

23 Compilation and analysis of parallel corpora in the extraction of terminology, retrieved for translation teaching purposes

Maria Matsira

ABSTRACT

Translator training is considered to be a difficult process. The teaching of basic translation theories to translation students, the familiarization with translation strategies, along with the supply of a rudimentary terminology, which will help them to accomplish the task of translation, only go so far in equipping them with the basic skills a translator needs.

Corpora (learners' corpora, LSP corpora, dictionaries corpora) serve a whole range of linguistic activities, including the extraction of terminology for translator training purposes as well as the extraction of translation phraseology which is typical in translated texts.

The interest of this paper is mainly focused on the compilation and the analysis of a parallel corpus of popular science texts, i.e. articles appearing in a wide circulation popular science magazine and their translations.

The idea is based on the fact that translation teachers use articles of this genre as teaching material, and the goal of this study is to introduce a concrete methodology for extracting terminology for translation teaching purposes, which can be easily understood and then used by both Greek translation teachers and students.

Συγκέντρωση και ανάλυση παράλληλων σωμάτων κειμένων κατά την εξαγωγή ορολογίας, για σκοπούς διδασκαλίας της μετάφρασης

Maria Matsira

ΠΕΡΙΛΗΨΗ

Η εκπαίδευση των μελλοντικών μεταφραστών είναι απαιτητική υπόθεση. Η διδασκαλία των βασικών μεταφραστικών θεωριών, των στρατηγικών που ακολουθεί ο μεταφραστής, καθώς και η χρήση μιας στοιχειώδους ορολογίας, που θα τους βοηθήσει να αντιμετωπίσουν στις μεταφραστικές δυσκολίες ενός κειμένου, θα λέγαμε ότι δεν συνιστούν το σύνολο των στοιχείων εκείνων που κάνουν ένα μεταφραστή ικανό και ανταγωνίσιμο.

Τα σώματα κειμένου χρησιμοποιούνται για πλήθος γλωσσολογικών εφαρμογών (διδασκαλία ξένης γλώσσας, σύνταξη λεξικών, κλπ.), συμπεριλαμβανομένης και της αλίευσης όρων για σκοπούς που αφορούν κυρίως τη διδασκαλία της μετάφρασης.

Το ενδιαφέρον αυτής της εργασίας εστιάζεται στη σύνταξη και την ανάλυση ενός παράλληλου corpus εκλαϊκευμένων επιστημονικών κειμένων –άρθρων εκλαϊκευμένου αγγλικού περιοδικού και των μεταφράσεων τους στα ελληνικά.

Η εν λόγω ιδέα προήλθε από το γεγονός ότι οι καθηγητές μετάφρασης χρησιμοποιούν συχνά άρθρα αυτού του είδους, ως εκπαιδευτικό υλικό και σκοπός αυτής της μελέτης είναι να εφοδιάσει καθηγητές και φοιτητές με μία εύχρηστη μέθοδο εξαγωγής ορολογίας από σώματα κειμένου (corpora) .

0 Introduction

The initial idea for this paper emerged from the author's own experience as a BA student in the Department of Foreign Languages, Translation and Interpreting, Corfu, Greece. There, in the two first years of their studies, the students become acquainted with the principles of translation practice, starting with short general interest texts and going on to more specialized texts in the third and fourth years of university.

The texts used in the translation teaching process of the two first years are mainly newspaper and magazine articles, or passages taken from websites, etc.; whereas the texts used in the two final years are characterized by a considerable amount of technicality: lawsuits, court decisions, and EU directives for the law translation module; articles in financial publications, or articles defining and analyzing complex financial notions for the economic translation module; and finally manuals for various devices, machines and gadgets, academic articles on physics, chemistry and biology, and articles from popular science magazines for the module of technical translation.

The teaching procedure aims to enable students to acquire all the necessary material that will help them to accomplish the translation task as accurately as possible. The research includes searching in dictionaries, encyclopedias and other sources, as well as searching in web engines. The data collected from this procedure can also be used later on, during the exams, in handy forms, such as glossaries and notes.

However, no matter how appealing this appears to be as a teaching process, we would argue that this is insufficient on its own to teach students how to translate. It needs a more solid systematic foundation. Sample texts used for translation teaching purposes are regularly replaced by new ones to keep pace with technological developments.

The challenge for the translation teachers is twofold: on the one hand to provide their students with a method to retrieve terminology, using the limited resources they have at their disposal; and on the other hand, to be confident that this method can be reliable enough and its outcome valid.

This challenge can be met by the use of corpora and more precisely parallel ones. These can easily serve as valuable databases, from which useful information could be extracted by students. They can be easily built up by anyone and also analyzed if somebody has the appropriate software program, and as Bowker and Pearson explain in their book *Working with Specialized Language, A practical guide to using corpora*: "A corpus can provide you

with both linguistic and conceptual information” and “you can consult a parallel corpus in much the same way as you consult a bilingual dictionary, but a corpus will provide more collocational and stylistic information than a dictionary”. [11]

Below, we are going to see how a small bilingual, parallel, English-Greek corpus that is constituted from popular science articles, can contribute to the extraction of terminology for translation teaching aims. To this purpose, the bilingual, parallel corpus was compiled from scratch; two software programs were used for the alignment and the statistical analysis; and three approaches were tested for extracting candidate terms.

1 Defining Terminology

Let us now turn to consider definitions of 'terminology'. Researchers have used a number of definitions:

“Terminology extraction (TE) tasks deal with the identification of terms which are frequently used to refer to the concepts in a specific domain.” [20]

According to most researchers, there are some standard methods for automatic terminology extraction [20]:

- i. Term extraction via morphological analysis: POS tagging and shallow parsing
- ii. Term weighting with statistical information.
- iii. Term extraction via syntactical analysis, which is primarily based on the first method and it definitely requires before POS tagging in order to be accomplished.

The automatization of the terminological extraction, however, still faces serious problems, like (1) recognition and identification of complex terms (2) identification of the terminological nature of a lexical unit (4) appropriateness of a terminological unit to a specific domain. [3]

Some researchers, however, developed an alternative way to cope with the problem of term definition, as appears in the following phrase: “[...] we give back to the user an active role in the extraction progress. That is instead of encoding a static definition of what might or might not be a term, we let the user specify his own. ” [19]

Another question that arises within terminology extraction is the issue of single-word terms (mono-lexical terms) and multi-word terms (poly-lexical terms) [20, 21]. Terminologists think preferentially of nouns when they consider domain-specific concepts [7, 15]. However, these nouns can sometimes be NPs that are constituted by several part of speech combinations,

such as Noun-Noun collocations or Noun-Adjective collocations [13]. These can also be terminologically relevant, since “in general language, many collocates in noun-verb or noun-adjective collocations have a collocational meaning, i.e. are not understood in the same meaning as in contexts outside the collocation.” [13]

2 Corpus Design and Alignment

The corpus used in this study was initially compiled for the needs of the author’s postgraduate research, for the MPhil (B) Corpus Linguistics in the University of Birmingham, during the academic year 2006-2007. The material for the building up of this corpus came from six past issues of Scientific American (November 2006-April 2007)¹, and included the original version in English and the translated one in Greek. The whole corpus is constituted from 92 articles (46 English and 46 Greek). The size of the English corpus is approximately 138.684 words (tokens) while the size of the Greek corpus is approximately 167.739 words (tokens).

The material was collected in two ways: the English part through the Internet and the Greek part through the laborious task of scanning –since there was no way to get access to the electronic issues of the Greek Scientific American. For the scanning a Greek OCR (Abby Fine Reader 8.0) was used; some of the editing had to be done manually.

After the data collection and editing, the next step was the alignment of the two corpora, which would enable us first to compare and then to attempt to extract candidate terms. For the alignment, a software program: Multiconc [16], created in the University of Birmingham, was used. Minmark 2.0 -a Multiconc tool- aligned the texts on paragraph level:

“It is difficult to employ this approach at sentence level since a skilled translator may well translate one sentence by two, or two by one, three by two, and so on. This is the central problem of text alignment.” [16]

Due to the fact that our source material was a popular science magazine, we often came across phenomena, such as omission or adaptation in the translated text, which were purely decisions of the Greek editor, always with regard to the target audience. However, Multiconc

¹ The selection of the articles was made on the basis of the Greek translations, since the problem was the difficulty of collecting data for the Greek corpus. Once enough material was collected for the Greek corpus, the original articles in English were traced and formed the English corpus, but they generally appeared 2-3 months earlier than their translations.

can provide parallel concordances at sentence level, or, when no match appears at sentence level, the user can select a paragraph-level assignment.

The corpus was divided into sub-corpora according to topic. The topics and sub-topics are indicated in the header information accompanying each article in Scientific American. Therefore, taking that into account, we ended up with 7 sub-corpora: Planetology/Cosmology, Psychology, Physics, Technology, Energy/Climate/Geology, Medicine and Biology/Anthropology. However, we should mention here that the subcorpora contain different numbers of articles, since the collection of the material was made randomly, the only criterion being their appearance in the issues of Scientific American (Greek edition) between November 2006 and April 2007.

This division into sub-corpora was made in order to facilitate terminology extraction; that is, all articles dealing with a given area were gathered into one sub-corpus, so as to help the researcher to collect terms that belong to the same scientific field and organize them afterwards into glossaries or specialized lexicons.

3 Corpus Analysis and Terms Extraction

In the analysis of the corpus, Wordsmith 3.0 was used to extract wordlists, keywords and concordances. Wordsmith 3.0 is not the latest version (the latest is Wordsmith 4.0) of this software, however this is the only one that works with Greek, which is why it was used in this study.

The methods we followed here, however, are not purely automatic –maybe one could call them semi-automatic- since the means we had at our disposal for Greek were somewhat limited. For minor languages² like Greek, taggers are limited [21], and consequently the statistical analysis can only be done in terms of frequency.

“The relative frequency of a lexical unit in two different corpora is strongly linked to the importance of the unit in the corpora. The more frequently it appears in a corpus, the more likely it is to be significant in this corpus” [15]; although “alone, the frequency is not a robust metric to assess the terminological property of a candidate, but it does carry useful information, as does also the length of terms”[19].

First of all we created wordlists for every article as well as for its translations, this helps in the comparative analysis of the original and its translation as well as across the articles

² By minor language, we refer to a language that is not spoken by many people.

within every subcorpus. This helped us to get some reliable results about what is domain-specific within the corpora.

For the extraction of keyword lists, we used as reference corpora, the wordlist of the written component of the British National Corpus³[2] for our English analysis corpus and the wordlist of the written component of the CGT⁴[8] (Corpus of Greek Texts) for our Greek analysis corpus. The reason these two general corpora were selected as reference corpora, in order to get keyword list for our analysis corpora is that they are big enough –albeit not domain-specific. The keyword lists we retrieved were also for every article of each subcorpus separately and for the whole of the articles included in a subcorpus.

The way keywords are calculated is, according to Wordsmith 3.0 manual, the following [22]: the frequency of each word in the smaller of the two wordlists is compared with the frequency of the same word in the reference wordlist. All words which appear in the smallest wordlist are considered, unless they are in a stop list⁵. The keyness is a very important element of the Wordsmith tool because it computes one item's frequency in the small wordlist, the number of running words in the small wordlist, the item's frequency in the reference corpus, the number of the running words in the reference corpus and finally cross-tabulates all these. The element of keyness was used extensively in this study and a part of the results was actually based on it.

Looking at the frequency of a specific word in the subcorpus and at its keyness gives us some information about which words could be considered as candidate terms; the collocations of these words can then be checked.

At first sight, the keyword lists appear to be very similar. This is also an indication that the translations are close enough to the originals and our results are reliable. Therefore, for the planetology/cosmology subcorpus, we get keyword lists –when sorting the results by keyness–, where words like *supernova*, *star*, *explosions*, *simulations*, *bubbles*, *deflagration*, *dwarf*, *core*, *neutrinos*, *thermonuclear*, *collapse* etc. appear on the top of the English keyword list and words like *σουπερνόβα*, *άστρο*, *εκρήξεις*, *προσομοιώσεις*, *φουσαλίδες*, *νετρίνα*, *κατάρρευση*, etc. appear on the top of the Greek wordlist. When we re-sort the keyword lists, by the frequency the linguistic items appear in the analysis corpus, we get

³ BNC size: approximately 100 million words.

⁴ CGT size: approximately 22 million words.

⁵ The function words are words, and more precisely grammatical ones, like the, and, am etc, which are not semantically interesting and are usually not included in any study.

almost the same results, with, at the top of our list, words like *star*, *supernova*, *explosions*, *core*, *white*, *collapse*, etc. for the English list and words like *σουπερνόβα*, *άστρο*, *εκρήξεις*, *ενέργεια*, *προσομοιώσεις* etc. for the Greek list.

From the results, we conclude that these words are representative in the planetology/cosmology subcorpus and some of them may be terms as well. In order to verify this, we check the concordances of the words which we think may be candidate terms.

In checking the concordances of these words, we came across three cases:

- Checking words like *supernova*, *neutrinos*, or *deflagration* within the concordance lines, we come to the conclusion that these words constitute terms. They rarely appear outside a non-scientific context (high degree of technicality), they also obey the general rule that terms are, most of the times, nouns and they appear in technical dictionaries. [6]
- For words like *stars*, *explosions* and *bubbles*, on the other hand whose degree of technicality is debatable we check the concordances one-by-one to see if there is anything unusual or special about them which could make them behave like terms. At the end, we came to the conclusion that they do not constitute terms, although they appear in a scientific context.
- Finally for terms like *core*, *dwarf*, *simulations* and *collapse* looking at the collocations, we discovered a very interesting fact. These words, although they do not appear to be technical, when they appear on their own, actually are in many of these cases, as a constituent of a multi-word phrase. Thus, we observe that, for instance, *core* collocates very often with *collapse* or that *core* is also followed by the prepositional set: *of+supernova* or collocates with the adjective *stellar*. In one of the cases, also, we came across the compound word *core-collapse*, which in the Greek translation is given periphrastically (with a prepositional set *από+κατάρρευση σουπερνόβα*). From the above, thus, we concluded that while *core* appears to be a technical term, *collapse* on its own is not. So, we wondered how one can decide whether *core collapse* is a technical term or not. They may constitute a collocation, but this would involve other criteria which are outside the scope of this study. Another example is the noun *dwarf*, which collocates only with the adjective *white*: a fact which makes them inseparable and thus a multi-word term, which is also included in technical dictionaries, Finally, we check the word

simulations and we notice that *simulations* can either collocate with the noun *computer* preceding it or can just appear alone in a phrase, implying again *computer simulations*. The Greek translator, trying to be close to the original, translates *simulations* to *προσομοιώσεις* and *computer simulations* to *υπολογιστικές προσομοιώσεις*. Adding *simulations* or *computer simulations* to the list of candidate terms is also a decision which needs to be taken in common by the teacher and the students, according to how important the term is in the context of the translation task.

4 Evaluation of the methods and future applications

In this paper we have tried to demonstrate a method for extracting terminologically relevant collocations from an English-Greek parallel corpus. Although, we did not end up with a full list of terms:

“An important point to remember is that, although terminology in all its forms, from the special language glossaries and dictionaries to the big databases available, is of considerable interest to the translator, the world of terminology is not always primarily interested in the translator”. [14]

we hope that our method can show a relatively simple way of retrieving terminologically relevant collocations, by just using the parameters of user-friendly corpus tools.

The main purpose here was not only to retrieve terminology, but to supply students with a valuable and handy method: “a creation of a domain-specific parallel corpus, a kind of archive of past translations and their originals, which serves as a reference as well as terminology source for further translations within the domain” [21].

Working with pos-tagged texts would improve the results and reduce the efforts of post-processing, but the tagging itself tends to be a very time-consuming phase in the process of corpus building, especially for minor languages. Nonetheless, for the time being, we are restricted to working with the limited means we have at our disposal and in the future, we envisage a greater availability of accurate and efficient taggers for Greek and we expect to see a greater interest on the part of translation departments working from and into Greek.

References

- [1] Blank, Ingeborg, “Terminology Extraction from Parallel Technical Texts” In Jean Veronis, *Parallel Text Processing: Alignment and Use of Translation Corpora*, Kluwer Academic Publishers, 2000, pages 237-252.

- [2] British National Corpus: <http://www.natcorp.ox.ac.uk/>
- [3] Cabre Castellvi, Maria Teresa; Bagot, Rosa Estopa; Palatresi, Jordi Vivaldi, "Automatic Term Detection: A review of current systems", In Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme, *Recent Advances in Computational Terminology*, John Benjamins Publishing Company, 2001, pages 53-87.
- [4] "Chambers Dictionary of Science and Technology", Chambers Harrap Publishers Ltd, New York, 2002.
- [5] Chen Yijiang; Zhou Changle; Shi Xiaodong, "Automatic Extraction of Chinese Terms", In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference*, 30 Oct.-1 Nov. 2005, pages 281-286.
- [6] Chung, Teresa Mihwa; Nation Paul, "Identifying technical vocabulary", *System* Vol.32, Issue 2, June 2004, pages 251-263.
- [7] Cmejrek Martin; Curin Jan, "Automatic Extraction of Terminological Translation Lexicon from Czech-English Parallel Texts", *International Journal of Corpus Linguistics*, Vol.6 (special issue), 2001, pages 1-12.
- [8] Corpus of Greek Texts: <http://www.ucy.ac.cy/sek/>
- [9] "Dictionary of Technology and Sciences English-Greek, Greek-English", Stafylidis Publisher, Siemens Tech, Athens, 2005.
- [10] van der Eijk, Pim, "Automating the Acquisition of Bilingual Terminology", In *EACL*, 1993 pages 113-119.
- [11] Fictumova, Jarmila, "Technology-enhanced Translator Training", In *Second International Workshop on Language Resources for Translation Work, Research and Training*, Geneva 2004, Coling 2004, 2004, od s. 31-35, 6s. [online] <http://isg.urv.es/cttt/cttt/research.html>.
- [12] Gamper, Johann, "Encoding a Parallel Corpus for Automatic Terminology Extraction", In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen/Norway, June 1999, pages 275-276.
- [13] Heid Ulrich, "Extracting Terminologically Relevant collocations from German Technical Texts", In: Sandrini, Peter (ed.): *5th International Congress on Terminology and Knowledge Engineering*, (TKE '99), Innsbruck, Wien:TermNet, 1999, pages 212-221.
- [14] Maia, Belinda, "Using Corpora for Terminology Extraction: Pedagogical and computational approaches", In B. Lewandowska-Tomasczczyk (ed) 2003 *PALC 2001 – Practical Applications of Language Corpora*. Lodz Studies in Language, Frankfurt: Peter Lang, pages 147-164.
- [15] Lemay Chantal; L'Homme, Marie-Claude; Drouin Patrick, "Two methods for extracting "specific" single-word terms from specialized corpora: Experimentation and Evaluation", *International Journal of Corpus Linguistics*, 10 (2), pages 227-253.
- [16] Multiconc: <http://artsweb.bham.ac.uk/pking/multiconc/lingua.htm>

- [17] Sager, Juan C., "A Practical Course in Terminology Processing", John Benjamins Publishing Company, 1990, pages 39-54.
- [18] Sager, Juan C., "Language Engineering and Translation: Consequences of Automation", John Benjamins B.V., 1994, page 180.
- [19] Patry, Alexander; Langlais Philippe, "Corpus-Based Terminology Extraction", In *Proceeding of the 7th International Conference on Terminology and Language Engineering*, Copenhagen, Denmark, August 17-18, 2005, pages 313-321.
- [20] Penas, Anselmo; Verdejo, Felisa; Gonzalo, Julio, "Corpus-based terminology extraction applied to information access", In *Proceeding of the Corpus Linguistics Conference 2001*, 2001.
- [21] Vintar Spela, "Using Parallel Corpora for Translation-Oriented Term Extraction", *Babel* 47:2, 2001, pages 121-132.
- [22] Wordsmith 3.0: <http://www.lexically.net/wordsmith/version3/manual.pdf>
- [23] Zanettin, Federico, "Bilingual Comparable Corpora and the Training of Translators", In *Meta*, XLIII, 4, Special Issue, *The corpus-based approach: a new paradigm in translation studies*, 1998, pages 616-630.

Maria Matsira

Translator: Graduate from the Department of Foreign Languages Translation and Interpreting, Corfu, Greece.

Postgraduate research student in the MPhil (B) Corpus Linguistics, University of Birmingham, United Kingdom.

Address: 31 Kazazi street Kalamaria

55133 Thessaloniki Greece

+30 2310434735 +30 6947216435

E-mail: mmatsira@yahoo.gr, mariamatsira@hotmail.com