Cahiers du plurilinguisme européen
ISSN : 2105-0368

2 | 2010
Outils de traduction – outils du traducteur ?

Source text quality and computer-assisted translation

Franco Bertaccini and Mara Rocchi

https://www.ouvroir.fr/cpe/index.php?id=278

DOI : 10.57086/cpe.278

Electronic reference
Franco Bertaccini and Mara Rocchi, « Source text quality and computer-assisted translation », *Cahiers du plurilinguisme européen* [Online], 2 | 2010, Online since 09 mars 2023, connection on 25 mars 2023. URL : https://www.ouvroir.fr/cpe/index.php?id=278

Source text quality and computer-assisted translation

Franco Bertaccini and Mara Rocchi

OUTLINE

TEXT

Introduction

1 The purpose of the present report is to determine whether CAT tools can enable multiple translators to produce a homogeneous target text. This paper contains a detailed description of all the stages of the experiment, as well as a brief discussion of the final result.

2 The experiment, born as a graduation thesis project, was carried out at the Advanced School of Modern Languages for Interpreters and Translators (SSLMIT) in Forlì. The whole work was coordinated by Franco Bertaccini, professor of terminology at the SSLMIT, and involved two Italian translation students, Mara Rocchi and Simona Ruggeri, both of whom had majored in English for their first degree.

3 The students were required to translate different portions of a software tutorial from Italian to English, starting with a common translation memory and terminology database, which were to be updated and shared at the end of each individual translation session.

Method

4 The first step of the project consisted in selecting the text to be translated. Taking into account the fact that the experiment focused on the efficiency of CAT tools, and that the most suitable texts for this kind of approach are those belonging to the legal, scientific or

technical fields (Bowker, 2002: 113), it was decided to translate part of the tutorial of Instant Developer, a software program created by an Italian company to design Web applications.

5 The reasons for this choice lay in the fact that the selected text presented two main functions of interest. On the one hand, since it was conceived as a product for programmers, the text was mainly informative and contained terminology specific to the field of information technology. On the other hand, being a tutorial, it presented phraseology that is common to directive texts.

6 The part of the tutorial selected for the translation consisted of about 20,000 words and was divided into six chapters, each containing a number of lessons relating to a particular subject. Special care was devoted to the division of the text. It seemed that dividing it simply into two blocks might compromise the final result of the experiment. As a consequence, the first half of the text was translated alternately, while the second half of the text was divided into two blocks of two chapters each. This made it possible to obtain a complete overview of the translation behaviour of each student and to check whether the final reader would react differently to these portions of text.

7 After selecting and dividing the source text, the students were asked to carry out a preliminary reading of some portions of it. A number of problems of different nature were detected at the grammatical level, as well as problems in the logical sequence of information and a certain inconsistency in the use of specialized terminology.

8 On this subject, Francesco Sabatini, president of the *Accademia della Crusca*[1], emphasizes the lack of standardization in the Italian language of information technology. On the one hand, this may be due to the fact that the "development of computers took place almost exclusively in English-speaking countries" (Covington, 1981: 66). On the other hand, evidence has been provided that "new software technologies continue faster than the work of the linguists of the language academies, and translators have to come up with new terms" (Mendiluce-Cabrera & Bermudez-Bausela, 2006: 452). These factors often lead to the existence of a number of different terms identifying the same concept, as well as an excessive – and often unjustified – use of anglicisms.

9 Besides highlighting a number of grammatical and syntactic errors, this step helped the students decide what kind of corpus to compile. As Bowker and Pearson (2002: 38) point out, a special-purpose corpus can help translators deal with one of the main challenges of their job, that is to say, the need to become "mini-experts" in a subject field as fast as possible. A parallel corpus seemed to be the best choice in this situation, as it would have allowed the students quickly to retrieve the most common terms, phrases and collocations of the selected subject field and to obtain conceptual information by looking at terms in context using bilingual concordances.

10 Another thing that was established at this point was the size of the corpus and the type of texts to be used to create it. As far as corpus size is concerned, taking into account that there were two people involved in the project and that they had planned to spend a good amount of time on this phase, an attempt was made to build a corpus that was substantial and customized at the same time. However, it would have been impossible to analyse the documentation existing in the field of information technology, even by narrowing the topic. For this reason, it was decided to retrieve documentation that was representative of the programming language.

11 As regards the selection of texts to be included in the corpus, it was agreed that specific software product documentation should be used in order to ensure that both the conceptual and linguistic needs were met. Since subject-field experts have their own preferences, it was also decided to consult the firm for advice. A clear preference was expressed for the tutorial of either *Microsoft Visual Studio or IBM Rational Rose* and this suggestion narrowed the scope of the search to two websites, solving two main problems. First of all, selecting one of these tutorials would not pose the problem of authorship, as both Microsoft and IBM are well known IT leaders and are subject-field experts par excellence. Moreover, as it had been planned to build a parallel corpus, there were good probabilities of finding both the English and Italian versions of the tutorials on these websites. In this way, it was possible to build a bilingual corpus composed of a collection of original English texts and their translation into Italian.

12 A brief analysis of the official websites of Microsoft and IBM was carried out and, since both provided high-quality texts, the main factors

considered in this phase were the website's user-friendliness and the table of contents of the tutorials. Microsoft was finally selected and, among the different versions of *Visual Studio* available, *Visual Studio .NET* was chosen for the terminology purposes of the present project. At the time when the selection was made, this was the latest release of *Visual Studio*. The corpus about to be built was therefore up-to-date with the latest know-how in software technology.

13 More than 300 pages per language were downloaded from the tutorial of *Visual Studio .NET*, for a total of 1,071,993 words. As can be imagined, this phase took a lot of time due to a series of factors. Bowker and Pearson (2002: 62-66) provide a comprehensive list of the problems that may arise when using texts in electronic form rather than printed texts to build corpora. These problems will not be discussed in detail in the present paper, but a couple of them are worth mentioning. As Bowker and Pearson (2002: 63) point out, "the very nature of the Web is that it makes use of hyperlinks, which means that although a web site as a whole may contain a lot of information, each individual page may contain relatively little data"», which indeed proved to be the case. Statistical data extracted from the English corpus show that a total of 518,011 words are spread over 312 pages, giving an average of 1660 words per page. As far as the Italian corpus is concerned, a total of 553,982 words are spread over 312 pages, giving an average of 1775 words per page.

14 Another problem posed by texts downloaded from the Web is that they are encoded using HyperText Markup Language. As a consequence, when downloading the texts from the Web, it was not possible to directly copy them to a word processor as they would have retained their graphics and formatting. For this reason, each page had to be saved as a plain text file first, then copied to a word processor and finally saved in .rtf format.

15 It should be emphasized that this step did not consist of a mere mechanical task of downloading and saving pages. In this phase, the students performed the real selection of texts by quickly reading the articles and determining which ones should be included in the corpus. This also allowed them to gain some background knowledge and to understand some of the key concepts in the subject field they were required to become familiar with.

16 Once all the texts were downloaded and saved, they were ready for alignment. *SDL Trados WinAlign* was the alignment program selected for this purpose. Since the students were using *Trados 7 Freelance*, they had to create a number of *WinAlign* projects containing 10 file pairs at most. On the whole, it can be argued that the alignment process was quite satisfactory as it did not pose any serious problems. In many cases, when a space between a full stop and the beginning of a new sentence was missing, the segments needed to be manually split and linked. In other cases, the alignment was incorrect due to the fact that entire sentences had not been translated into Italian. In these cases, the non-translated segments were identified and removed from the project, and the correct links between the remaining segments were restored manually.

17 The aligned segments were then stored in a new translation memory, which included a total of 27,661 translation units. At the beginning of the project, it had been established that the translation memory should contain at least 20,000 translation units and, as their total amount exceeded the expected quantity, the project coordinator confirmed that the collected material was enough and that there was no need to add new texts. Moreover, the size of this TM was expected to increase during the translation process, as it had been planned to enlarge it interactively during each translation session. Looking at the size of the newly created translation memory, it seemed that there were good bases for a successful outcome of the project. However, a number of other steps had to be taken to create the terminology database before starting the translation process.

18 Due to the size of the TM, it was necessary to use *SDL MultiTerm Extract* to extract bilingual terminology automatically. The automatic extraction resulted in more than 10,000 term-pairs that needed to be checked and validated. At this point, the validation phase began and was performed according to three different stages.

19 The first stage was the validation of the extracted term-pairs, provided that they were correct. If there was no correspondence between the two terms, bilingual concordances were used to check the appropriate translation and to correct it. In some cases, concordances made it possible to find out that more than one translation existed for a given term. In these cases, the term-frequency was ana-

lysed in order to determine whether or not multiple translations should be entered in the terminology database. However, when more than two translations existed for a certain term, the tendency was to enter only the two more frequent solutions, as it was thought that an excessive number of translation possibilities might have caused inconsistencies and confusion during the translation process. Moreover, since bilingual concordances would be available during translation, the omitted possibilities could be checked, and possibly used, in that phase.

20 The second stage was the completion of partially extracted term-pairs. In practice, there were a lot of incomplete term-pairs, where only the English term had been extracted, but its translation was missing. Even in these cases, we found that the appropriate translations could be retrieved by means of bilingual concordances and, when multiple translations were available, the parameter mentioned above was used during compilation. What seemed to be somewhat strange is that most incomplete term-pairs involved individual words which should not have posed problems in the way that compounds or phrases might do.

21 The third stage was the insertion of new term-pairs. It can easily be imagined that a program for automatic terminology extraction will never be able to retrieve all the necessary term-pairs relating to a certain subject field. For this reason, while concordances were being consulted for the above-mentioned purposes, any potential term or phrase of interest was checked and, if not present, was manually added to the list.

22 Besides inserting, completing and validating term-pairs, an attempt was made to ensure consistency within the terminology database. A series of operations was performed in order to standardize the format of some abbreviations. For example, it had been noticed that file formats were expressed in different ways (XML, xml, .xml) and, after having analysed the bilingual concordances to identify the most frequent style, it was decided to adopt a common version for all of them. Since particular care had to be taken in the validation of each term-pair and a number of criteria had to be followed, this proved to be the longest and most laborious phase of the project, and it took more than a month to validate and complete the list of terms.

23 At the end of this phase, everything was ready for the creation of the terminology database. The term-pairs were exported to *SDL MultiTerm* 7, where a new database was created for that purpose.

24 With a translation memory composed of 27,661 translation units and a terminology database with 7,027 entries, the students could now start the translation process.

25 Professor Bertaccini suggested that the translation process should consist of different steps to be performed every day. First, some articles would be translated by each student with the aid of the translation memory and terminology database generated. Then, the translated texts would be aligned, loaded into the translation memory, and finally exchanged by the students at the end of each translation session. In this way, the translation memory would always be up-to-date and the students could benefit from each other's work in real time.

26 However, from the very first day, a number of problems arose during the translation process that made the students think that this method would not work. Probably, the main problem was that there were a number of expressions where the order of constituents gave rise to different interpretations. For example, in a sentence like il *DB Code è il nome fisico sul database della tabella*, it might be assumed that the DB Code is a physical name stored in the table database. However, since this is not plausible, it can be easily understood that the database is the real location of the DB Code and that sul database (in the database) should be placed at the end of the sentence, which is something a CAT tool will never be able to do.

27 Another problem was represented by anglicisms. For example, words like "background", "folder", "font" "menu bar" "scrollbar", "statement", "template" or "toolbar" were often used instead of the Italian terms *sfondo*, *cartella*, *carattere*, *barra del menù*, *barra di scorrimento*, *istruzione*, *modello* and *barra degli strumenti*. While it must be recognised that anglicisms partially helped the students during the translation process, it should be pointed out that they also compromised the efficiency of the translation memory. In a paper devoted to French computer terminology, Gray (1985: 806) argues that the speakers of a certain language should borrow specialized terminology only when there are terminological gaps in their language. However, since the Italian version of the above-mentioned terms not only exists but is

also of common usage, it may be maintained that the author misused this lexical resource.

28 The situation was worsened by the fact that various terms appeared in both their English and Italian versions. For example, the word "folder" also appears as *cartella*, which is the appropriate Italian translation for folder, and even as *cartellina*, a diminutive of *cartella* that should have been avoided in this kind of text. This example introduces another problematical feature of the source text, that is to say the use of synonyms. The most representative example of this tendency is represented by the Italian for "assignment statement". In fact, five different versions were detected in the source text for this compound, four of which appear in the excerpt reported below (bold mine):

> *In questo caso è disponibile la sola* ***assegnazione*** *per la variabile Order Status. […] Dato che abbiamo scelto* ***l'operazione di assegnazione****, ci viene richiesto di editare l'espressione che deve essere assegnata alla variabile Order Status. […] La stessa operazione (creazione di uno* ***statement di assegnamento****) poteva essere fatta direttamente tirando la variabile sul body della procedura o su uno dei blocchi. In questo caso Instant Developer crea un nuovo* ***statement di assegnazione*** *come ultima istruzione del blocco su cui tiriamo la variabile*[2].

29 The fifth version is *statement di tipo assegnazione*. It is interesting to note that none of these five options is the appropriate translation for assignment statement, as the most appropriate one would be *istruzione di assegnazione*.

30 Finally, it should also be mentioned that a number of cases were detected in which a wrong term had been used. For example, *categoria preferita* (preferred category) was found instead of *categoria predefinita* (default category). In another case, the author wrote *campo di pannello* (panel field) instead of *campo di tabella* (table field).

31 The factors analysed so far caused a reduction in the efficiency of CAT tools, which, in most cases, could not retrieve the desired terminology and phraseology. The translation memory was particularly affected by grammatical and terminological problems, and segments were almost never found. On the other hand, the students were able

to benefit from the terminology database, as this proved effective in retrieving and checking terminology.

32 In order to preserve the quality of the translation memory and terminology database that had been created from the documentation of *Visual Studio .NET*, it was decided to create a copy of the translation memory in which translated segments could be loaded, and to build a new terminology database in which new terms and phrases extracted during the translation process could be entered. Taking into account the faulty expressions and inexact formulations of the source text, the students did this with the awareness that the new terminology database and the updated translation memory would never and could never be used for purposes other than this translation.

33 Once the translation process was complete, the next step should have consisted in submitting the translation to a native English speaker, who was supposed to evaluate whether the presence of more than one translator could be perceived. However, it was clear that there was work still to be done on the target text, as too many problems had interfered with the translation process. The students opted for a peer revision, which was carried out on the text according to the order of appearance of the articles. Among the adjustments made to the text, the main task consisted in standardising some article sections, such as abstracts and introductory sentences, in order to improve the coherence and cohesion of the text. Special care was also taken in standardising the use of case, as this was another factor that had been neglected in the source text. As far as terminology is concerned, when two or more translation options were available for a certain term or expression and the students had opted for two different translations, bilingual concordances were consulted in order to establish the most common and/or appropriate option. On the whole, it may be said that the peer revision involved major changes at the stylistic level and minor terminological adjustments. During the peer revision, the students also identified further sentences that might be interpreted in different ways due to the ambiguous order of constituents and tried to determine the most appropriate interpretation.

34 At this stage, it became clear that only a meeting with the expert could resolve some of the main translation problems, and it was de-

cided to contact the firm. The resolution of doubts with the aid of the specialist must be viewed as another fundamental step of the translation process, even if it took place after the peer revision. Besides throwing light on ambiguous sentences, the specialist's explanations helped the students understand concepts and mechanisms they had not been able to clarify before, confirmed some of their assumptions about verbs and terms that had been used inappropriately and highlighted a number of terms that had been created by the team of developers working at *Pro Gamma* and for which a translation did not yet exist. The target text was then established according to the specialist's explanations.

35 When everything was ready for the last step of the experiment, three articles were selected and submitted to Dr. Derek Boothman, a native English speaker and professor of translation at the SSLMIT, who was required to read them and to express his opinion about the coherence and cohesion of the text by filling in a questionnaire. One of the questions was "Do you think that this text was translated by an individual translator or by more than one translator?" and his answer was "I do not single out any stylistic differences. Therefore, this text may have been translated by an individual translator".

36 At first sight, it might be thought that the experiment was successful due to the fact that the final reader did not perceive the presence of two translators behind the target text. However, it is more appropriate to maintain that the outcome of this experiment was a partial failure and a partial success. As explained in the previous paragraphs, the students needed to schedule additional steps before submitting the text to the English reader and there are reasons to think that if the students had not carried out these steps, the target text would not have been homogeneous and the experiment would have failed. Therefore, it should be acknowledged that the peer revision performed *a posteriori* played a pivotal role in the final result of this experiment.

37 As far as CAT tools are concerned, it may be maintained that they did not work as expected and that this lack of efficiency is mainly to be ascribed to the low quality of the source text. The translation memory was particularly affected and segments were almost never found. On the other hand, the students often consulted the termino-

logy database to check or retrieve terms and phrases, and also made use of bilingual concordances, which proved very helpful in retrieving "information on common authentic usage not available in even the best bi- or monolingual domain-specific dictionaries, glossaries, databases and any other resources" (Friedbichler & Friedbichler 2000: 108). Given the size of the text to be translated and taking into account that the translation memory was enlarged interactively, it cannot be denied that these tools helped the students during the translation process.

Conclusion

38 This practical experiment produced interesting results that can be used to make a number of observations on the role played by CAT tools in technical translation, as well as on future actions to take to improve technical texts.

39 As far as CAT tools are concerned, it must be acknowledged that they have become absolutely necessary for translators. However, their importance should not be overestimated as, even in specialized texts, there remains a wide margin of interpretation and rephrasing which requires the presence of a human translator. Fortunately for translators, the most obvious conclusion to this experiment is that "the task of translation [...] requires human capabilities which, for the time being at least, cannot easily be simulated by a computer program" (Somers, 1997: 194).

40 The first remark concerns the quality of technical texts, which plays a very important role for translators. So far, it has been shown that a low-quality source text can heavily compromise the efficacy of CAT tools, thus increasing the risk of producing an incoherent target text. On 30 September 2000, on the occasion of the International Translation Day, Esteves Ferreira (2001: 11) pointed out that technical translations cover more than 80% of worldwide translations, which means that the majority of technical texts are written to be translated. As technical documentation is mainly authored by subject-field experts, not only should they care about the transmission of contents, but also understand that technical writing implies objective reading. Ambiguity and subjective reading in these kinds of text may cause serious problems in the transmission of contents from one language to

another. Therefore, when writing technical documentation, subject-field experts should become subject-field writers and they should not forget to pay attention to both content and form.

BIBLIOGRAPHY

BOWKER Lynne, 2002, *Computer-Aided Translation Technology: A Practical Introduction*, Ottawa, University of Ottawa Press.

BOWKER Lynne, PEARSON Jennifer, 2002, *Working with Specialized Language: a Practical Guide to using Corpora*, London, Routledge.

COVINGTON Michael A., 1981, "Computer Terminology: Words for New Meanings", in *American Speech*, 56, 1, p. 64-71.

ESTEVES FERREIRA Joao, 2001, "La technologie, le traducteur et son client", in *Traduire*, 188-189, p. 10-21.

FRIEDBICHLER Ingrid, FRIEDBICHLER Michael, 2000, "The Potential of Domain-Specific Target-Language Corpora for the Translator's Workbench", in BERNARDINI Silvia, ZANETTIN Federico, (ed.), *I corpora nella didattica della traduzione: atti del Seminario di studi internazionale*, Bertinoro, 14-15 novembre 1997, Bologna, Clueb, p. 107-116.

GRAY Eugene F., 1985, "French Computer Terminology", in *The French Review*, 58, 6, p. 805-810.

MENDILUCE-CABRERA Gustavo, BERMÚDEZ-BAUSELA Montserrat, 2006, "Sci-Tech Communication: Is There a Process of Internationalization in English and Spanish?", in *META*, 51, 3, p. 445-458.

PUNTO INFORMATICO, PI: *L'esperanto può cambiare l'informatica?*, [http://punto-informatico.it/servizi/ps.asp?i=1346503], last accessed 12 April 2008.

SOMERS Harold, 1997, "Practical Approach to Using Machine Translation Software: 'Post-editing' the Source Text", in *The Translator*, 3, 2, p. 193-212.

NOTES

1 PUNTO INFORMATICO, PI: *L'esperanto può cambiare l'informatica?*, [http://punto-informatico.it/servizi/ps.asp?i=1346503], last accessed 12 April 2008.

2 "In this case, for the Order Status variable, the assignment operation is the only one available. [...] As you have chosen the assignment operation, you are required to specify the expression to be assigned to the Order Status variable. [...] You could have performed the same operation (creation

of an assignment statement) by dragging the variable onto the procedure body or one of its blocks. If you had done this, Instant Developer would have created a new assignment statement at the end of the block onto where the variable was dragged."

ABSTRACTS

English

This paper presents a project aimed at establishing whether, with the aid of CAT tools, it is possible to obtain a target text where the presence of two or more translators is not perceived by the reader. The project involved two Italian students, who were required to translate from Italian to English the tutorial of *Instant Developer*, a programme used to create Web applications. Before beginning the translation, a translation memory and a termbase were created. First, the texts of *Microsoft documentation* on *Visual Studio 2006* were downloaded and aligned using *SDL Trados*, then the relevant terminology was extracted with *Trados Multiterm*. In this way, the students started working on the text with a common translation memory and termbase which were updated daily at the end of each translation session. Due to faulty expressions in the source text, the translation memory proved to be almost useless, as segments were hardly ever found. On the other hand, the termbase was frequently used to obtain and check terminology. However, this was not enough to produce a homogeneous target text, so that a cross-revision was required. The partial failure of this experiment was largely due to the faulty expressions in the source text, which greatly compromised the efficacy of the CAT tools. For this reason, it can be said that the effectiveness of CAT tools should be acknowledged but not overestimated as, even in specialized texts, there remains a wide margin of interpretation and rephrasing which requires the presence of a human translator.

Français

L'étude présente un projet visant à déterminer si l'utilisation des outils de TAO permet d'obtenir un texte cible où la présence de deux ou de plusieurs traducteurs ne serait pas perçue par le lecteur. Le projet impliquait deux étudiants italiens chargés de traduire le tutoriel d'*Instant Developer*, un logiciel utilisé pour la création d'applications Web. Avant de commencer la traduction, une mémoire de traduction et une base terminologique furent créées. Les textes de *Microsoft documentation* sur *Visual Studio 2006* furent téléchargés et alignés avec *SDL Trados* et la terminologie pertinente fut extraite avec *Trados Multiterm*. Ainsi les étudiants commencèrent à travailler sur le texte avec une mémoire de traduction et une base de données communes mises à jour à la fin de chaque séance de traduction. Du fait de la présence d'expressions erronées dans le texte source, la mémoire de traduction s'est révélée quasiment inutile, les segments n'étant jamais retrouvés. D'autre part, la base terminologique fut fréquemment utilisée pour l'ob-

tention et la vérification des termes, ce qui ne suffit toutefois pas pour produire un texte homogène. Une révision croisée fut alors décidée. L'échec partiel de cette expérience peut largement être attribué aux expressions erronées figurant dans le texte source, ce qui a lourdement compromis l'efficacité des outils de TAO. Pour cette raison, nous pouvons affirmer que l'efficacité des outils de TAO doit être reconnue sans être surestimée, puisque même dans les textes spécialisés il reste une marge très large d'interprétation et de reformulation demandant la présence d'un traducteur humain.

INDEX

Mots-clés
base de données terminologiques, mémoire de traduction, qualité du texte source, rédaction technique, TAO

Keywords
CAT, source text quality, technical writing, terminology database, translation memory

AUTHORS

Franco Bertaccini
University of Bologna. Franco Bertaccini is a professor of terminology and director of the Laboratory of Terminology Research at the Advanced School of Modern Languages for Interpreters and Translators (SSLMIT) in Forlì. He is a member of the Scientific Council of the Italian Association for Terminology (Ass.I.Term) and works as a terminology and terminography consultant for the European Union.
IDREF : https://www.idref.fr/189304731
ISNI : http://www.isni.org/0000000117291084

Mara Rocchi
University of Bologna. Mara Rocchi graduated in translation at the Advanced School of Modern Languages for Interpreters and Translators (SSLMIT) in Forlì and is currently working as a freelance translator.
IDREF : https://www.idref.fr/26557630X