Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 67 – 73

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Comparable corpus approach to explore the influence of computer-assisted translation systems on textuality

Miguel Ángel Candel-Mora*

Universitat Politècnica de València, Camino de vera, s/n, Valencia 46022, Spain

Abstract

Computer-assisted translation tools have significantly influenced translators' workflow, especially with respect to productivity and consistency criteria. However, not much has been investigated about the effects and constraints that these tools have on translators' decision-making process during the language transfer stage. Therefore, the objective of this paper is to outline a framework for analysis to identify potential textual constraints due to the segmentation function of computer-assisted translation (CAT) systems and verify the extent of the effect of CAT tools on translated texts based on a comparable corpus approach.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: CAT tools; translation memories; editing; comparable corpus

1. Introduction

Computer-assisted translation (CAT) systems have their origin back in the 1970s (Somers, 2003:14), however it is not until the late 1990s when they start to be widely used by professional translators, mainly due to their competitive prices, as these tools have adapted to specific translation contexts and language pair needs, offering freelance, professional and enterprise versions; and because they have become an industry standard, and translation companies expect their translators to use them.

With respect to productivity and consistency criteria, the contribution of CAT systems on the translators' workflow is therefore unquestionable. Almost all CAT tool users agree that, in general, the use of a CAT tool helps

* Corresponding author. Tel.: +34 96 3877000 ext. 75341.
E-mail address: mcandel@upvnet.upv.es

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.420

them translate more efficiently. However, not much has been written about the effects and constraints that these tools have on the translators' decision-making process during the language transfer stage, where tools occupy a secondary position in favor of their productivity and efficiency. A quick consultation of the literature reveals a lack of research on CAT tools from the academic arena: while language industry research mainly focuses on aspects such as productivity, return on investment, and efficiency, research works on the integration and effects of these tools within the translation workflow still seem scarce.

The initial hypothesis for this empirical work lies in the assumption that the automatic division of the source text into sentences following the CAT system's default segmentation rules affects the translator's approach to the target text and it is at the revision or editing stage, outside the CAT system where these sentence shifts need to be rectified and adapted to more natural target language characteristics, especially with regards to textuality. According to Dragsted (2006: 237): "the sentence-by-sentence presentation inherent in TM systems therefore creates an unnaturally strong focus on the sentence, which affects the very task of translation (as well as the translation product)."

Text analysis has always been central for translation (Neubert & Shreve, 1992; Hatim & Mason, 1995; Reiss & Vermeer, 1996). According to Neubert & Shreve (1992:23): "The text-linguistic model of translation research maintains that an original text and a translation are different not only because their sentences are different … but also because there are constraints operating at a level beyond the sentence." Thus, the segmentation rules must necessarily have an effect on the textual characteristics of the target text. Neubert and Shreve (1992) also note that in the text-linguistic model meaning is not sentence-bound, and meaning equivalence is distributed throughout the text, not only isolated in words and sentences: meaning is also carried globally in the text.

Therefore, this paper proposes a preliminary study to explore the possibility of using a comparable corpus approach in order to identify the influence of computer assisted translation systems on textuality in translation. Neubert and Shreve (1992:83) also point out that authors often have multiple objectives when they communicate, from instruction manuals to instruct, to political speeches intended to convince, motivate, or dissuade. Given the number of variables involved in a study of this nature, this work focuses on identifying research possibilities first, and then proposes the design of specific corpora to implement this research methodology on specific types of texts.

With all this in mind, the main objective of this paper is to outline a framework for analysis to identify potential textual constraints due to the segmentation function of computer-assisted translation systems and verify the extent of the effect of CAT tools on the translated text based on a comparable corpus approach (Laviosa, 1997).

The methodology followed starts with the revision of the literature on the Text-Linguistic Model of translation research to identify potential textual constraints of CAT segmentation rules first, and then exploits the comparable corpus-based approach to investigate these constraints in translation.

The first part of this work is devoted to the revision of the literature on CAT tools (Section 2) and text studies applied to translation (Section 3) with the objective to provide a quick overview of the translation workflow with CAT tools and the description of their most characteristic functions, especially those related to the interaction and effects on the translator's linguistic decision-making processes, such as text segmentation rules.

Finally, the last section (Section 4) emphasizes the synergies between the findings from the literature review and identifies potential textual constraints of using CAT tools.

The approach here concentrates on the study of the text system from the perspective of the original text, in contrast with recent corpus-based translation studies (Laviosa, 2002:1) that seem to focus on the target text, or a comparison of both to investigate quality, translation evaluation, and universals of translation.

2. CAT tools workflow

The aim of a CAT tool or translation memory system (TM) is to allow translators to reuse previously translated work, therefore the types of texts that are best suited for working with a TM are those which are repetitive or which will be updated or revised (Bowker, 2005: 15). Regardless of the software manufacturer, most CAT tools include the following five basic functions that are active throughout the translation process: before translation starts, like the text segmentation, and project management statistics and analysis functions; during, like the text search algorithm and concordancer, or the terminology management features; or after the translation process, like the segment alignment component also used as a maintenance feature to feed the database of translation units with existing translations. As Esselink (2003: 80) points out:

> Naturally, TM tools are particularly useful on large-volume texts, which contain a lot of repetitive text and where translations can be created on a one-to-one sentence basis. Using TM tools to translate marketing text or adverts is not often a good idea, simply because those types of texts often require many adjustments, rewrites, and other modifications.

At first sight, the function most directly related to the interaction with the translator's decision-making processes in terms of language choice is the segmentation feature or the automatic division of the source text into sentences that CAT systems perform at the beginning of the translation according to segmentation rules, which can be modified depending on the translators' preferences but follow preprogrammed rules that identify where a sentence break occurs in the source text based on punctuation marks like full stops, exclamation marks, colons or tabs.

For Bowker (2005: 16) the result is a "sentence salad" rather than a text, since in order to maximize the recyclability of a text, translators working with a TM may choose to structure the sentences in the target text to match those in the source text, and they may choose to avoid using pronouns or other references:

> Therefore, if a translation is produced by recycling individual sentences from a variety of different texts – which may have different terminology, different styles, pronouns or deictics that are unclear without a larger context – the result may be more of a "sentence salad" than a coherent text. (2005: 16)

In order to verify that segmentation rules have in fact an effect on the translator's interaction with the text, a series of experiments were conducted on a small ad hoc comparable corpus of press releases in English and in Spanish.

The first test aimed at revealing the natural segmentation of texts: one English text (3081 words) was processed through a CAT tool (SDL Trados Studio) and after segmentation the total amount of segments to translate was 119. The text was translated into Spanish, and then revised and edited in a word processor for correction and fluency. When the document was finished, the text previously translated into Spanish was then processed with the same CAT tool – as if it were to be translated back into English – and after segmentation, the Spanish text had 107 segments: a difference of 12 segments less than the original text where it came from.

This simple experiment revealed that although the first draft of the translation followed the source text segmentation pattern, during the editing stage, the Spanish version was adapted to a more natural Spanish punctuation style, and the 119 segments were reduced to 107.

The second test was performed on original texts only, from the same ad hoc comparable corpus of press releases: a total of 40 press releases in Spanish and 40 in English – totaling 105.539 tokens – with the same textual characteristics and from equivalent sources in both languages. The average number of sentences per text in English was 37.7, while the amount in the Spanish corpus was 20.2. These preliminary analyses revealed that the average naturally-occurring number of sentences in the genre of press releases in English and in Spanish follow different

punctuation patterns and sentence length than those suggested by the segmentation rules of CAT tools, and therefore pave the way for further investigations on textuality, and the effects of segmentation of target texts.

3. Translation as text

Context and text analysis has always been central in translation (Neubert & Shreve, 1992; Hatim & Mason, 1995; and Reiss & Vermeer, 1996). As Muñoz points out (1995: 167) in order to understand you must have previous, linguistic and non-linguistic information. In translation it is essential to determine not only the characteristics of each individual text, but to assign it to an established classification to guide the decision-making process. According to Nida (2001: 24) "Although language is rightfully described as structurally linear, the understanding of language does not precede in merely one direction. The real meaning of a word may depend on a context that occurs on a following page. Among the text features most directly related to translation are the different functions of text type, target reader, context, communication channel and subject matter. For example, from the functional approach to translation (Reiss & Vermeer, 1996: 154) text types are defined as types of speech acts of supraindividual nature and subject to recurrent communicative speech acts, which have generated characteristic patterns in the use of language and their structure precisely due to this constant repetition. Therefore, the first step in establishing a hierarchy of levels of equivalence is to assign the text to be translated to a text type, since communicative translation involves to replace the conventions of the source culture with specific target culture conventions (Reiss and Vermeer, 1996: 138).

For example, among the differences that distinguish text-types are the intention to convey information and the informative function. According to Ciapuscio and Kuguel (2002: 56), texts with higher specialization are aimed at achieving acceptance of progress and influence experts, while more informative levels are aimed at achieving a positive attitude about science and attract interest.

Considering the specifications of a translation project of specialized texts, the first step consists in identifying the characteristics of the source text, in order to become familiar with the subject, identify potential transfer problems and define the conceptual structure of the text and, consequently guide the search for resources and reference materials, or, in other words, begin the decision-making process. Regarding formal aspects, it is essential to identify the type of publication of the source text, the audience of the original and the final purpose of that text.

From the study of professional and academic English, Alcaraz (2000: 135) notes that the technical vocabulary is not the main source of difficulty to understand a text, but the lack of familiarity of the target reader with the macrostructure of the genre, because everything has meaning in the text. Alcaraz, describes two types of macrostructure: primary, consisting of characteristic sections of the text type, and secondary, formed by the parts of each section. From the sentence level perspective, Alcaraz (2000: 23) defines register as the variety of a language designed to serve a communicative purpose in a specific professional or academic setting.

As in the translation of any text, technical or literary, full understanding of the source text requires full apprehension of the extralinguistic knowledge of the subject. The translator has to transfer the technical and scientific information to the target text, not only through precise terms but by selecting the appropriate expression according to the target audience and style of the target language. According to Neubert and Shreve (1992:14): "The translation of a technical text must… correctly reproduce the technical content of the original document both in all its details and in its entirety. It must also be linguistically correct, which means specifically that the common language components must be correct, even phraseologically correct."

Thus text typology and textual analysis applied to translation provide the adequate comprehension elements to decipher the source text and prepare the ground for the translation process and target text.

The aspects to consider when analyzing the text, which Neubert and Shreve (1992: 17) call "standards of textuality" include *intentionality*, or the interaction between the communicative purpose of a text and the readers' need for such information; *informativity*, or the knowledge contained in the text; *acceptability*, centered on the receiver and that helps classify models or patterns that apply to each type of text; *situationality*, or factors describing the social and communicative context of the text, such as time, the geographical aspect of the text, social status, age and gender of the target reader; *coherence*, i.e., the logical relationships between information units; *cohesion*, or expression of semantic relations and the structure of information within the text through linguistic, lexical or grammatical elements; and *intertextuality*, or relationships between one type of text and other texts of the same type.

For Muñoz (1995: 239) cohesion is manifested through the grammatical elements of the text, such as the connection between distant lexical items by anaphora, cataphora and ellipsis; the connection of textual segments such as connective elements with specific syntactic patterns, and specific uses of the lexicon such as repetitions, substitutions, etc. Cohesion is part of a language system, of its grammar, for example, the connection between a pronoun and its antecedent is a grammatical device.

Vázquez-Ayora (1977) states that the target text must meet the following conditions to be considered correct: the translation must be correct from a technical standpoint; terminologically correct with respect to the general language; it has to represent the technical language in its proper context, meet the requirements of text types and should take into consideration the characteristics expected in the target culture.

In the field of academic and professional English studies, the concept of genre (Alcaraz, 2000: 133) is also used for the typological classification of professional and academic texts that share the following conventions: communicative function, organization (macrostructure), modality and discursive techniques, lexical-semantic level, and sociocultural contexts.

Finally, the typology presented by Ciapuscio and Kuguel (2002: 44-45) integrate the different views on specialized text for existing typologies and note that text types have their origin in multidimensional systems of prototype representations at various levels. Text typology for these authors falls into four levels: function, situation, semantic content and form. Among the advantages of this classification is the distinction between different degrees of specialization based on linguistic and textual criteria.

4. Potential textual constraints of CAT tools segmentation rules

Thanks to the advances in technology and corpus linguistics studies, there seems to prevail a research model based on translation as product. Corpus linguistics studies are also relatively new in translation, although its research methods are fully consolidated (Baker, 1995; Candel-Mora & Vargas, 2013; Laviosa, 2002; Olohan, 2004). However, despite the amount of possibilities for study and analysis techniques of the corpus-based approach, it seems that most research focuses on the exploitation of parallel corpora to study the effect of the original text on the translated text and more specifically, on translation universals, translationese or the language of translation (Baker, 1995; Laviosa, 2002; Olohan, 2004). According to Olohan (2004: 35) "In Translation Studies the most common corpora used are: corpora of comparable original texts in two or more languages corpora of original texts and their translations into two or more other languages.

In this work, the comparable corpus-based approach has been chosen because it allows to identify patterns that occur independently of the language of origin. A comparable corpus here is defined as a corpus that consists of texts in more than one language, not containing translations and built using the same design criteria and the same sampling techniques. During the preliminary literature review, the different approaches of the interdisciplinary approach to the study of translation revealed a lack of consensus in the definition of comparable corpus, as some authors include in translations and their original texts (Laviosa, 2002).

Following the aim of this work to outline the potential of using a comparable corpus to study the textuality constraints of using CAT tools, among the corpus analysis techniques to obtain textual data, Biber et al. (1998: 23) suggest the following: key words in context (KWIC) to study the different meanings of the same word in different contexts, analysis of frequencies, non linguistic associations, and typical collocations to apprehend the typical features of certain types of texts, or field of expertise, and the different grammatical functions of a word.

Biber et al. (1998) stress the importance of the evidence obtained from the corpus because they describe the actual use in real contexts, which has its most immediate application in translation, such as sentence length, concordances and co-text, frequency lists, lexical density, and keywords.

From the corpus-based translation studies approach, according to Olohan (2004:73), concordances are commonly used to study distinctions between different usages, patterns and meanings, for which a wider co-text is necessary to understand the functions of items in discourse; frequency lists to isolate unusual or creative items; and lexical density which has been studied to demonstrate simplification in a corpus. From the point of view of the study of translation universals, the role of lexical density has also been identified by authors like Baker (1999: 184): "the individual texts in an English translation corpus are more like each other in terms such as lexical density, type-token ratio and mean sentence length than the individual texts in a comparable corpus of original English."

Unquestionably, the first visible effects of segmentation is the interference of source on target language, whether in terms of translationese, or the language of translation, or at the lexical level, as distributional anglicisms – in terms of Vázquez-Ayora (1997: 102) for the pair of languages of the case under study. That is, when the translator does not select the most suitable correspondence offered by the target language and copies the most similar form which has a high frequency of usage and then creates an anomaly that spreads throughout the entire translation.

Therefore, according to the literature review on corpus-based translation studies, the most frequent potential textual constraints of CAT tools segmentation rules are: coherence, cohesion, orthotypography, anaphoric and cataphoric references, linearity, and readability, with variable level of difficulty in their analysis.

Coherence and cohesion are probably the most difficult features to identify through corpus techniques, unless the corpus in use is annotated, however it can be performed manually if necessary. According to Thornbury (2010: 272) "Corpus tools cannot easily detect cohesive ties, such as pronominal reference, unless they have been tagged as such." Although as pointed out by Bowker (2005: 15), both features – cohesion and coherence – are closely associated with the CAT tool translation workflow "Following claims of increased productivity, the second most often cited benefit of working with a TM system is that it improves translation quality by increasing consistency. A translator who is working on a long document is able to maintain consistency throughout the text."

Orthotypography may fall into the category of distributional anglicisms and it is clearly as an interference from the English source text like in the case of the serial comma, which in Spanish does not exist, the use of the colon, capital and lower case letters or quotation marks, to mention some of the most relevant (de Sousa, 2010).

Polysemy and anaphoric references – within the same texts and inter texts of the same type – is definitely subject to appear due to the limited display of information on the text. Especially when because of the extension of the text, the translation project is distributed between different translators.

Anthony Pym's (2010: 3) research on the effects of technology on translation highlights *linearity* and notes that CAT tools break texts into units and its linearity is repeatedly interrupted: "The translating mind is thereby invited to work on one segment after the other, checking for terminological and phraseological consistency but not so easily checking, within this environment, for syntagmatic cohesion."

As for readability, Nida (2001: 24) notes that the "process of reading is essentially based on the principle of reading by contexts rather than by lines, since so frequently the meaning of words depends on what follows rather

than on what precedes.

6. Conclusions

Assuming the key role of the text and the existence of constraints beyond the sentence level, this work followed the text-linguistics model of research in translation and the use of comparable corpus analysis techniques as this type of corpus allows identifying patterns that occur independently of the language of origin.

Based on the preliminary literature review conducted on text-linguistics and corpus-based translation studies, the framework for analysis of potential textual constraints of CAT tools segmentation rules may entail complications during the transfer process in the following features: coherence, cohesion, orthotypography, anaphoric and cataphoric references, linearity, readability, all of them with variable level of difficulty in their analysis.

It can be concluded that the CAT tools editing interface is not suitable for text revision towards textuality and the ideal texts for further research would include texts with intra and inter textual references like press releases.

Further research is suggested also to study translation strategies that require further processing in terms of syntax and change in sentence structure.

References

Alcaraz, E. (2000). *El inglés profesional y académico*. Madrid: Alianza Editorial.

Baker, M. (1995). Corpora in translation studies: an overview and some suggestions for future research, *Target* 7: 223-43

Baker, M. (1999). The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators. *International Journal of Corpus Linguistics* 4(2), 281–298.

Bowker, L. (2005). Productivity vs Quality? A pilot study on the impact of translation memory systems. *Localisation Focus* 4(1): 13-20.

Biber, D., Conrad S., & Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Candel-Mora, M. A., & Vargas-Sierra, C. (2013). An Analysis of Research Production in Corpus Linguistics Applied to Translation. *Procedia* 95: 317–324.

Ciapuscio, G. & Kuguel, I. (2002). Hacia una tipología del discurso especializado: aspectos teóricos y aplicados. In García Palacios, J. & Fuentes Morán, T. (Eds.) *Texto, Terminología y Traducción.* (pp. 37-74). Salamanca: Almar.

de Sousa, J. M., (2003). Los anglicismos ortotipográficos en la traducción. *Panace@*. 11:1-5

Dragsted, B. (2006). Computer-aided translation as a distributed cognitive task. In Dror, I, & Harnad, S. (Eds.) *Cognition Distributed. How cognitive technology extends our minds* (pp. 237–256). Amsterdam and Philadelphia: John Benjamins.

Esselink, B. (2003). Localisation and translation. In Somers, H. (Ed.) Computers and translation (pp. 67–86). Amsterdam: John Benjamins.

Hatim, B., & Mason, I. (1995). *Teoría de la traducción: Una aproximación al discurso*. Barcelona: Ariel.

Laviosa, S. (1997). How comparable can "comparable corpora" be?, *Target* 9: 289-319

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

Muñoz Martín, R., (1995). Lingüística para traducir. Barcelona: Teide

Neubert, A., & Shreve, G. (1992). *Translation as Text*. Kent, Ohio: Kent State University Press.

Nida, E. (2001). *Contexts in Translating*. Amsterdam: John Benjamins Publishing.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London/New York: Routledge.

Pym, Anthony. (2011) What technology does to translating. *Translation and Interpreting* 3(1):1-9

Reiss, K., & Vermeer, H. (1996). *Fundamentos para una teoría funcional de la traducción*. Madrid: Akal.

Somers, H. (2003). Translation memory systems. In Somers, H. (Ed.) *Computers and translation* (pp. 31–47). Amsterdam: John Benjamins.

Thornbury, S. (2010). What can a corpus tell us about discourse? In O'Keeffe, A., & McCarthy, M. (eds.) *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

Vázquez Ayora, G (1977). *Introducción a la traductología*. Washington: Georgetown University Press.