>>> revista tradumàtica
Traducció i Tecnologies de la Informació i la Comunicació
Número 07: L'aplicació dels corpus linguístics a la traducció : desembre 2009 :
http://www.fti.uab.cat/tradumatica/revista : ISSN: 1578-7559

Principles of corpus linguistics and their application to translation studies research

Gabriela Saldanha
Centre for English Language Studies, University of Birmingham

1. Introduction

Corpora have been put to many different uses in fields as varied as natural language processing, critical discourse analysis and applied linguistics, to mention just a few. As is to be expected, within each of those areas corpora fulfil different roles, from providing data to build statistical machine translation systems to revealing ideological stance in politically-sensitive texts. 'Corpus linguistics' is understood here in a more restricted sense, linked to British traditions of text analysis that see linguistics as a social science and language as a means of social interaction where meaning is inextricably linked to the cultural and historical context in which it is produced. This article focuses specifically on the principles of corpus linguistics as a research methodology, and looks at the implications of this specific approach to the study of language in translation studies.

2. A corpus defined in corpus linguistics terms

Because there is no unanimous agreement on the necessary and sufficient conditions for a collection of texts to be a corpus, the term 'corpus' can be seen in the literature referring sometimes to a couple of short stories stored in electronic form and sometimes to the whole world wide web. In order to discuss the fundamental principles of corpus linguistics, it is important to first establish certain limits around what can and cannot be considered a 'corpus-based' study of translation.

Different definitions of corpus emphasise different aspects of this resource. The definition offered by McEnery and Wilson (1996: 87), for example, emphasises representativeness: "a body of text which is carefully sampled to be maximally representative of a language or language variety". The problem with making representativeness the defining characteristic of a corpus is that it is very difficult to evaluate and it will always depend on what the corpus is used for. A way around this problem is found in the definition offered by Bowker and Pearson (2002: 9): "a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria". Bowker and Pearson's definition is more flexible than McEnery and Wilson's, even if the assumption is still that the corpus is intended to be "used as a representative sample of a particular language or subset of that language" (Bowker and Pearson, 2002: 9). However, in making selection criteria and not representativeness the defining characteristic, Bowker and Pearson allow for a certain flexibility that reflects more accurately the fact that corpus representativeness is always dependent on the purpose for which the corpus is used and on the specific linguistic features under study. For example, a corpus that represents accurately the distribution of a common feature – say, pronouns – in a certain language subset may not represent accurately a rarer feature, such as the use of reported speech, in the same subset. Generally, corpora are intended to be long-term resources and to be used for a variety of studies, so representativeness cannot be ensured at the design stage.

According to Bowker and Pearson's definition, selection criteria is one of four aspects that differentiate a corpus from other collections of texts; the others are size, authenticity of the data and means of storage. Authentic data is generally understood as naturally occurring data, that is, not originally created or elicited for the purpose of linguistic analysis. The reference to means of storage in Bowker and Pearson's definition is instrumental in

differentiating current corpus linguistics from a longer-established tradition of manually analysing collections of texts – in some cases also relatively extensive – for purposes of extracting data. Regarding size, Bowker and Pearson only indicate that a corpus should be 'large'. Giving more precise indications of size is problematic because whether a corpus is 'large' will depend on what it tries to represent. As a common-sense criterion, Bowker and Pearson suggest: "a greater number of texts than you would be able to easily collect and read in printed form" (2002: 10).

An even more flexible 'definition' of corpus is offered by Leech (1992: 106): "a helluva lot of text, stored on a computer". Here, the emphasis is obviously on size and medium, but no criterion is offered as to what differentiates a corpus from other collections of texts; Leech seems to imply that there is no need for such a distinction. A similarly flexible approach is taken by Kilgarriff and Grefenstette (2003: 334): "A corpus is a collection of texts when considered as an object of language or literary study." It can be argued that the focus on linguistic study can be taken for granted in corpus linguistics, so this does not really add a constraint to what can be considered a corpus. Still, Kilgarriff and Grefenstette make a very good point, which is that we should not confuse the question "What is a corpus?" with "What is a good corpus (for certain kinds of linguistic study)?" (ibid). The conclusion we can draw is that if we are concerned about what makes a 'good corpus' then sometimes size (if large enough, and Kilgarriff and Grefenstette are talking about the whole world wide web as a corpus) can outweigh the benefits of carefully selected criteria.

3. The object of study in corpus linguistics and translation studies

Corpus linguistics is not a linguistic theory but a methodology that can be applied to a wide range of linguistic enquiries; however, there is more to corpus linguistics than the use of corpora. Some scholars consider it to be a research paradigm in its own right (Tognini-Bonnelli, 2001; Laviosa, 2002), on the basis that doing research using corpora generally entails some basic assumptions as to what is the object of enquiry and how it should be studied. Much of the work done within corpus linguistics, particularly in Britain, is informed by Firthian and neo-Firthian approaches to language, which see language as essentially a communication tool (rather than, for example, a cognitive process) and are concerned with practical applications of linguistic research (see Stubbs, 1996). The use of corpora in translation studies research was first proposed as particularly adapted to the purposes of empirical descriptive translation studies (Baker, 1993). Some of the principles underlying corpus linguistics are shared by descriptive translation studies, and this has been, as Laviosa (2004) points out, key to the success story of corpora in Translation Studies.

Corpus linguistics and descriptive translation studies focus on **'attested' language production.** Corpus linguistics uses authentic, or naturally occurring, texts (as opposed to intuitive, invented, isolated sentences). This goes hand in hand with what Toury recommends as a starting point in descriptive studies: "a study in translation activities which have already yielded their products would start with the *observables;* first and foremost, the translated utterances themselves, along with their constituents" (1995: 36). In other words, the focus is on performance rather than competence: both corpus linguistics and descriptive translation studies are interested in the full range of varieties of language production, including spontaneous, non-edited language use as well as edited, usually written, language, and neither grammaticality or translation quality are necessarily prerequisites. This does not mean that one of the criteria for compiling a specialized corpus cannot be translation quality, for example if it is to be used as a resource for assessing translations or for translation training, but they should be 'attested', real translations, rather than translations created for the purpose of translation assessment or translator training.

Descriptive translation studies encouraged moving away from the traditional comparison of translations against source texts, which entailed evaluating degrees of equivalence and faithfulness, usually from a prescriptive perspective. The object of a descriptive approach is instead to explain translated texts in their own terms and not as mere reproductions of other works. In other words, it aims to establish **distinctive features of translated texts**, so that the principles governing their production can then be explained and predicted (Toury, 1995). This requires finding linguistic patterns that are repeated across large numbers of translations, for which purpose electronic corpora are particularly suitable. However, the most rigorous counting of linguistic features is meaningless unless we can provide a relative norm of comparison. Local features have to be seen in relation to other features, and texts have to be considered against the background of other texts; this is another principle of corpus linguistics: **texts and text types are studied comparatively across text corpora.**

In translation studies, cross-linguistic comparison has been the default method of analysis. However, the increasing availability of different types of corpora puts at our disposal more sophisticated ways of assessing whether the frequency of a linguistic feature in a particular text is part of a more general trend in similar texts or is actually a distinctive feature of that particular text. Translational norms, like any other social norms, are essentially probabilistic; they are dependent on genre, text function, register and so on; and in order to account for these effects, comparative study across texts is essential. There are a variety of translational and non-translational corpora that can be used for this purpose, and these need not be described here.

The description of patterns and regularities of behaviour is in itself of little interest unless we are able to associate it with extra-linguistic factors of production. Both the neo-Firthian tradition in linguistics, and the systemic approaches to the study of translation (Hermans, 1999), which encompass descriptive translation studies, insist on the relationship between observable language phenomena and the non-observable norms and situations that affect translators/speakers' choices; in other words, they see a **connection between everyday routine and cultural transmission** (Stubbs, 1996). Closely linked to this assumption is another principle in neo-Firthian linguistics, highlighted particularly in the work of John Sinclar and Michael Halliday, and that is the **interdependence of form and meaning**. Corpus linguistics has demonstrated that lexical choices more often than not entail the choice of a specific grammatical form or structure, and vice versa. Halliday coined the term 'lexico-grammar' to refer to this phenomenon. Other linguistic traditions have tended to see grammar as autonomous and independent of meaning (Chomsky, 1957: 17), but Halliday stresses that

> "*all* types of option, from whatever function they are derived, are meaningful. … and if we attempt to separate meaning from choice we are turning a valuable distinction (between linguistic functions) into an arbitrary dichotomy (between 'meaningful' and 'meaningless' choices) (1971: 338).

This principle is particularly relevant to the study of translations because the argument that form and meaning are interdependent is at the heart of the translatability debate. The unavoidable impact on the meaning of the text by a change in form is also a keystone in the argument for translator's right to co-authorship of translated texts.

The theoretical principles described in this section have important methodological implications for corpus linguistics, which are described below. A corpus, judged by any of the criteria set out in Section 2, is still only a resource, and will only show us what we are capable of finding.

4. Methodological considerations

4.1 Corpus-based and corpus-driven approaches

Tognini-Bonelli (2001) has distinguished between corpus-based and corpus-driven studies, the main difference being that the former approach starts with a pre-existing theory which is validated using corpus data, while the latter

> builds up the theory step by step *in the presence of the evidence*, the observation of certain patterns leads to a hypothesis, which in turns leads to the generalisation in terms of rules of usage and finally finds unification in a theoretical statement (ibid: 17).

One of the disadvantages of using corpora as a testing ground for pre-existing hypotheses, in order to find quantitative data to support a certain theory, is that corpus linguistics has offered insights into language that have challenged the underlying assumptions behind many well established theoretical positions in the field, such as the division between lexis and grammar discussed above. Studies that are too strictly embedded in specific linguistic theories forego the potential to challenge theories and descriptions that were formulated before large corpora became available to inform language study. According to Tognini-Bonelli, corpus-based linguistics gives priority to the pre-existing theoretical statement and, rather than account for the variability of naturally occurring language, it attempts to "insulate it, standardise it and reduce it" (ibid: 67).

Although it may be useful for clarification purposes, the distinction proposed by Tognini-Bonelli is far too simplistic. As Tognini-Bonelli herself acknowledges, there is no such a thing as pure induction (ibid: 85), and intuition inevitably plays a part in any kind of research, from the selection of the phenomenon to be investigated to the interpretation of the results. Besides, there are no grounds to assume that corpus-based research will not be committed to the integrity of the data as a whole or aim to be comprehensive with respect to corpus-evidence, as Tognini-Bonelli seems to suggest (ibid: 84). There are several corpus-based translation studies that are examples to the contrary (see, for example: Kenny, 2001; Olohan, 2003; Saldanha, 2004). Olohan (2003) focuses on contractions in translated and non-translated English and her study at first reveals that contractions are much more common in non-translated English, which would suggest that translated language is possibly more formal. A more in depth exploration, however, shows that the overall frequencies average out important differences among individual texts or groups of texts, for example, by one translator. Looking at the work of certain translators (Peter Bush and Dorothy S. Blair) in more detail, she finds that overall frequency patterns suggest clear differences in the choices made by each translator, but again, a more detailed exploration starts to show the influence of the authors' style and genre conventions. Saldanha (2004) shows that the use of pre-existing hypotheses is not a problem in itself, as long as the exceptions to the norm are also accounted for and as long as we are prepared to revise our theories in the light of the data when this is required.

One of the main problems encountered by linguists who are committed to accept and reflect the evidence offered by authentic instances of language in context, is that it is not actually possible to find and account for every possible pattern that is prominent in a given text or texts. Thus, the corpus-driven linguist has to resign him or herself to plodding through the detail (Sinclair, 1991: 27). The alternative approach is to start with a potential explanation and then try to find evidence for or against it. The problem with this approach is that, given the great diversity of linguistic features and functions in a text, we run the risk of looking too narrowly into those areas where confirmatory evidence is likely to be found and, consequently, of focusing on those results that confirm the hypothesis and ignoring those that contradict it. An analysis of specific linguistic features necessarily shows a partial view of the data, so it is important that the selection of the features themselves is as impartial as possible.

4.2 Quantitative and qualitative approaches to data analysis

The use of corpora in linguistic research generally involves classifying and counting linguistic features and has therefore been considered to belong to the realm of quantitative analysis. The possibility of accounting for every occurrence of a specific item in a text in a systematic manner has allowed research in translation studies to move beyond the mere enumeration of examples, an approach that does not prove the validity of hypotheses or theories, but is nevertheless more tempting "since it starts functioning even with a very limited corpus, and even with an arbitrary one" (van Doorlaser, 1995: 247). In the context of a large corpus, on the other hand, the more interesting examples tend to become 'diluted' and, although the data may be more substantial, the conclusions are likely to be more modest. The literature on translation universals, for instance, has moved from impressive claims to cautious suggestions about 'features of translations' generally formulated with a considerable amount of qualification.

Quantitative methods in corpus linguistics vary widely, and can go from simple frequency counts, to simple but powerful calculations (type-token ratio, lexical density), to complex statistical techniques including significance tests. There are different views on the usefulness and reliability of significance tests in corpus linguistics. Many linguists highlight the need to demonstrate that any differences or similarities revealed are not due to chance, especially since sampling procedures cannot always guarantee representativeness, and some argue that corpus linguists should "collectively increase the level of statistical sophistication of our analyses" (Gries, online). However, the statistical tests used in corpus linguistics are generally those designed for use in the social sciences (Meyer, 2002: 120), and transferring the methodology to a field where the nature of the data is essentially different presents some problems. For example, the most powerful tests used in the social sciences (parametric tests) assume that the data are normally distributed, which is often not true of linguistic data (Oakes, 1998: 11; McEnery and Wilson, 1996: 70). Non-parametric tests, such as chi-square, on the other hand, are unreliable with small frequencies. Besides, as Danielsson (2003) points out, statistical tests in many cases do not show anything that cannot be revealed by simply comparing raw frequencies. Danielsson argues that, if something is recurrent in a text, it is there for a reason, but it cannot be expected that the reason may be discovered in a simple calculation, because "the distribution of words in texts is far more complex than a mathematical formula can perceive" (ibid: 114).

However, the use of corpora does not exclude qualitative analysis, and a combination of both approaches is necessary in order to provide a richer picture of the translational phenomena under observation, and in particular, to be able to offer explanations. Quantitative analysis "enables one to separate the wheat from the chaff" (McEnery and Wilson, 1996: 62-63); while qualitative analysis, which does not require the data to fit into a finite number of categories, enables very fine distinctions to be drawn (ibid: 62).

Quantitative and qualitative methods can be combined in a number of ways. In-depth qualitative analysis can form the basis for hypotheses that are afterwards tested through quantitative methods. Alternatively, the information obtained from the analysis of the translations themselves can be verified against the information obtained from external sources (and vice versa). This procedure, commonly known in the social sciences as triangulation, not only strengthens the evidence but is a crucial complement of corpus analysis if we are to explore potential motivations for translational behaviour in terms of the translators' cultural and ideological positions, or in terms of the context of situation or culture. In order to establish the connection between everyday routine and cultural transmission mentioned above, it is necessary to go beyond the textual data and look at extratextual material.

5. Conclusion

Linguistics has gradually moved from using words and clauses as the unit of analysis to considering texts as a whole and finally to seeing texts as instances of discourse that are constantly engaged in the dynamic representation and construction of knowledge and ideology. Nowadays, the trend in translation studies is towards foregrounding the social, cultural and political context of translation, and corpora are being used in areas that, by their very nature, require a more nuanced approach than we have seen so far, such as issues of style and ideology in translation (Baker, 2000; Munday, 2008; Saldanha, 2005; Winters, 2007, 2009). Despite corpus linguistics' concern with the relation between micro-linguistic events and macro-social structures, corpus analysis tools draw attention to patterns at the micro-linguistic level, and few of them facilitate access to extra-linguistic information about the texts. A development in this direction in corpus analysis and visualization techniques would therefore be welcome. In the meantime, if we are to benefit from the increased rigour achieved by the use of corpora and at the same time look beyond the text to contextualize our data, corpus analysis still needs to be combined with other methods.

References

Baker, M. (1993). "Corpus Linguistics and Translation Studies: Implications and applications" in Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds) *Text and Technology*, Amsterdam and Philadelphia: John Benjamins, 233-250.

Baker, M. (2000). "Towards a Methodology for Investigating the Style of a Literary Translator", *Target* 12(2): 241-266.

Bowker, L. and J. Pearson (2002). *Working with Specialized Language*: *A practical guide to using corpora*, London and New York: Routledge.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Danielsson, P. (2003). "Automatic extraction of meaningful units from corpora: A corpus-driven approach using the word *stroke*", *International Journal of Corpus Linguistics* 8(1): 109-127.

Gries, S. (online). "Useful statistics for corpus linguistics", http://www.linguistics.ucsb.edu/faculty/stgries/research/UsefulStatsForCorpLing.pdf, accessed September 2009.

Halliday, M.A.K. (1971). "Linguistic Function and Literary Style: An Inquiry into the Language of William Golding's *The Inheritors*", in Seymour Chatman (ed) *Literary Style: A Symposium*, London and New York: Oxford University Press, 330-365.

Hermans, T. (1999). *Translation in Systems: Descriptive and Systemic Approaches Explained.* Manchester: St. Jerome Publishing

Kenny, D. (2001). *Lexis and Creativity in Translation: A Corpus-based Study,* Manchester, St. Jerome Publishing.

Kilgarriff, A. and G. Grefenstette (2003). "Web as Corpus", *Computational Linguistics* 29(3): 333-47.

Laviosa, S. (2002). *Corpus-based Translation Studies: Theory, Findings, Applications,* Amsterdam and New York: Rodopi.

Laviosa, S. (2004). "Corpus-based Translation Studies: Where does it come from? Where is it going?", *Language Matters, Studies in the Languages of Africa* 35(1): 6-27.

Leech, G. (1992). "Corpora and theories of linguistic performance", in Jan Svartvik (ed) *Directions in corpus linguistics,* Berlin: Mouton De Gruyter, 105-122.

McEnery, T. and A. Wilson (1996). *Corpus Linguistics*, Edinburgh: Edinburgh University Press.

Meyer, C.F. (2002). *English Corpus Linguistics: An Introduction*, Cambridge: Cambridge University Press.

Munday, J. (2008). *Style and Ideology in Translation: Latin American Writing in English,* London and New York: Routledge.

Oakes, M. (1998). *Statistics for Corpus Linguistics,* Edinburgh: Edinburgh University Press.

Olohan, M. (2003). "How frequent are the contractions? A study of contracted forms in the Translational English Corpus", *Target* 15(1): 59-89.

Saldanha, G. (2004). "Accounting for the Exception to the Norm: a Study of Split Infinitives in Translated English", *Language Matters, Studies in the Languages of Africa* 35(1): 39-53.

Saldanha, G. (2005). *Style* of *Translation: An exploration of stylistic patterns in the translations of Margaret Jull Costa and Peter Bush.* Unpublished PhD Thesis. Dublin: School of Applied Language and Intercultural Studies, Dublin City University.

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

Stubbs, M. (1996). *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture,* Oxford and Malden: Blackwell Publishers.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work,* Amsterdam and Philadelphia: John Benjamins.

Toury, G. (1995). *Descriptive Translation Studies and Beyond,* Amsterdam and Philadelphia: John Benjamins.

van Doorslaer, L. (1995). "Quantitative and Qualitative Aspects of Corpus Selection in Translation Studies", *Target* 7(2): 245-260.

Winters, M. (2007). "F. Scott Fitzgerald"s *Die Schönen und Verdammten* : A Corpus-based Study of Speech-act Report Verbs as a Feature of Translators" Style", *Meta* 52(3): 412-425.

Winters, M. (2009). "Modal particles explained: How modal particles creep into translations and reveal translators' styles", *Target* 21(1): 74–97.