the target language equivalents in a specialized corpus. In this paper, we focus more specifically on English and French collocations in the domain of the environment. We tested the method manually using 26 English terms and the collocations in which these terms appear and sought to locate translations of these collocations in French. Results show that this strategy for finding translations of collocations is promising and can help terminologists locate and validate collocates in a given language more quickly. With some adaptations, the method could be automated, but human validation is required, especially during step 3.

Keywords: terminology, collocations, terminological resource, environment, translation

1. Introduction

It is now recognized that adding collocations to terminological resources is extremely useful for certain types of users (translators, technical writers, or any user wishing to know how to insert a term adequately in a specialized text or finding out more about specialized usage). However, there are still few terminological resources that contain large sets of collocations. Some printed dictionaries are available for specific fields of knowledge: stock exchange (Cohen 1986) and business (Binon et al. 2000). A few electronic resources are also available. The Canadian term bank Termium (2018) includes collocates in some term records. IATE contains different kinds of "phrases and formulaic expressions" (Fontenelle 2014: 35). EcoLexicon (2018) lists verbal collocates in some of its entries and classifies them semantically. A resource called ARTES encodes collocations linked to scientific language (Pecman 2012). In our own resources – the DiCoEnviro (2018) and the DiCoInfo (2018) – collocations are listed along with other paradigmatic lexical relations (synonyms, antonyms, morphologically related terms, etc.) in English, French and Spanish (a few Portuguese terms are also included in the DiCoEnviro). Collocations are encoded and the meaning of collocates explained using the system of lexical functions (Mel'čuk 1996).

Collecting collocations from corpora and encoding them in specialized resources is time consuming and this might partly explain why few specialized resources list them. Methods were developed over the years (e.g. Kilgarriff and Tugwell 2001; Kilgarriff et al. 2012) to identify relevant word combinations automatically in running text, but combinations extracted must still be validated by lexicographers or terminologists.

This paper investigates a method for finding translations of specialized collocations and help terminologists locate valid collocations more quickly. Furthermore, the method is designed to balance lists of collocations between languages in multilingual resources. Often (and it is the case with the resource that we are currently compiling, i.e. the DiCoEnviro), entries are written in each language separately. Hence, collocations can be listed in a first language but their translation might not be available in another language. This is unfortunate since tools such as lexical functions can be used to access and retrieve equivalent collocations in different languages.

In this paper, after a brief overview of how collocations are described in our resources (Section 2) we present our method (Section 3) along with an experiment to test its usefulness in the context that we just described. We first tested the method manually in order to verify its potential for automation (Section 4). We examined 26 English terms in the field of the environment. Our test took into account 82 collocations for the 26 English terms. The identification and validation of equivalent collocations was carried out for French. Results are commented in detail in Section 5.

2. Collocations in the DiCoEnviro

As was mentioned above, collocations are listed in our resources and encoded using the system of lexical functions, LFs (Mel'čuk 1996). LFs take into account the syntactic structure of the collocation, its general and abstract meaning and, finally, the relation between the collocation and the argument structure of the keyword. For instance, assuming that the term *habitat* has the following argument structure: a habitat: ~ used by X, the collocation *occupy a habitat* would be encoded as follows:[1]

$$\text{Real}_1(habitat) = occupy$$

[1] Real_i represents collocates that denote the typical activity associated with the key word. In addition, Real_i is used when the key word is first complement (other LFs denoting typical activities are used when the keyword has another syntactic functions). Finally, the subscript "1" refers to the argument of *habitat* since it realizes the subject of *occupy*.

From the point of view of encoding, LFs have several advantages. First, they allow us to take into account different properties of collocations (syntactic, semantic and argument structure) and thus classify collocations accordingly. Furthermore, they are language-independent. Hence collocations in different languages that have the same meaning are encoded with the same LF.

$Real_1$(*habitat*) = *occupy*, *inhabit*
$Real_1$(*habitat*) = *peupler*

This kind of encoding can be used to establish equivalent relations between collocations in different languages without having to translate them one by one. The DiCoEnviro (and the DiCoInfo, for that matter) allows users to retrieve translations of collocations when they are available in the resource (L'Homme et al. 2012).

However, LFs can be quite difficult to decipher for users who are not familiar with the system. Therefore, different proposals were made to make them more transparent. In the online interface of our resources, LFs are explained with paraphrases that are superimposed on LFs and are designed to translate them in natural language. Our paraphrases are adapted from the proposal made by Mel'čuk and Polguère (2007). Hence, although collocations are encoded by terminologists with LFs, users only view the associated paraphrases in the online textual version (Table 1).[2]

Collocation	**LF**	**Explanation**
occupy a habitat	$Real_1$	The species uses a h.
inhabit a habitat	$Real_1$	The species uses a h.
peupler un habitat	$Real_1$	L'espèce utilise un h.
the habitat disappears	$FinFunc_0$	The h. ceases to exist
disappearance of a habitat	$S_0FinFunc_0$	Noun for "The h. ceases to exist"
loss of a habitat	$S_0FinFunc_0$	Noun for "The h. ceases to exist"
rétablir un habitat	$Caus_{@}De\_nouveauFunc_0$	Qqn ou qqch. remet un h. dans son état antérieur

Table 1: Encoding of collocations in the DiCoEnviro

3. The problem: imbalance between lists of collocations in different languages

When compiling a terminological resource, the different steps of the methodology are often carried out separately in different languages: specialized corpora are compiled for each language; terms are extracted and identified in each language; each corpus will be searched to retrieve relevant information for terms in that language, and so on.

This clear separation of the workflow in different languages is necessary to ensure that the information collected truly reflects usage in each language and not translation strategies. Furthermore, it prevents resorting to parallel corpora and thus translated texts for one of the languages.

It does, however, have a drawback. Indeed, corpora built may differ from one language to another. Hence, the content of these corpora might not be completely comparable leading to the addition of different kinds of information in a term record. This does not mean that the information given on terms is contradictory. However, the data recorded might not completely overlap when comparing entries in different languages. This problem can be observed in the lists of collocations compiled in the English and French versions of the DiCoEnviro (2018) as shown in Table 2 for the term pair *habitat* (En) and *habitat* (Fr).

habitat.1.en	**habitat.1.fr**
conserve $_{1a}$ ~ *preserve* $_{1a}$ ~ *protect* $_{1a}$ ~	*conserver* $_1$ *un* ~ *protéger* $_1$ *un* ~
	restaurer $_1$ *un* ~ *rétablir* $_{1b}$ *un* ~
alter $_{1a}$ ~ *degrade* $_{1a}$ ~	*dégrader* $_{1b}$ *l'*~
	améliorer l'~
	modifier l'~
the ~ *disappears* $_1$	
introduce $_1$ *... into a* ~	
	détruire l'~
inhabit $_{1a}$ ... *occupy* $_{1a}$ ...	*peupler* $_1$ *un* ~
conversation $_1$ *of a* ~ *protection* $_1$ *of a* ~	*conservation* $_1$ *d'un* ~ *protection d'un* ~
~ *regeneration*	*restauration* $_1$ *d'un* ~ *rétablissement* $_1$ *d'un* ~
	appauvrissement de l'~ *dégradation de l'*~
	amélioration de l'~
	modification de l'~
degradation $_1$ *of a* ~ *deterioration of a* ~	
disappearance $_1$ *of a* ~ *loss* $_1$ *of* ~	
recession of a ~	
	expansion de l'~ *extension de l'*~
change in a ~	
destruction of a ~	*destruction de l'*~

Table 2: Collocations recorded for the English term *habitat* and its French equivalent *habitat*

Besides the contents of the corpora, there might be other reasons for this imbalance. For instance, some lexical items might display a higher level of polysemy in one language than in another, leading to difficulties in locating relevant collocates for a specific term. The experience of terminologists might not be the same either and some of them might not spot relevant collocates as easily as others. All in all, we calculated the following discrepancies between English and French collocations in the DiCoEnviro (Table 3). We can see that most collocations do not have an equivalent one in the other language: between 66% and 77% depending on the language considered.

[2] Recently a new representation was added to the DiCoEnviro so users can visualize all lexical relations (including collocations) in the form of a graph (L'Homme et al. 2018). The graph shows both the explanation and the original lexical function.

	English collocations		French collocations	
With equivalent collocations	302	34%	315	23%
Without equivalent collocations	596	66%	1052	77%
Total	898		1367	

Table 3: Current imbalance between English and French collocations in the DiCoEnviro

4. Methodology

To identify and validate missing equivalent collocations in a target language we defined a method that comprises the following steps (we will illustrate them using examples taken from Table 2):

1. Locate a term in language A for which collocations are listed.

   e.g. *habitat* in English

2. Locate the equivalent in language B for this term in language A. Equivalents are stated explicitly in term records.

   e.g. *habitat* in English → *habitat* in French

3. Retrieve collocations of the term in language A.

   e.g. *habitat* in English:
   occupy a ~,
   introduce … in a ~,
   conserve a ~
   etc.

4. For each collocation, retrieve the lexical function used to describe it.

   e.g. *habitat* in English:
   occupy a ~ ($Real_1$),
   introduce … in a ~ (Labreal@$_1$),
   conserve a ~ (Caus@ContPredVer)
   …

5. For each collocation in language A, locate a collocation in language B that has the same lexical function. This step leads to two different situations:

   Situation 1: An equivalent collocation is listed in language B.

   e.g. *occupy a ~* ($Real_1$) → *peupler un ~* ($Real_1$)

   Situation 2: No equivalent collocation is found in language B.

   e.g. *introduce … in a ~* (Labreal@$_1$) → ?

The remainder of the method applies to Situation 2.

6. For each collocation in language A, take the collocate and search for its equivalents in an online bilingual dictionary.

   e.g. *introduce*

7. Retrieve the equivalents of this collocate from the bilingual dictionary.

   e.g. *introduce* → *introduire, initier, présenter, faire connaître*

8. Search each equivalent in language B and the equivalent of the keyword in language B in a specialized corpus.

   e.g. *introduce* → *introduire + habitat*
   initier + habitat
   présenter + habitat
   faire connaître + habitat

9. When a threshold number of contexts contain a term and a translation of the collocate, this can be considered a candidate translation of the collocation in language A.

   e.g. *introduce … in a habitat* → *introduire + habitat*

10. Encode the Language B equivalent collocate in the entry using the same LF as in English.

    e.g. *introduire + habitat*:
    Labreal$_{@1}$(habitat) = *introduire*

5. Manual validation of the method

We tested our method on a sample of terms and carried out part of the steps manually to assess its potential automation.

5.1 List of terms

We selected our keywords from a list of general English environmental terms collected for another experiment that consisted in identifying general environmental terms as opposed to terms that are linked to a specific subfield of the domain (Drouin et al. 2018). Among these 126 terms, 56 had French equivalents and 26 had recorded English collocations without French equivalents. The resulting list contains 26 terms[3] shown in Table 4.

animal.1.en → animal.1.fr	land.2.en → terre.4.fr
bird.1.en → oiseau.1.fr	oil.1.en → pétrole.1.fr
carbon.1.en → carbone.1.fr	plant.1.en → plante.1.fr
climate.1.en → climat.1.fr	polulation.2.en → population.2.fr
ecosystem.1.en → écosystème.1.fr	sea.1.en → mer.1.fr
effect.1.en → incidence.1.fr	species.1.en → espèce.1.fr
fish.1.en → poisson.1.fr	stratosphere.1.en → stratosphère.1.fr
forest.1.en → forêt.1.fr	temperature.1.en → température.1.fr
fuel.1.en → carburant.1.fr	threat.1.en → menace.1.fr
habitat.1.en → habitat.1.fr	tree.1.en → arbre.1.fr
impact.1.en → impact.1.fr	vehicle.1.en → véhicule.1.fr
land.1.en → terre.2.fr	waste.1.en → déchets.1.fr
ocean.1.en → océan.1.fr	water.1.en → eau.1.fr

Table 4: Term sample used for the manual validation

5.2 Extraction of collocations in English and French

For each term, we extracted all the lexical relations that were encoded as collocations from the English version of the database along with their lexical function. We

[3] Note that some lexical items are polysemous. We extracted them and their associated collocations separately.

proceeded to identify equivalent collocations in French based on the lexical functions. We obtained a table similar to that presented in Table 2 for each term.

We thus obtained for the 26 English terms:

- 180 English collocations;
- 98 English collocations with one or more French translation;
- 82 collocations without a French translation.

5.3 Searching for translations of collocates

We selected all 82 English collocations that did not have an equivalent in French. We extracted the collocates and searched for French translations in a bilingual resource. In this experiment, the translations were those produced by Google Translate.[4]

Equivalents labeled in Google Translate as "frequent" and "less frequent" were extracted (rare equivalents were not retrieved). Hence we obtained from 0 to 8 French equivalents for each English collocate (for a total of 211 equivalents). Examples are given in Table 5. The only English collocate that did not produce an equivalent was the verb *to power* (indeed, the only two French equivalents suggested for the verb by the bilingual resource were labeled as "rare").

Lexical function	**Collocation in English**	**Translations of collocate in French according to Google Translate**
Habitat.en.1 → habitat.fr.1		
FinFunc$_0$	*~ disappears*	*disparaître*
Labreal$_{@1}$	*introduce … into a ~*	*introduire, déposer, presenter*
S$_0$Degrad	*degradation of a ~*	*Degradation*
S$_0$Degrad	*deterioration of a ~*	*détérioration, degradation*
vehicle.en.1 → véhicule.fr.1		
Fact$_2$	*the ~ runs on …*	*fonctionner, passer, gérer, diriger, courir, tourner, marcher, faire fonctionner*

Table 5: Some equivalents suggested by Google Translate for collocates of *habitat* and *vehicle*

5.4 Validating translations of collocates

For each translation produced by the bilingual resource, we searched for contexts in a specialized corpus on the environment that contained both the French key words and the translations of the collocates.

The corpus was a large extract of the PANACEA corpus, an automatically compiled corpus that has a French component containing environmental texts (Prokopidis et al. 2012). The corpus is a compilation of web pages dealing with different topics related to the environment and covers various genres, i.e. official (governmental) reports, popularization, blogs, etc. (according to Bernier-Colborne 2014).

We searched for occurrence of both keywords and collocates using an in-house concordancer called *Intercorpus* (2018). The extract we used (approx. 231 Mb) represented about half the original corpus and was deemed sufficient to obtain representative results.

Contexts were searched using truncation for key words and collocates and a distance of 5 words or less was allowed between the two character strings. Contexts were considered relevant only if there was an actual link between the key word and the candidate collocate. For instance, the following context was considered relevant for *animal* and *vivre* (as a possible translation for *animal lives in …*):

> *C'est aussi parce que ces **animaux** **vivent** dans les forêts tropicales qu'il est important d'agir rapidement* (PANACEA/18159.txt)

However, the following two contexts were not considered:

> *Pour les plantes, il s'agit des conditions de sol et de microclimat propres à la station où elles **vivent**. Grâce à leur mobilité, les **animaux** peuvent utiliser divers types d'abris présents dans leur domaine vital.* (PANACEA/2051.tx)

> *L'ectofaune épizoaire, qui **vit** à la surface d'un **animal**, est une autre forme d'épifaune.* (PANACEA/ 41.txt)

6. Results

The number of valid contexts found in the reference corpus was recorded for each potential collocate as shown in Table 6.

habitat.en.1 → habitat.fr.1		
FinFunc$_0$	*~ disappears*	disparaître (28) -- --
Labreal$_{@1}$	*introduce … into a ~*	-- -- introduire (6), deposer (0), présenter (0)
S$_0$Degrad	*degradation of a ~*	dégradation (203) -- --
S$_0$IncepPred [MAN:différent]	*change in a ~*	modification (39) évolution (11) changement (8), variation (0)
vehicle.en.1 → véhicule.fr.1		
Fact$_2$	*the ~ runs on …*	fonctionner (35) -- passer (0), gérer (0), diriger (0), courir (0). tourner (0), marcher (0), faire fonctionner (0)

Table 6: Frequency of equivalents in the corpus (PANACEA)

[4] We first searched for equivalents in BabelNet (2018). However, for a small set of collocates no translation was available for French.

Three frequency categories were established: A) 20 and over occurrences; B) between 10 and 19 occurrences; C) up to 9 occurrences. It was assumed that valid collocates should appear with 20 occurrences and over in our reference corpus. It was also assumed that the last category would contain invalid translations. Results obtained in each category are detailed and commented below.

Among the results obtained, 53 French equivalents suggested by Google Translate were found in at least 20 contexts. These French equivalents were suggested for 40 source collocations (for a possibility of 82 that our test sample contained). In nearly all these cases, the translations were valid. This confirms our hypothesis according to which valid collocates would be found in that category.

For some source collocations, multiple valid translations were found although with varying frequencies of occurrence. For *conserve* (in *conserve an ecosystem*), the three French equivalents *protéger* (207), *préserver* (106), and *conserver* (20) were validated in the reference corpus and are all valid translations. The collocation *management of water* led to a slightly different situation. Three French equivalents suggested for *management* in the bilingual resource were found over 20 times in the reference corpus, i.e. *gestion*, *administration*, *direction*. However, the first one would qualify best as a valid equivalent and has by far the highest number of occurrences. Results obtained for *conserve an ecosystem* and *management of water* are reproduced in Table 7.

ecosystem.en.1 → écosytème.fr.1		
Caus@ContPredVer	*conserve an ~*	protéger (28), préserver (106), conserver (20)
		--
		évoluer (0)
water.en.1 → eau.fr.1		
PermFunc$_0$	*management of ~*	gestion (369), administration (41), direction (35)
		--
		management (0)

Table 7: Frequency of equivalents for the collocations *conserve an ecosystem* and *management of water* in the corpus (PANACEA)

In the 10-19 category of results, 19 equivalents suggested by Google Translate were found in the reference corpus (for 18 source collocations). In some of these cases, a valid French equivalent was already recorded in the 20 and over category. For instance, *pose* (in *pose a threat*) can be translated with *poser* (48). Among the other equivalents suggested by Google Translate, *créer* was also found in the corpus, but with only 10 occurrences.

In other cases, there was no French equivalent with over 20 occurrences in the corpus. However, a less frequent suggestion could be a plausible translation. For example, the only French equivalent proposed by our bilingual resource for *accumulation* (in *accumulation of carbon*) was *accumulation*. It was found only in 11 contexts but still remains a valid translation. Finally, the 10-19 category did contain invalid translations. For *grow* (*the plant grows*), the bilingual resource proposed *devenir* (among other translations). *Devenir* appeared in 11 contexts, but was never a valid translation for *grow* considered from the point of view of the environment. On the other hand, *croître* that appears in the same category is the valid translation. Results obtained for *pose a threat*, *accumulation of carbon* and *plant grows* are reproduced in Table 8.

threat.en.1 → menace.fr.1		
Oper$_1$	*pose a ~*	poser (48)
		créer (10)
		présenter (9)
carbon.en.1 → carbone.fr.1		
S0IncepPredPlus@ [@:lieu]	*accumulation of ~*	--
		accumulation (11)
		--
plant.en.1 → plante.fr.1		
Fact$_0$	*plant ~*	produire (31)
		croître (12), devenir (11)
		grandir (0), devenir (0), augmenter (0)

Table 8: Frequency of equivalents for the collocations *pose a threat*, *accumulation of carbon* and *plant grows* in the corpus (PANACEA)

The final category 0-9 contained 95 cases where no attestations of the equivalents suggested by our bilingual resource were found. In nearly all these cases, equivalents were invalid translations in the context of a collocation and could be discarded immediately. For example, *avilir* was suggested for a translation of the English verb *degrade (for degrade an ecosystem)*, but can certainly not be considered a valid translation in the context of *degrade an ecosystem*. Of course, it was never found in our environmental corpus.

In this category, 34 additional suggestions were made by the bilingual resource but were only found a few times in the reference corpus. Many of these suggestions were invalid translation, thus confirming our assumption about candidates with low frequencies. A few suggestions could correspond to valid translations, but did not occur very frequently (along with the key word) in our reference corpus. This was the case with two of the French equivalents suggested for *disturb* (in *disturb an ecosystem*), namely *déranger* and *troubler*. Results obtained for *degrade and disturb an ecosystem* are reproduced in Table 9.

ecosystem.en.1 → écosytème.fr.1		
Caus@Degrad	*degrade an ~*	degrader (26)
		se degrader (11)
		avilir (0)
Caus@NonFact$_0$	*disturb an ~*	perturber (42)
		--
		déranger (3), troubler (2), inquiéter (0)

Table 9: Frequency of equivalents for the collocations *degrade an ecosystem* and *disturb an ecosystem* in the corpus (PANACEA)

7. Discussion

In addition to the quantitative results commented in the previous subsection, the method yielded some qualitative

results that we did not anticipate when we started this project.

First, the corpus clearly showed that the same English collocate could translate differently in French. For instance, Google Translate suggested four different equivalents for the verb *disturb*: namely *déranger*, *inquiéter*, *perturber* and *troubler*. *Déranger* is preferred when *animal* is the key word (23 occurrences); while *perturber* is preferred with *écosystème* (42 occurrences). This, combined with the fact that some French equivalents were never found in the corpus along with given terms, shows that a validation with a specialized corpus remains necessary and is a strong aspect of the method.

Secondly, the corpus could reveal a clear preference for a given equivalent in the context of a collocation. For instance, *conserve* (in *conserve an ecosystem*) translates into French (according to the corpus) as *protéger un écosystème* (207 occurrences). Other equivalents are possible, but less frequent: *preserver un écosystème* (106 occurrences) and *conserver un écosystème* (20 occurrences).

The two observations above show that even collocations if specialized corpora often have a compositional meaning, usage influences the choice of a collocate and must be taken into consideration.

Thirdly, a human validation of the occurrences found in the corpus is necessary. For instance, some English collocates are highly polysemous and lead to French equivalents that are not synonyms or not even remotely semantically related. Our bilingual resource suggested the following French equivalents for the verb *occupy* (*the animal occupies …*): *habiter*, *occuper*, *prendre*, *remplir*. In this case, only *habiter* and *occuper* would be accurate translations. However, in the corpus, *animal* was found in contexts with *prendre* and *remplir* as shown below:

> *Si l'animal prend la fuite à quatre reprises, il est en danger de mort.* (PANACEA/381.txt)

> […] *car il s'agit souvent de plantes et animaux non-autochtones, ne pouvant pas remplir les fonctions qu'ils rempliraient dans la nature, ni ne pouvant remplacer les écosystèmes locaux détruits ou dégradés par les activités humaines.* (PANACEA/2659.txt)

The method also has some limitations. We listed four below:

- In a few cases, accurate translations were unavailable in bilingual resource. For instance, for the English term *warming*, the French equivalents suggested by Google Translate were *chauffage* and *échauffement*. The correct translation is the field of the environment is *réchauffement*. In order to correct this limitation, we could consider using more than one bilingual resource as long as they can be accessed freely.

- Some equivalents were suggested by our bilingual resource, but could not be found in the corpus. For instance, three French equivalents were suggested for the verb *thrive* (*prospérer*, *se developer*, *réussir*). None could be found along with specific terms of our set in the corpus.

- We hypothesized that valid translations of collocates would be found with 20 occurrences and over in the reference corpus. Although we confirmed this hypothesis to a large extent, for many source collocations (half of our sample), no equivalent suggested by the bilingual resource could be found with a sufficient number of attestations. This limitation could perhaps be corrected by using different resources: using a different bilingual resources or more than one bilingual resource, and increasing the size of our reference corpus.

- There was a non-negligible number of cases for which only a few occurrences of both key word and collocate could be found in the specialized corpus. Even if the corpus used (PANACEA) is very large, it covers many different areas of the environment. Perhaps a more focused and specialized corpus would increase the number of occurrences of some collocations. We could also use some of the corpora we compiled manually to increase the number of occurrences of keywords and collocates.

8. Conclusion and future work

In our opinion, our method produced a sufficient number of valid translations for our English collocations and could be used with some adaptations to complete other missing translations of collocations. Some suggestions were made above to correct some of its limitations (use of another more focused and specialized corpus, use of other bilingual resources, etc.). Our next step would consist in automating the method step by step. However, it seems that human validation cannot be avoided for this kind of work.

One strength of our method that we did not anticipate when we embarked on this project is that it allowed us to identify some clear preferences for some translations of collocates. It would draw terminologists' attention to phenomena that would be missed otherwise.

There are few directions that we can take in the near future. We could also apply this method the other way around for finding English translations for French collocations. We could also validate its potential for populating versions in other languages for which we have few collocations (our resource also has Spanish and Portuguese components). Looking back on the method, we could also extend it to all collocates in a source language and not exclusively to collocations for which there are no equivalents. This could lead us to find and fill other gaps in descriptions in different languages.

9. Acknowledgements

This work was supported by the Social Sciences and Humanities Research Council (SSHRC) of Canada. We would like to thank two anonymous reviewers for their useful suggestions on both the method and the paper.

10. Bibliographical References

Bernier-Colborne, G. (2014). Analyse distributionnelle de corpus spécialisés pour l'identification de relations sémantiques, In Actes de SemDis : enjeux pour la sémantique distributionnelle, Marseille, France, pp. 238-251.

Binon, J., S. Verlinde, van Dyck and J., Bertels, A. (2000). *Dictionnaire d'apprentissage du français des affaires*, Paris: Didier.

Cohen, B. (1986). *Lexique de cooccurrents. Bourse – Conjoncture*, Brossard (Québec) : Linguatech.

Drouin, P., L'Homme, M.C. and Robichaud, B. (2018). Lexical Profiling of Environmental Corpora. In Language Resources and Evaluation, LREC 2018, Myazaki, Japon.

Fontenelle, T. (2014). From Lexicography to Terminology: a Cline, not a Dichotomy. In Abel, A., Vettori, C. and Ralli, N. (Eds.). 16th Euralex Conference 2014. Proceedings, Bolzano, Italy, pp. 25-45.

Kilgarriff, A. and Tugwell, D. (2001). WORD SKETCH: Extraction, Combination and Display of Significant Collocations for Lexicography. In Proceedings of the Workshop on Collocations: Computational Extraction, Analysis and Exploitation, ACL-EACL 2001, Toulouse, pp. 32-38.

Kilgarriff, A., Rychly, P., Kovar, V. and Baisa, V. (2012). Finding Multiwords of More Than Two Words. In 15th Euralex Conference, Proceedings, Vatvedt Fjeld, R. and J. Matilde Torjuse (eds.). Oslo, Norway, 693-700.

L'Homme, M.C., Robichaud, B. and Leroyer, P. (2012). Encoding collocations in DiCoInfo: from formal to user-friendly representations. In S. Granger and Paquot, M. (Eds.). Electronic Lexicography. Oxford, Oxford University Press, pp. 211-236.

L'Homme, M.C. Robichaud, B. and Prévil, N. (2018). Browsing the Terminological Structure of a Specialized Domain: A Method Based on Lexical Functions and their Classification. In *Language Resources and Evaluation, LREC 2018*, Myazaki, Japon.

Mel'čuk, I. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In Wanner, L. (Ed.) Lexical Functions in Lexicography and Natural Language Processing, Amsterdam / Philadelphia: Benjamins, pp. 37-102.

Mel'čuk, I. and Polguère, A. (2007). *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. Bruxelles: De Boeck.

Prokopidis, P, Papavassiliou, V. Toral, A., Poch, M. Frontini, F., Rubino, F. and Thurmair, G. (2012). *Final Report on the Corpus Acquisition & Annotation subsystem and its components* (https://repositori.upf.edu/handle/10230/22514). Accessed 23 February 2018.

Pecman, M. (2012). Étude lexicographique et discursive des collocations en vue de leur intégration dans une base de données terminologiques. *JoSTrans. The Journal of Specialised Translation* 18.

11. Language Resource References

ARTES : Aide à la Rédaction de TExtes Scientifiques. http://www.eila.univ-parisdiderot.fr/recherche/artes/ Accessed 11 January 2018.

BabelNet. http://babelnet.org/ Accessed 5 January 2018.

DiCoEnviro http://olst.ling.umontreal.ca/dicoenviro Accessed 11 January 2018.

DiCoInfo. http://olst.ling.umontreal.ca/dicoinfo Accessed 11 January 2018.

EcoLexicon. Terminological knowledge base. http://ecolexicon.ugr.es/en/index.htm Accessed 12 December 2017.

Google Translate. https://translate.google.com Accessed 28 February 2018.

IATE. *Interactive Terminology for Europe* http://iate.europa.eu/about_IATE.html Accessed 19 February 2018.

Polguère, A. & E. Chièze. *Intercorpus* http://olst.ling.umontreal.ca/intercorpus/ Accessed 11 January 2018.

PANACEA. http://panacea-lr.eu/en/info-for-researchers/ Accessed 11 January 2018.

Termium. 2018. http://www.btb.termiumplus.gc.ca Accessed 11 January 2018.