Investigating the Use of TM Tools in Translating IBC into Chinese: A Pilot Study on SDL Trados Studio

Hao Wu*
School of Foreign Languages, Tianjin Chengjian University, Tianjin, China
*Corresponding author e-mail: hao.wu_scar@tcu.edu.cn

Abstract. Being one of the most widely used software applications in the localisation of digital information, translation memory (hereinafter referred to as “TM”), along with other translation technologies constitutes a new yet important field of interdisciplinary study between computer science and translation. By assessing the application of TM in the subdomain of Civil Engineering in scientific translation, this pilot study sheds light on the strengths and shortcomings of TM and suggests possible areas for further research.

1. Introduction
In this essay, I propose to investigate the use of TM in SDL Trados Studio 2014 (hereinafter referred to as “Trados”) for scientific translation in the subdomain of Civil Engineering. I will begin with a brief introduction on the evolution of TM tools, key features pertinent to TM in Trados and the rationale for choosing International Building Code 2012 (hereinafter referred to as “IBC”) [1] as the source material. Subsequently, I will proceed to account for my use of TM tools in translating IBC and evaluate the strengths and shortcomings of TM system in Trados. I will conclude by proposing field of future research.

2. An Overview of TM tools and Rationale for Choosing IBC as Source Text
Nowadays, TM systems are the most widely used software applications in the localisation of digital information [2]. In this section, I will account for the choice of TM systems and provide rationale for choosing IBC as source text. A brief introduction of the key features in Trados will be given for a better understanding of this research. As TM systems works better in specialised text, in which specialised terms occurs more often and the consistency of terminology is of vital importance, IBC can provide a repetitive source text of this sort for the purpose of this research.

2.1. Evolution of TM Tools and Trados
The use of translation technology as an aid to traditional human translation can be dated back to 1971, when Krollmann proposed to have a “linguistic data bank” [3], with one of the subbanks being translation archive, to better support the translators’ work. Since then, translation tools for professional translators have appeared and developed. The necessity of translation technology has been testified over time.

Among the most famous commercial workstations for professional translators, not only was Trados the first to use the term “translation memory”, it was also the first to incorporate TM and alignment facilities into its workstation [4]. TM was later termed as the repository of previously translated texts divided into aligned translation segments for further use [5]. The concept behind TM is
to leverage previous translations and alleviate translators from repetitive work. TM also has the advantage of reducing cost and time, and maximising consistency.

Trados is popular among Chinese translators as well, with 85% market share [6]. In years of practice, Trados has proven to be most productive when translating texts from specialised field which demands high consistency of the repetitive terms [7].

What make Trados stand out are the key features pertinent to TM in Trados, such as Context Match, Concordance Searching, Alignment tools and Autosuggest.

Other than matches from TM, context match from the context is suggested for the sake of leveraging previous translations. Context matches are very similar to 100% matches and can be counted as exact matches. Fuzzy matches include all matches with match rate below 100% [8]. Depending on the parameters set beforehand, fuzzy matches are shown to the translator with auto suggestions one can refer to or follow [5]. Concordance searching is designed to reveal matches from TM which are below sentence level. Alignment tools are used when creating and adding TM resources from legacy bilingual documents.

The use of Multiterm can be integrated in translation project as well. Together with Autosuggest function, the integration of prepared termbase can contribute to productivity and accuracy. Different from matches from TM, Autosuggest functions on a fragment basis, while TM allow matches on a sentence (segment) basis [8].

2.2. Rationale for Choosing IBC as Source Text
TM system works better in specialised areas where texts are repetitive and functional [5]. The translation of terminology is of vital importance in that inconsistency of terminologies would contaminate the quality of these documents. Therefore, in these texts, consistency is always a priority [9, 10]. While Merkel claims that the results of TM tools is satisfactory as long as the source text is repetitive and of good quality [11], others argues that whether the technical text is repetitive or not, TM tools would always contribute to the consistency of terminologies [5].

The text I chose is extracted from International Building Code 2012, a recognised technical specification used by civil engineers. The rationale for choosing the 153 sentences (2,605 words) from the first 12 sections of Chapter 16 Structural Design is that this part is less descriptive and consists of more repetitive texts to different extent. I decided to narrow down the volume of the translation work to retain working memory of myself to ensure that I can detect the inconsistent segmentations between translation units.

3. My Use of TM Tools and Evaluation of TM Tools in this Project

3.1. My Use of Trados in Translating IBC
To better process the data collected, I opted to use Sketch Engine. Being one of the most powerful online corpora, Sketch Engine is available in many languages including English and Chinese. Terms are extracted beforehand using one of Sketch Engine’s core functions, terminology extraction. Terminology Extraction functions on the basis that certain keywords and terms occur more frequently than they do in general corpora, like BNC [12]. A corpus is compiled based on the sections pertinent to “load” in Chapter 16 and is interrogated for specialised terminologies. While I consult professionals in the field of civil engineering, some of the terms used in English do not have equivalent forms in Chinese. In those cases, literal translation is used for the sake of suggesting new terms to be introduced. There are certain limitations in this algorithm as well. Some co-occurrences are not necessarily terms. Of the 100 terms extracted, 46 are considered as legitimate terms (cf. Figure 1). For more general terminologies, I compile 246 terms from English for Structural Engineers (V1.0), an achievement of East China Architectural Design Institute (Y0707S3317) [13]. The 292 extracted keywords and terms are imported in Multiterm and serve as supplementary information while I translate.
Figure 1. Terms extracted Using Compiled Corpus in Sketch Up.

I create a new TM with nothing filling in. I attempt to witness the changes as I translate. Starting from “Section 1607” (59th sentence), matches from TM and context matches become more frequent and the results are more usable and reliable (cf. Figure 2). Of all the segments, 141 (2352 words) are classified as context match, taking up 90.29% of the source text (cf. Table 1).

Table 1. Analysis Report for this Pilot Study.

| Type                  | Segments | Words | Characters | Percent  | Recognised Tokens | Tags |
|-----------------------|----------|-------|------------|----------|-------------------|------|
| Perfect Match         | 0        | 0     | 0          | 0.00%    | 0                 | 0    |
| Context Match         | 141      | 2352  | 12495      | 90.29%   | 453               | 190  |
| Repetitions           | 0        | 0     | 0          | 0.00%    | 0                 | 0    |
| Cross-file Repetitions| 0        | 0     | 0          | 0.00%    | 0                 | 0    |
| 100%                  | 8        | 112   | 614        | 4.30%    | 15                | 4    |
| 95-99%                | 3        | 112   | 643        | 4.30%    | 36                | 28   |
| 85-94%                | 0        | 0     | 0          | 0.00%    | 0                 | 0    |
| 75-84%                | 0        | 0     | 0          | 0.00%    | 0                 | 0    |
| 50-74%                | 0        | 0     | 0          | 0.00%    | 0                 | 0    |
| New                   | 1        | 29    | 151        | 1.11%    | 0                 | 0    |
| Total                 | 153      | 2605  | 13903      | 100%     | 504               | 222  |
After translating the proposed texts, I compile a corpus from the target text and interrogate it for key words and terms. I examine the bilingual documents manually to spot inconsistency between repeated segments. The analysis of the issue of inconsistency, along with evaluation of overall experience is presented in next section.

3.2. Evaluation of the Use of TM in Translation Process

The use of CAT tools in translation has longed been controversial. On the one hand, compared to Machine Translation and traditional human translation, the use of computer aid translation improves quality and productivity. The introduction of TM, in particular, frees translator from repetition work for more intellectual demanding ones. The consistency in terms improves the cohesion and coherence in the translation work; thus improves readability of the translated work. Clients benefit from the use of TM tools as well. The rates of pay for certain translation project may be altered according to percentage matches [14]. However, on the other hand, TM has its limitations. For instance, lack of proper maintenance and updated terminologies [14] would result in reduced translation quality. There are also discussions of whether blind faith in TM would result in inaccuracy of translation; whether the repetition of the same pattern reinforces the conventional usage in languages, etc. In an attempt to testify the commonly-held perception that “they [TM tools] reduce the cost of translation, save time and remove inconsistency” [15], I translate selected sections of IBC and examine this issue while evaluate my overall experience. Black Box Approach¹ is used in the evaluation process.

Terms are prepared beforehand to reduce the possibility of error propagation, which can also result from bad quality control of TM system [9, 14].

Overall, the use of TM in this project is satisfactory in that it increases productivity while maintains fair degree of quality. However, there are certain problems arising from my use of Trados.

Matches between segments with section numbers can be problematic, since these numbers are not consistent. Segment 40 shows a match rate of 80% from TM, which is from segment 32. Problem may arise if difference in section numbers is overlooked when one is concentrating on the word “alternative” (cf. Figure 3).

Inconsistent segmentation can be problematic in later retrieval of TM data. In my practice, Trados failed to suggest legacy translated segments. In segment 147, “highvelocity” is not spotted as matches or terms for the right spelling convention is “high-velocity” in TM and prepared termbase. The match rate between segment 138 and 147 should be higher if it is not for the difference between “high-velocity” and “highvelocity” (cf. Figure 4).

¹ Black Box Approach is based on testing only whether a given system achieves its objectives, i.e. the evaluators are only familiar with the input and output material, which serve as their sole basis for appraising the system’s performance. In other words, they only know how the system does what it does and they are only concerned with the results it produces and how they related to the tool’s objective.
The segmentation can result in pitfalls in the quickpaste function as well. For instance, the section number 1607.7.4.1 is divided into three paste-able units, which could be overlooked by translators.

There are certain issues regarding inconsistency as well. Moorkens [16] raised the issue of inconsistency in TM, which would contaminate the quality of translated work. Possible causes to this issue may be inadequate management and maintenance from the users’ side, as well as incorrect segmentation, erroneous match suggestion and inadequate integration of terminology of Trados.

Moorkens proposed a framework of 4 categories of Translation Unit (hereinafter referred to as “TU”) inconsistency at segment level, as follows:

| Four TU categories | Description | Example |
|--------------------|-------------|---------|
| 1                  | Inconsistent source --> Inconsistent target | “areas of special flood hazard” (segment 122), “special flood hazard areas” (segment 129) |
| 2                  | Inconsistent source --> Consistent target | “in the design” (segment 28), “in design” (segment 37) |
| 3                  | Consistent source --> Inconsistent target | “applicable load” (segment 12, 50) |
| 4                  | Consistent source --> Consistent target | “Other load” (segment 27, 29, 36, 38, 49, 51) |

For example, between segment 28 and 37, there is a match rate of 98%. Since Quickpaste is used in translating the former segment (28), the difference in numbers does not cause the match rate to be 98%, but the difference between “in design” and “in the design”. It can be categorised in Category 2.

The data analysed in this research is rather small, compared to Moorkens’ collected data, therefore, there are not many cases of inconsistency spotted in this analysis. Another reason behind this is that the material selected for translation is of high quality, contains few complex syntax such as passive voice and long sentences [16]. Reduced number of translator involved could contribute to consistency as well. As otherwise translations by different translators would be collected and merged afterwards and this requires for internally agreed translation decisions and well-maintained flows of TM to ensure consistency of terms.

TM users have to balance between the save in time and cost and the quality of translated work. Moorkens suggest accepting only exact matches and high fuzzy matches (98% or higher) in order to retain the internal consistency of TM and assure the quality of target text [16].

4. Conclusion

TM tools like Trados have shown their feasibility and applicability for translating technical texts. According to my use of Trados and my evaluation of it, the axioms that TM tools save time and cost while maximising consistency [16] are perpetuated. The inconsistency among segments are not much and would not create error propagation if TM tools is used with due care. As Bowker puts it, one has to be critical enough with the use of TM to ensure the productivity and accuracy of target text [17].

However, this practice is limited in that the translated work involved in this study is rather short and the word limits in this essay prevents me from further explore and account for the collected data in
Another version, for instance, International Building Code 2009 could be translated, using the TM I created from this project to see how TM works between different versions of the same code, where would be high repetition rate between the two and recall for extra care to use matches and autosuggestions. Further study based on Moorkens’s qualitative phase would prove helpful in the context of Chinese translation of English technical texts. Interviews would be held with professional Chinese translators from English scientific texts to testify the observations in this study and Moorkens’s statement from a non-European perspective, to prove the replicability of the axiom of TM tools. Last but not least, an interdisciplinary study in the field of human-machine coupling would prove fruitful as the application of translation technologies in translation and interpreting activities receives increasingly more attention and calls for further research. More theoretical issues could be addressed from the perspective of computer science, in order to evaluate and in turn, enhance the performance of TM, and translation technologies as a whole.

References
[1] ICC, International Building Code, International Code Council, Washington DC. (2012)
[2] R. Uwe, State of the art in translation memory technology, in G. Rehm, F. Sasaki, D. Stein, A. Witt (Eds.), Language technologies for a multilingual Europe, Translation and Multilingual Natural Language Processing 4, Language Science Press, Berlin, 2018, pp. 55-84.
[3] F. Krollmann, Linguistic Data Banks and the Technical Translation, Meta: Journal des traducteurs [Meta: Translators' Journal], 16(1-2) (1971) 117-124.
[4] J. Hutchins, The Origins of the Translator's Workstation, Machine Translation. 13(4) (1998) 287-307.
[5] J. Moorkens, Measuring Consistency in Translation Memories: A mixed-methods case study, Dublin City University, Dublin, 2012.
[6] Y. Li, P. Lou, An Investigation Report on Translation Industry, Chinese Science and Technology Translators Journal. 4 (2007) 65.
[7] D. Qian, Computer-aided Translation: A coursebook, Foreign Language Teaching and Research Press, Beijing, 2011.
[8] SDL plc, SDL Trados Studio 2014 Getting Started, Part 1: Translating [Training manual]. (2013)
[9] A. Pym, The Moving Text: Localization, translation and distribution, vol.49, John Benjamins Publishing, Amsterdam and Philadelphia, 2004.
[10] F. Gow, Metrics for Evaluating Translation Memory Software, MA thesis, University of Ottawa, Ottawa, 2003.
[11] M. Merkel, “Checking Translations for Inconsistency–A Tool for the Editor”, Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96), Montreal, 1996, October 2-5.
[12] A. Kilgarriff, “Terminology Finding, Parallel Corpora and Bilingual Word Sketches in the Sketch Engine”, Proceedings ASLIB 35th Translating and the Computer Conference, London, 2013.
[13] L. Bao et. al, English for Structural Engineers (V1.0) (Y0707S3317), East China Architectural Design Institute, Shanghai, 2010.
[14] L. Bowker, Computer-aided Translation Technology: A practical introduction, University of Ottawa Press, Ottawa, 2002.
[15] M. Olohan, Translators and Translation Technology: The dance of agency, Translation Studies. 4(3) (2011) 342-357.
[16] J. Moorkens, Consistency in Translation Memory Corpora: A mixed methods case study, Journal of Mixed Methods Research. 9(1) (2015) 31-50.
[17] L. Bowker, Productivity VS Quality? A Pilot Study on the Impact of Translation Memory Systems, Localisation Focus. 4(1) (2005) 13-20.