31 The importance of corpora in translation studies: a practical case

Montserrat Bermúdez Bausela[1]

Abstract

This paper deals with the use of corpora in Translation Studies, particularly with the so-called '*ad hoc* corpus' or 'translator's corpus' as a working tool both in the classroom and for the professional translator. We believe that corpora are an inestimable source not only for terminology and phraseology extraction (cf. Maia, 2003), but also for studying the textual conventions that characterise and define specific genres in the translation languages. In this sense, we would like to highlight the contribution of corpora to the study of a specialised language from the translator's point of view. The challenge of our particular study resides in combining in a coherent way different linguistic issues with one aim in mind: looking for the best way to help the student acquire and develop their own competence on translation, and that this is reflected in the professional field.

Keywords: translation studies, *ad hoc* corpus, specialised languages.

1. Introduction

This paper shows how the compilation of an *ad hoc* corpus and the use of corpus analysis tools applied to it will help us with the translation of a specialised text in English. This text could be sent by the client or used by the teacher in the classroom.

1. Universidad Alfonso X el Sabio, Villanueva de la Cañada, Madrid, Spain; mbermbau@uax.es

How to cite this chapter: Bermúdez Bausela, M. (2016). The importance of corpora in translation studies: a practical case. In A. Pareja-Lora, C. Calle-Martínez, & P. Rodríguez-Arancón (Eds), *New perspectives on teaching and working with languages in the digital era* (pp. 363-374). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2016.tislid2014.448

The corpus used for the present study is a comparable bilingual (English and Spanish) specialised corpus consisting of texts from the field of microbiology. Once our corpus is operative to be exploited using corpus processing tools, our aim is to study terminological, phraseological and textual patterns in both the English and the Spanish corpus to help us make the best informed decision as to the most appropriate natural equivalents in the Target Language (TL) in the translation process (cf. Bowker & Pearson, 2002; Philip, 2009). We intend to do so thanks to word lists, concordance, collocates and cluster searching. All these utilities are provided by the lexicographical tool *WordSmith Tools*.

2. Background

As Bowker and Pearson (2002) highlight, a corpus is a large collection of *authentic* texts, as opposed to 'ready-made' texts; they are in *electronic* form, which allows us to enrich them as we go along, and they respond to a *specific set of criteria* depending on the goals of the research in mind.

There are many fields of study in which linguistic corpora are useful, such as lexicography, language teaching and learning, sociolinguistics, and translation, to name a few. Using García-Izquierdo and Conde's (2012) words, "[i]n any event, regardless of their area of activity, most subjects feel the need for a specialised corpus combining formal, terminological-lexical, macrostructural and conceptual aspects, as well as contextual information" (p. 131). The use of linguistic corpora is closely linked to the need to learn Languages for Specific Purposes (LSPs). In this sense, translators are among the groups who need to learn and use an LSP, since they are non-experts of the specific field they are translating and they need to acquire both a linguistic and a conceptual knowledge in order to do so.

From the observation of specialised corpora, it is possible to identify specific patterns, phraseology, terminological variants, the frequency of conceptually relevant words, cohesive features and so forth. The access to this information will allow the translator to produce quality texts. Vila-Barbosa (2013) argues

that Corpus Linguistics can be applied to the study of translation, among other disciplines. The line of research focusing on Corpus Translation Studies (CTS) stems from the descriptive approximations of Translation Studies, which consider the text as the unit of study depending on the context in which it is produced.

3. Methodology, corpus design and compilation

Cabré (2007) mentions the type of specialised texts that we need to include in our corpus so that it is balanced. Among the most relevant criteria highlighted by this author, we identify the topic, level of specialisation, textual genre, type of text, languages, sources, and, in the case of multilingual corpora, the relation established between the texts in the different languages. We could also add the communicative function, which is really implicit in the rest of the criteria mentioned by the author.

The whole process begins by choosing a specialised text in the Source Language (SL). It may be the text that the teacher and the students are working with in the classroom, or the actual text sent by the client to be translated. It could belong to any field: scientific, technical, legal, business, etc. In our particular case, we have taken as our Source Text (ST) the article entitled "Antibacterial activity of Lactobacillus sake isolated from meat" by Schillinger and Lücke (1989). We have chosen this one in particular because we think that it is a good example of a highly specialised text, scientific in this case, which is confirmed not only by its specialised terminology, but also by its macrostructure. It is an academic and professional type of discourse in which both the sender and the recipient are experts (high degree of shared knowledge) and it is an expositive and explicative type of text.

3.1. Corpus compilation in English

What we first need to know is the field of study and the level of specialisation of the ST. With this aim in mind, we have generated a wordlist (using the software *WordList*, provided by *WordSmith Tools*) of the most frequent words

in the text, which will provide us with the specific terminology (*bacteriocin, strain, culture, agar, bacteria, plasmid, supernatant,* etc.). In order to start building our corpus, we search on the Internet for texts that include a number of the above mentioned terms. Each text has been saved individually in TXT format (the format supported by *WordSmith Tools*). All files have been stored in a folder named MEAT_INDUSTRY CORPUS with two subfolders, for the English and the Spanish texts. On most occasions, the texts were in PDF format and had to be converted into TXT, which implied a thorough and laborious cleaning process.

All the results obtained in our search are specific papers published in Journals. This is important since the results are going to be equally comparable with the ST regarding topic, level of specialisation, textual genre and type. The degree of reusability of our corpus is very high, since it has been created with the aim to be further enlarged and enriched with each new translation project.

The following are some interesting facts of the English compilation corpus:

- *Accuracy and reliability*: All the chosen texts (and this applies to both the English and the Spanish corpus) have passed a strict quality control, since they are published in well-known journals that have a peer-review process. Awareness has always been raised regarding the quality of the information found on the Internet. Harris (2007) points out the CARS Checklist (Credibility, Accuracy, Reasonableness and Support) as the criteria designed to guarantee high quality information on the Internet. We believe that even though we can never lower our guard, if the previous terminological job is done accurately and precisely, the results will very likely be knowledgeable, authentic and trustworthy, also due in great part to the development of the current search engines.

- *Limited accessibility*: It has not been an easy task to have free access to the academic texts. Therefore, apart from the free-downloadable ones, we have also included texts made up by *Abstracts*, which were, on all occasions, free.

- *Text originality*: Olohan (2004) defines bilingual or multilingual comparable corpora as "comparable original texts in two or more languages" (p. 35). But, can we be sure that all the texts that make up our corpus were originally written in English? However, even if these texts are *covert translations* (House, 2006), they are presented to the scientific community as originals, and they are totally acceptable and functional translations working in the target system as if they were originals. In fact, Baker (1995) does not refer to comparable corpora of texts as 'original' texts in two or more languages, since it is very hard to determine if they have really been written in the SL or they are translations in themselves. Apart from this, English is the *lingua franca* in scientific communication and it is the most frequent language of scientific scholarly articles published on the Internet.

3.2. Corpus compilation in Spanish

We now start building the Spanish corpus by searching for texts in Spanish that include the equivalents in Spanish of some of the most frequent and representative terms in the ST in English (we have searched for texts that included *bacteriocina, cepa, cultivo, agar, bacteria, plásmido, sobrenadante,* etc.). Some of the issues raised in the compilation of the Spanish corpus have been:

- *Wider variety of textual genres in the output*: We have not only gathered scientific articles, but also PhD theses and final year dissertations, which considerably enlarges the size of the Spanish corpus compared to the English one.

- *Cleaning*: The Spanish texts have required more 'cleaning' than the English texts. This is due to the fact that they included parts in English, such as the abstracts, the acknowledgments, or part of the bibliography.

We include in Table 1 statistical information regarding our corpus, where we can observe, among other data, the running words in the corpus (tokens) versus the different words (types), thus obtaining the resulting type/token ratio.

3.3. Asking the corpus the 'right' questions

The translator becomes a bit of an expert with each new translation brief. It is important to understand the meaning behind the term and learn something about the subject. In this context, corpora are of great importance, since we can search the corpus to find this kind of information (Table 1).

Table 1. Corpus statistical information

	English corpus. Statistical details	**Spanish corpus. Statistical details**
Number of files	29	27
Tokens	67.844	363.424
Types	6.466	18.994
Ratio Type/Token	10.73	5.87
Number of sentences	4.991	16.149

Sometimes it is also difficult for translators to locate equivalents, or to choose among several possible ones. Even if we are not using a parallel corpus, we can still identify a terminological equivalent, sometimes even guided by our intuition: we might suspect what the correct equivalent is, but we need to check it in our corpus. What we can do is generate a concordance and verify if our intuition was right. Towards this end, we recommend using an asterisk. This particular wildcard substitutes an unlimited number of characters. Like this, we will be able to rule out an incorrect equivalent and check the different varieties of the term.

The most frequent word in the ST has been *bacteriocin*, with a frequency of 0.98%. A corpus can help us identify terms shown in context, and the most frequent patterns of use. From the different concordance lines, collocates and clusters (retrieved thanks to the software *Concord*, a functionality provided by *WordSmith Tools*), we obtain relevant grammatical and lexicographical information. We show a very brief example of the terminological equivalents and the patterns found for *bacterio**.

The terminological English variants are:

- bacteriocin (401 entries), bacteriocins (238 entries);
- bacteriocinogenic (42 entries);
- bacteriocidal (1 entry).

The terminological Spanish variants are:

- bacteriocinas (1070 entries), bacteriocina (554 entries);
- bacteriostático/bacteriostática (31 entries);
- bacteriocinogénicas/bacteriocinogénicos (23 entries);
- bacteriolítica/bacteriolítico (13 entries);
- bacteriocidal (2 entries).

Please refer to Table 2 to see the most common patterns of *bacterio**.

Table 2. Contrastive study of the use of *bacterio** in English and Spanish

English	**Spanish**
bacteriocinogenic + noun (bacteriocinogenic activity, bacteriocinogenic strain)	noun + bacteriocinogénica/o (actividad bacteriocinogénica, cepa bacteriocinogénica)
bacteriocin + noun (bacteriocin activity, bacteriocin inhibition)	noun + bacteriocinas (actividad de las bacteriocinas, inhibición a las bacteriocinas)
Bacteriocin(s) + participial form (bacteriocins produced by, bacteriocin isolated from)	Bacteriocina(s) + participial form (bacteriocinas producidas por, bacteriocinas sintetizadas por)
bacteriocins + verb in passive voice (bacteriocins were first discovered, bacteriocins were defined by)	bacteriocinas + verb in active voice (las bacteriocinas presentan, las bacteriocinas inhiben)
bacteriocin + ing form (bacteriocin-producing strains, bacteriocin-producing lactococcus)	bacteriocinas + 'de' + type (bacteriocinas de Lactococcus, bacteriocinas de bacterias ácido lácticas)

We also learn about the most common verbs that are collocates of 'bacteriocina(s)' in the Spanish corpus: 'producir', 'codificar', 'aislar', 'presentar', etc.

All this information is of utmost importance for the translation of the text. A corpus can help us reflect the most natural style in our Target Text (TT). As Philip (2009) claims, TL norms should be borne in mind "when reproducing any idiosyncratic usage or innovative expressions that the SL text might include" (p. 59).

4. Using corpora in translation: an example

We would like to show an example of the direct contribution of corpora to translation practice. Let us look at this sentence taken from the abstract of the article we are using as our ST and suppose we need to translate it into Spanish:

> "In mixed culture, the bacteriocin-sensitive organisms were killed after the bacteriocin-producing strain reached maximal cell density, whereas there was no decrease in cell number in the presence of the bacteriocin-negative variant".

There are certain issues that catch our attention, such as how we could translate the following compound nouns:

- bacteriocin-sensitive organisms (see pattern 1);
- bacteriocin-negative variant (see pattern 2);
- bacteriocin-producing strain (see pattern 3).

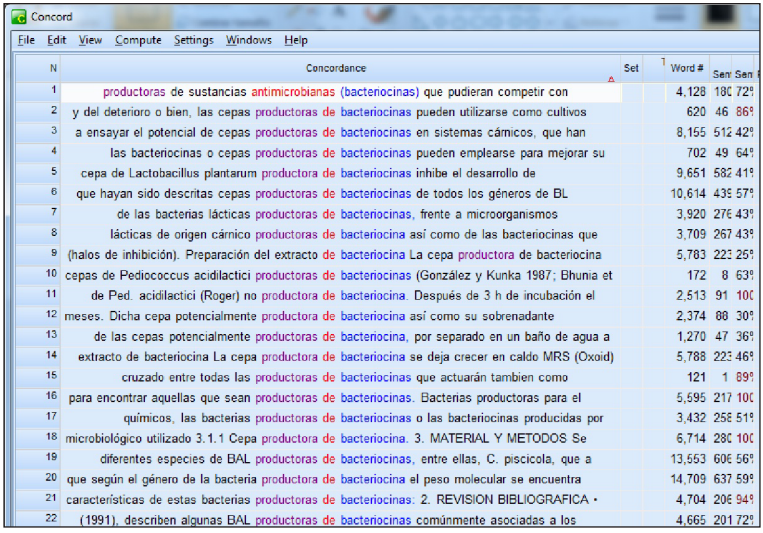
Pattern 1: the first thing we do is conduct a concordance search in the Spanish corpus using 'sensible*' as our search word and including a context word, 'bacteriocina*'. A context word is used to check if it typically occurs in the

vicinity of our search word in a specified horizon to the right and left of the search word. Also, we use a wildcard, the asterisk, in order to look for all the possible variants. We obtain a result of 10 concordance lines, from which we can deduce that the most frequent expression in Spanish is 'organismos sensibles a las bacteriocinas'.

Pattern 2: we conduct a concordance search using 'bacteriocina' as our search word and include the context word 'negativa'. In the outcome, we observe the concordance line: 'variante negativa para bacteriocina'.

Pattern 3: we look for the search word: 'bacteriocina*' and include the context word: 'productora*'. The results are astounding: 56 lines of concordances and in all of them we can observe that in Spanish the noun phrase 'cepa productora de bacteriocina' is very frequent (Figure 1).

Figure 1. Concordance lines of *bacteriocina**, context word *productora**


Concord

File Edit View Compute Settings Windows Help

N	Concordance	Set	Word #	Sen	Sen
1	productoras de sustancias antimicrobianas (bacteriocinas) que pudieran competir con		4,128	180	72%
2	y del deterioro o bien, las cepas productoras de bacteriocinas pueden utilizarse como cultivos		620	46	86%
3	a ensayar el potencial de cepas productoras de bacteriocinas en sistemas cárnicos, que han		8,155	512	42%
4	las bacteriocinas o cepas productoras de bacteriocinas pueden emplearse para mejorar su		702	49	64%
5	cepa de Lactobacillus plantarum productora de bacteriocinas inhibe el desarrollo de		9,651	582	41%
6	que hayan sido descritas cepas productoras de bacteriocinas de todos los géneros de BL		10,614	435	57%
7	de las bacterias lácticas productoras de bacteriocinas, frente a microorganismos		3,920	276	43%
8	lácticas de origen cárnico productoras de bacteriocina así como de las bacteriocinas que		3,709	267	43%
9	(halos de inhibición). Preparación del extracto de bacteriocina La cepa productora de bacteriocina		5,783	223	25%
10	cepas de Pediococcus acidilactici productoras de bacteriocinas (González y Kunka 1987; Bhunia et		172	8	63%
11	de Ped. acidilactici (Roger) no productora de bacteriocina. Después de 3 h de incubación el		2,513	91	100
12	meses. Dicha cepa potencialmente productora de bacteriocina así como su sobrenadante		2,374	88	30%
13	de las cepas potencialmente productoras de bacteriocina, por separado en un baño de agua a		1,270	47	36%
14	extracto de bacteriocina La cepa productora de bacteriocina se deja crecer en caldo MRS (Oxoid)		5,788	223	46%
15	cruzado entre todas las productoras de bacteriocinas que actuarán tambien como		121	1	89%
16	para encontrar aquellas que sean productoras de bacteriocinas. Bacterias productoras para el		5,595	217	100
17	químicos, las bacterias productoras de bacteriocinas o las bacteriocinas producidas por		3,432	258	51%
18	microbiológico utilizado 3.1.1 Cepa productora de bacteriocina. 3. MATERIAL Y METODOS Se		6,714	280	100
19	diferentes especies de BAL productoras de bacteriocinas, entre ellas, C. piscicola, que a		13,553	606	56%
20	que según el género de la bacteria productora de bacteriocina el peso molecular se encuentra		14,709	637	59%
21	características de estas bacterias productoras de bacteriocinas. 2. REVISION BIBLIOGRAFICA •		4,704	206	94%
22	(1991), describen algunas BAL productoras de bacteriocinas comúnmente asociadas a los		4,665	201	72%

As mentioned previously, specialised translation is not only about terminology, but also about style. Our translation should resemble other texts produced within

that particular LSP. It must be stylistically appropriate as well as terminologically accurate. In this sense, we came across a difficulty in the translation of 'the bacteriocin-sensitive organisms were killed'. We did not find in our corpus any example of concordance of 'organismos eliminados' or 'fueron eliminados'. As it seems, we had come across the appropriate collocate but not the appropriate style. The verb 'eliminar' in the Spanish corpus follows the grammar pattern: verb + object (eliminar microorganismos) and in a large number of the cases, the noun 'eliminación' is used. Suggested translation:

> "En un cultivo mezclado, la eliminación de los organismos sensibles a la bacteriocina se produjo después de que la cepa productora de bacteriocina alcanzara la máxima densidad celular, mientras que no hubo disminución en el número de células en presencia de la variante negativa para bacteriocina".

5. Conclusions

There is a number of ways in which specialised corpora can help the translator. We can generate word lists to identify the field and level of specialisation of the ST. We can use them to learn about the subject we are translating, and about the most common lexical and grammatical patterns through the retrieval of concordances, collocates and clusters. Furthermore, it is an invaluable source regarding style: choosing the appropriate textual conventions and norms that the recipient of the TT expects to find reflected on the text is a guarantee that the text will have a high degree of acceptability. As Corpas-Pastor (2004, p. 161-62) points out, it involves a great development in the documentary sources for the translator, since the proper selection, assessment and use of those sources let the translator focus on developing strategies to consult the corpus and extract valuable information, optimizing time and effort. We believe that corpora help the student acquire and develop their own competence on translation, and that their use perfectly responds to the specialised translator's needs.

References

Baker, M. (1995). Corpus linguistics and translation studies: implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: in honour of John Sinclair* (pp. 17-45). Amsterdam/Philadelphia: John Benjamins.

Bowker, L., & Pearson, J. (2002). *Working with specialized language. A practical guide to using corpora*. London: Routledge. Retrieved from http://dx.doi.org/10.4324/9780203469255

Cabré, M. T. (2007). Constituir un corpus de textos de especialidad: condiciones y posibilidades. In M. Ballard & C. Pineira-Tresmontant (Eds.), *Les corpus en linguistique et en traductologie* (pp. 89-106). Arras: Artois Presses Université.

Corpas-Pastor, G. (2004). *La traducción de textos médicos especializados a través de recursos electrónicos y corpus virtuales*. Actas del II Congreso. Las palabras del traductor. Toledo, 2004. El español, lengua de traducción. Congreso internacional de ESLETRA. Retrieved from http://cvc.cervantes.es/lengua/esletra/pdf/02/017_corpas.pdf

García-Izquierdo, I., & Conde, T. (2012). Investigating specialized translators: corpus and documentary sources. *Ibérica, 23*, 131-156.

Harris, R. (2007). *Evaluating internet research sources*. Radnor Township School District. Retrieved from http://radnortsd.schoolwires.com/cms/lib/PA01000218/Centricity/ModuleInstance/2137/Evaluating_Internet_Research_Sources.pdf

House, J. (2006). Covert translation, language contact, variation and change. *SYNAPS, 19*, 25-47.

Maia, B. (2003). What are comparable corpora? In *Proceedings of pre-conference workshop multilingual corpora: linguistic requirements and technical perspectives* (pp. 27-34). Lancaster: Lancaster University.

Olohan, M. (2004). *Introducing corpora in translation studies*. London: Routledge.

Philip, G. (2009). Arriving at equivalence. Making a case for comparable general reference corpora in translation studies. In A. Beeby, I. Patricia-Rodríguez, & P. Sánchez-Gijón (Eds.), *Corpus use for learning to translate and learning corpus use to translate* (pp. 59-73). Amsterdam/Philadelphia: John Benjamins. Retrieved from http://dx.doi.org/10.1075/btl.82.06phi

Schillinger, U., & Lücke, F. K. (1989). Antibacterial activity of Lactobacillus sake isolated from meat. *Applied and Environmental Microbiology, 55*(8), 1901-1906.

Vila-Barbosa, M. M. (2013). Corpus especializados como recurso para la traducción: análisis de los marcadores de la cadena temática en artículos científicos sobre enfermedades neuromusculares en pediatría. *Onomázein, 1*(27), 78-100.

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; Voillans, France, info@research-publishing.net

New perspectives on teaching and working with languages in the digital era
Edited by Antonio Pareja-Lora, Cristina Calle-Martínez, Pilar Rodríguez-Arancón

Typeset by Research-publishing.net
Cover design and frog picture by © Raphaël Savina (raphael@savina.net)

ISBN13: 978-1-908416-34-6 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method, with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-35-3 (Ebook, PDF, colour)
ISBN13: 978-1-908416-36-0 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: mai 2016.