16 Using a parallel corpus as a dictionary resource: studying idioms in an EN-GR parallel corpus

Sofia Trypanagnostopoulou, Janet DeCesaris

ABSTRACT

While parallel corpora have been widely used in several language applications, they have not been extensively exploited as a tool in bilingual lexicography. However, the amount of information that a parallel corpus offers to the lexicographer can considerably improve the quality of bilingual dictionaries. In the present paper we explore the potentials of a parallel corpus as a lexicographic resource, focusing on the examination of idiomatic expressions of the language pair English-Greek. Idioms are often underrepresented in lexicography, especially in bilingual dictionaries. As a first step to address this issue we decided to examine idiomatic expressions in a bilingual context. For this purpose we created a bilingual corpus using as a source the English TED talks and their Greek translations (http://www.ted.com/). After having compiled and aligned the corpus, we were able to extract the idiomatic expressions contained and to identify the main translation strategies used. Then, we used a sample of these idiomatic expressions in order to examine their lexicographic representation in bilingual English-Greek dictionaries. Our main focus was to investigate firstly if these expressions are included or not, what kind of information do the dictionaries provide (register, grammatical information, translation equivalents, notes, examples, etc.) and to propose improvements based on the data we found in our corpus.

Χρήση σωμάτων κειμένων ως λεξικογραφική πηγή:μελέτη ιδιωματικών εκφράσεων σε Ελληνο-αγγλικό σώμα κειμένων

Σοφία Τρυπαναγνωστοπούλου, Janet DeCesaris

ΠΕΡΙΛΗΨΗ

Ενώ τα παράλληλα σώματα κειμένων (ΣΚ) έχουν ευρέως χρησιμοποιηθεί σε διάφορες γλωσσικές εφαρμογές, δεν έχουν αξιοποιηθεί εκτενώς ως εργαλεία στη δίγλωσση λεξικογραφία. Το πλήθος, ομως, των πληροφοριών που προσφέρει ένα παράλληλο ΣΚ στο λεξικογράφο, μπορεί να βελτιώσει σημαντικά την ποιότητα ενός δίγλωσσου λεξικού. Στην παρούσα εργασία εξετάζουμε της δυνατότητες των παράλληλων ΣΚ ώς λεξικογραφικές πηγές, εστιάζοντας την προσοχή μας στη μελέτη των ιδιωματικών εκρφράσεων του γλωσσικού ζεύγουςαγγλικά-ελληνικά. Τα λεξικά -και ιδιαίτερα τα δίγλωσσα- συχνά δεν καλύτπυν επαρκώς τις ιδιωματικές εκφράσεις. Ως πρώτο βημα για να προσεγγίσουμε το ζήτημα, αποφασίσαμε να εξετάσουμε τις ιδιωματικές εκφράσεις σε ένα δίγλωσσο περιβάλλον. Για το σκοπό αυτό, δημιουργήσαμε ένα δίγλωσσο ΣΚ χρησιμοποιώντας ως πηγή τις αγγλικές ομιλίες TED και τις ελληνικές μεταφράσεις τους. (http://www.ted.com/). Αφού συντάξαμε και αντιστοιχίσαμε το ΣΚ, προχωρήσαμε σε εξαγωγή των ιδιωματικών εκφράσεων και εντοπίσαμε τις μεταφραστικές στρατηγηκές που χρησιμοποιήθηκαν. Στη συνέχεια χρησιμοποιήσαμε ένα δείγμα των εκφράσεων αυτών, προκειμένου να μελετήσουμε τη λεξικογραφική τους απεικόνηση στα δίγλωσσα αγγλο-ελληνικά λεξικά. Σε πρώτο στάδιο μελετήσαμε το κατά πόσο οι εκφράσεις αυτές εμπεριέχονται στα λεξικά ή όχι, τι είδους πληροφορίες παρέχονται (επίπεδο ύφους, γραμματικές πληροφορίες, μεταφραστικά ισοδύναμα, σημειώσεις, παραδείγματα, κλπ.) και προτείναμε βελτιώσεις με βάση τα δεδομένα που συλλέξαμε από το ΣΚ.

0 Introduction

Phraseology and especially idiomatic expressions are an important part of every language. They have strong cultural bonds andthey are considered to be one of the most difficult parts in language acquisition, both for native and non-native speakers. Undoubtedly, they require special attention in lexicography, because of their distinctive characteristics, and the lack of compositionality of their meaning. However, they are often underrepresented in dictionaries, especially in bilingual ones. In our attempt to address the issue, we will examine how the use of a parallel a parallel corpus as a lexicographic resource can improve the representation of idiomatic expressions, in bilingual English-Greek dictionaries.

The lexicographic representation of phraseology is usually problematic due to various factors. One of the first difficulties is the arbitrary borders among the various types of phraseological units. Despite the long academic discussion about its terminology and classification, to date there is no standard taxonomy of phraseology. Another difficult issue for the lexicographer is the process of selecting which idioms are important to be included in the dictionary and which are not. The position of the idioms in the macrostructure and microstructure of a dictionary has always been a matter of discussion, since accessibility to phrases in the dictionary is not straightforward, yet is clearly of primary importance from the perspective of users.

Equally significant is the issue of variation of the idiomatic expressions in terms of grammar, syntax and morphology. As mentioned by Moon (1996), while corpus evidence shows that the form of idiomatic expressions is not fixed, this variability is under-represented in dictionaries. Also, as idioms are used both in written and oral speech, and in various contexts, they might be used in several registers. This usually creates problems for non-native speakers, who are some of the main addressees of bilingual dictionaries. Last but not least is the issue of translation equivalence. The translation of idiomatic expressions in bilingual dictionaries is challenging. Idioms often have a metaphorical base, and thus they have a strong connection to the culture associated with a particular language. According to Mona Baker (1992:68-69), significant difficulties for translators include the lack of equivalence in the target language in the case of culture specific idioms or expressions, the use of an idiom in a different context in the target language, and finally the fact that an idiom may be used in the source language in both a literal and idiomatic sense at the same time.

Despite the importance of the English language in Greek society, there are limited options for users of English-Greek dictionaries and even fewer options when considering dictionaries of phraseology. The existing dictionaries are used both for decoding and encoding purposes, while English is used as an intermediate language for other users. Speakers of third languages (i.e. Arabic) might use an English-Greek dictionary due to a lack (or the poor quality) of an Arabic-Greek dictionary.

1 The role of parallel corpora in bilingual lexicography

A lexicographic study today is almost unthinkable without the use of real language data. While parallel corpora have been widely used in several fields such as translator training, machine translation, contrastive linguistics, and various language applications, they have not been extensively exploited as a tool in bilingual lexicography. However, we believe that the amount of information that a parallel corpus offers to lexicographers can considerably improve the quality of bilingual dictionaries.Firstly, by having access to the source texts, lexicographers can see the actual use of an idiom and can extract information about variation, frequency, morphology, usage pragmatic or cultural correlations, etc.In addition, a parallel corpus always offers a translation equivalent in target language. A parallel corpus consists of texts that have already been translated; therefore, we can find there the solutions that the translators have proposed to real translation problems. For this reason, we consider that the equivalents proposed by the translators could be good candidates to be included in a bilingual dictionary. Finally, a parallel corpus is a rich source of phrases, which can be used as examples of use.

Of course, we should keep in mind that the translations provided in the corpus are highly context-dependent and this means that they might be a solution which can work only in the specific context; therefore, they might not be appropriate candidates as dictionary equivalents. Aparallel corpus usually gives us only a single version of a translation, and the quality of the translator determines the quality of the translation. One solution to this would be to include various versions of translations of the same texts, which is impractical. Salkie (2008) suggests the following: (a) know which is the source language; (b) be sure that the translator was a skilled professional; (c) guarantee that the translation was checked; and (d) know that the translation was published by an organization which takes quality seriously. Additionally, we should make sure to use a balanced corpus of contemporary language. Parallel corpora diminish the role of human intuition; however, this does not mean that lexicographers should exclusively rely on them. As Sinclair (1985) points out, personal introspection is a factor that will inevitably play a big part in the decisions made by lexicographers – in evaluating evidence rather than creating it.

2 Exploratory Study

As a first step to address this issue, we decided to examine idiomatic expressions in a bilingual context and to compare our findings with two English-Greek dictionaries. For this purpose we created a parallel corpus of English texts and their Greek translations. After having compiled and aligned the corpus, we were able to extract the idiomatic expressions contained and to identify the main translation strategies used. Then, we used a sample of these idiomatic expressions in order to examine their lexicographic representation in bilingual English-Greek dictionaries, and compared them with the information provided by our corpus.

2.1 Corpus description

We used as a source for our parallel corpus a number of English TED talks and their Greek translations (http://www.ted.com/[1]).We chose to collect TED talks because of the following characteristics and the advantages they offer. They are of free access in the internet and without copyright, while they are provided both in audiovisual (sound and subtitles) and in text form (original script and translations). There is a certain guarantee of the quality of the translations, as they provide the name of the translator and the reviser. Because the texts were spoken, they are likely to contain several, and possibly many, multiword expressions. Despite the oral language, the register it not very low and this means that and expressions used are appropriated to be included in a bilingual dictionary. Finally, a corpus from these texts would be very wide in terms of language use and vocabulary as the subjects of TED talks vary from art, technology, design, politics and everyday life stories. Our corpus is consisted of 30 TED talks with their translation into Greek. The total number of English words is 35.503 (average 1183.43 words per text) and the total of Greek words is 33.033 (average 1101.1 words per text).

2.2 Extraction and classification of phraseological Units

We extracted the phraseological Units manually and we classified them based on the taxonomy proposed by Moon (1998) into: **Anomalous collocations**, which are problematic in terms of

[1] TED (Technology, Entertainment and Design) is a global set of conferences owned by the private non-profit Sapling Foundation, under the slogan "ideas worth spreading”. It started out (in 1984) as a conference bringing together people from three worlds: Technology, Entertainment, Design.

lexicogrammar and include ill-formed collocations, cranberry collocations, defective collocations and phraseological collocations (*by and large, of course*); **Formulae**, which are problematic in terms of pragmatics and include simple formulae, sayings, proverbs and similes (*I'm sorry to say, an eye for an eye, enough is enough);* and **Metaphors,** which areproblematic in terms of semantics and include transparent metaphors, semi-transparent metaphors and opaque metaphors: (*behind someone's back, kick the bucket).*

From the total of 220 phraseological units extracted, 52 are anomalous collocations, 70 formulae and 76 metaphors. Of course the margins between the categories are not very clear; so many times a phraseological unit can be classified to more than one group. For the purpose of the present study, we focused on idiomatic expressions, namelythe metaphorical-based phraseology, which includes mainly the third category (metaphors), but also some formulaeor anomalous collocations that have metaphorical meaning. More specifically we examined: 76 metaphors (*change my mind)*, 21 formulae (*come to the conclusion*) and 1 anomalous collocation (*lo and behold*). On the other hand, some phrases appeared more than one time in the corpus, mainly in variations. So, in total we examined 119 phraseological elements, including the multiple entries.

In the corpus we identified the following translations strategies: (1) Equivalence: (same meaning – same form) even though it is difficult to have perfect equivalence, we marked as "equivalence" the cases where the translator uses an idiom of the target language with the same meaning as the one of the original phrase. From the total of 119 phrases found in the corpus, 63 were translated by an equivalent expression: *with open arms* με ανοιχτές αγκάλες. (2) Paraphrase: (same meaning – different form) when the translator transfers the meaning, but uses a lexeme in the TL which is not idiomatic. In our corpus, 28 phrases were translated by paraphrase: *They were dressed to the nines* Είχαν εξεζητημένο ντύσιμο. (3) Literal Translation: (different meaning-same form) when the translator uses parallel lexemes in the TL, but the meaning is not the same and the final result lacks functionality. In our corpus, we recorded 11 cases of literal translation: *little yellow brick road* - κίτρινος πέτρινος δρόμος. (4) Omission: when the translator totally omits the phraseological unit. Four of the 119 phrases were omitted in their Greek translation: And then I went into geology, "*rocks for jocks*." This is easy. - Έπειτα μπήκα στη γεωλογία, αυτό ήταν πιο εύκολο.

We can observe that the corpus offers translation equivalents for almost half of thecases, while we have fewer cases of paraphrase and only a few literal translations and omissions. This is very important, because it shows the good quality of the translations, which means that the equivalents could be good candidates to be used by the lexicographers.

2.3 Comparison with dictionary representation of phraseological units

After extracting the phraseological units from the corpus, we examined their lexicographic representation in two bilingual English - Greek dictionaries: Oxford English-Greek Learner´s Dictionary (D. N. Stavropoulos & A. S. Hornby) and Collins English-Greek Dictionary. The main criteria for selecting the dictionaries were their overall satisfactory quality, the adequate amount of information they provide and their popularity in terms of usage.

Firstly, we examined if the expressions found in the corpus are included or not. The table below shows the phraseological coverage of the two dictionaries:

DICTIONARY	NOT INCUDED	INCLUDE SIMILAR EXPRESSION	INCLUDED	INCLUDED IN VARIATION	INCLUDED WITH DIFFERENT MEANING
OXFORD	43	5	43	6	1
COLLINS	49	3	34	6	6

As we can see from the table above, the Oxford dictionary does not include 43 out of 98 idioms found in the corpus, while it contains5 similar – but not exactly the same – expressions. For example, we cannot find the idiom: *need to take time to*, but we can find the phrase:*take your time*. In these cases, even though the expression provided by the dictionary is similar, it is not certain that the user will be able to understand it and use it properly. On the other hand, 43 out of 98 idioms are included in the corpus, while 6 are included in a variant form. For example, we cannot see the phrase: *stops dead in her tracks*, which we found in the corpus, but only the idiom: *stop dead*. Finally, one expression is included, but not with the same meaning as in the corpus: we can find the idiom: *at the bottom*, but only in its literal meaning (στοκάτωμέρος) and not in the metaphorical meaning as used in the text (deeply - καταβάθος). Likewise, in Collins dictionary 49 out of 98 idioms are not included, 3 are given in a similar expression, 34 are included, 6 include a variation and 6 are given with a different meaning.

What we can conclude from these numbers is that there is a rather low percentage of dictionary coverage in both dictionaries Of course we should keep in mind that it is not always possible or useful to include all the idiomatic expressions of a word in a bilingual dictionary, especially in a paper edition, where the limitation of space restricts the lexicographic options. However, among the idioms that we found in the corpus that are not present in the dictionaries, there are many phrases which have a discoursal function: *I didn't get the memo, take a second, you betcha*.

These types of idioms are problematic in terms of pragmatics and since they are frequently used, especially in oral speech, it is, we believe, very important to include them in a bilingual dictionary. Moreover, some of the idioms not found in the dictionaries, such as *little yellow brick road, Beam me up, Scotty* have strong cultural connotations. The user would not be able to understand and use these expressions without previous knowledge, thus it would be useful to get this information from the dictionary.Finally, some of the expressions have a highly opaque meaning: *It was off and running, lo and behold, meet that bar*. The user will not be able to find the meaning of such expressions by looking in the dictionary the meaning of their parts.

Our next step was to investigate what kind of information the dictionaries provide in the microstructure of their entries (register, grammar, usage label, translation equivalents etc.). in comparison with the data we collected from our corpus.

In some of the expressions Oxford uses the label (*idm*) for idiom, while Collins uses sometimes the label (*fig.*) for figurative. However, the criteria on this are not very clear. It seems that there is no standard categorization of phraseology in neither of the dictionaries. Regarding the register, Oxford gives some usage labels, such as (καθομ.) for Καθομιλουνένη, colloquial language. Collins, on the other hand, even though in the abbreviation list (in the introduction of the dictionary) includes various labels, such as *Fml* for formal, *Hum* for humorous, in the sample we examined we didn´t find any of them. In a bilingual context -where there is not always equivalence between the source and target language - information about the register is very important, as the user should be able to know how and where to use an expression. Collins dictionary provides a lot of grammatical information (*in the wake of* ως επακόλουθο (+GEN), *To give sb a bad name* δυσφημώ κν), while Oxford includes only a few (*can/can't/could(n't) help doing sth* μπορώ / δεν μπορώ να (αποφύγω να κάνω κάτι). Of course, this kind of information mentioned above can be also incorporated in the examples of usage. Both dictionaries include a lot of examples of idiomatic expressions, which are accompanied by their Greek translation: (Oxford) *The affair is no longer in my hands* η υπόθεση δεν είναι πλέον στα χέρια μου / στη δικαιοδοσία. (Collins) *we were accused of giving the country a bad name overseas* Κατηγορηθήκαμε ότι δυσφημήσαμε τη χώρα στο εξωτερικό.

If we compare the equivalents proposed by the dictionaries with the corpus, we observe that in many any cases the translation is different. Even though the equivalent of the dictionary is not bad or wrong, the translation might offer an additional equivalent, which could be good candidate to be included in the dictionary: *In the wake of* => (Corpus) στη συνέχεια, (Oxford) από πίσω, καταπόδι, μαζί, ως επακόλουθο, (Collins) ως επακόλουθο (+GEN). However, there are cases where the corpus can give an idiomatic equivalent, while the dictionary only gives an explanatory equivalent: *stops dead in her tracks* => (Corpus) κοκάλωσε, (Oxford) σταματάω απότομα. Sometimes the equivalent of the dictionary lacks naturalness. In this case the equivalent of the corpus, can give a solution: *be/get out of control* => (Corpus: It's out of your control) Δεν μπορείς να το ελέγξεις, (Oxford) αποχαλίνουμαι. In other cases, the corpus translation is closer to the register of the original than the proposed dictionary equivalent: *spend a lot of time* => (Corpus: spend so little time) περνάμε τόσο λίγο χρόνο, (Oxford) διαθέτω / αφιερώνω πολύ χρόνο για κτ. As mentioned above, in many cases the corpus provides different meanings than those of the dictionary. This automatically results in different equivalents: *at the bottom* => (Corpus) μέσα μου, (Oxford) κατά βάθος, (Collins) στο κάτω μέρος. And the same can also be seen in the variations of the expressions as different forms of an idiom could have different equivalents: *something is going wrong* => (Corpus) κάτι δεν πάει καλά, (Oxford: go wrong) α. Ακολουθώ λάθος κατεύθυνση, β. Αποτυγχάνω γ. Στραβώνω, χαλάω.

3 Conclusions

This paper is the first step of a larger study of corpus-based lexicographical representation of phraseology. More specifically, we used a small amount of data in order to see if interesting conclusions are possible and to examine and assess the suggested methodology. Our first results showed that the corpora can provide useful information about idiomatic expressions which is not included in the dictionaries we examined.

One of the important results was the rather low coverage of idiomatic expressions in the two dictionaries in comparison to our corpus findings. Before rushing to conclusions, we should keep in mind the small size of our corpus, thus the limited sample of phraseological units we examined. And of course,as mentioned above, it is not possible – or even useful – for a dictionary to include every possible phraseological unit of an entry-word. However, the very low percentage of dictionary coverage gives an indication about the quality of the dictionary in terms of phraseology representationand further research seems to be necessary. The limited use of real linguistic data is characteristic of many bilingual dictionaries, as opposed to monolingual dictionaries, in which corpus data has been extensively used for over thirty years now. However, as parallel corpora offer texts that have actually been translated, we can assume that they are likely to contain elements that one might look up in a dictionary. Therefore, it would improve the representation of phraseological coverage if the lexicographer would rely on a parallel corpus to select which idioms to include in the dictionary.

Another interesting point of this study is the fact that the corpus can reveal the variation of phraseological units. A significant number of the idioms we extracted from the corpusappeared in adifferent form than the formgivenin the dictionaries. The lexicographical challenge is therefore to examinehow this variation could be better represented in bilingual dictionaries. Of course, it is not possible to include all the variations of an expression in a bilingual dictionary, however, the corpus can give us a clue about the most important or the most frequently used forms. Furthermore, from the small sample of the phrases we examined, we observed that in many cases the corpus gave us different or even more successfultranslation equivalents than the ones proposed by the dictionaries.In some cases the dictionary provides an explanatory equivalent, while - as found in corpus- there is a functional equivalent.As we have seen in our corpus, almost half of the phrases were translated by an equivalent expression. These translations could be used from the lexicographer as functional equivalents, which can transfer not only the meaning, but also the idiomaticity of the original. Of course, we should keep in mind that translations are context specific and proposed by the translator with an audience in mind, while the dictionaries should include information that is not context-bound. However, a larger corpus could cover as many occasions as possible, and could give many equivalents to the lexicographer. This would automatically diminish the role of human intuition; the equivalents would be more functional and natural.

As mentioned above, the present study used a small amount of language data to extract and examine idiomatic expressions in a bilingual context. In order to achievemore accurate results, we should enrich our parallel corpus, including texts of literature, and audiovisual material (movies, documentaries, television series etc.), or the European Parliament corpus. Further study can be carried out on how this model could be used not only in general purpose bilingual dictionaries, but also in specialized dictionaries, and more specifically on the representation of phraseology in terminological resources.

4 References

[1] Baker, M. (1992) *In other words : a coursebook on translation.* London: Routledge.

[2] *Collins English-Greek Dictionary* (2nd ed., 2006).London: HarperCollins.

[3] Moon R. (1996). *Data, Description, and Idioms in Corpus Lexicography*.In Euralex '96 Proceedings. Gothenburg: Göteborg University, 245-256.

[4] Moon, R. (1998).*Fixed Expressions and Idioms in* English.A Corpus-Based Approach. Oxford: Oxford University Press.

[5] Salkie, R. (2008). "How can a lexicographer use a translation corpus?" In *Proceedings of*

the International Symposium of Using Corpora in Contrastive and Translation Studies.Zhejiang University, Hangzhou.

[6] Sinclair, J. (1985), Lexicographic Evidence. in: Ilson R. (ed.), *Dictionaries, lexicography and language learning*, ELT Documents 120, Oxford: Pergamon Press, 61—68.

[7] Stavropoulos, D. N. & A. S. Hornby. (2007) *Oxford English-Greek Learner's Dictionary*. Oxford: Oxford University Press.

Janet DeCesaris

Associate Professor of Translation and Interpreting
InstitutUniversitari de Lingüística Aplicada - UniversitatPompeuFabra
RocBoronat, 138 - 08018 Barcelona
E-mail: janet.decesaris@upf.edu

Sofia Trypanagnostopoulou

PhD Candidate - UniversitatPompeuFabra
RocBoronat, 138 - 08018 Barcelona
E-mail: sofitrip@yahoo.gr