Introducing linguistic transformation to improve translation memory retrieval Results of a professional translators' survey for Spanish, French and Arabic

Souhila Djabri
University of Alicante
sd89@alu.ua.es

Rocío Caro Quintana
University of Wolverhampton, UK
R.Caro@wlv.ac.uk

Abstract

Translation memory systems (TMS) are the main component of computer-assisted translation (CAT) tools. They store translations allowing to save time by presenting translations on the database through matching of several types such as fuzzy matches, which are calculated by algorithms like the edit distance. However, studies have demonstrated the linguistic deficiencies of these systems and the difficulties in data retrieval or obtaining a high percentage of matching, especially after the application of syntactic and semantic transformations as the active/passive voice change, change of word order, substitution by a synonym or a personal pronoun, for instance. This paper presents the results of a pilot study where we analyze the qualitative and quantitative data of questionnaires conducted with professional translators of Spanish, French and Arabic in order to improve the effectiveness of TMS and explore all possibilities to integrate further linguistic processing from ten transformation types. The results are encouraging, and they allowed us to find out about the translation process itself; from which we propose a pre-editing processing tool to improve the matching and retrieving processes.

1 Introduction

Computer-assisted translation tools are expanding by offering translators increasingly useful solutions; they are composed of several tools such as terminology databases, the integration of machine translation engines and translation memories (TM). Translation memories are the most relevant; their function is to store previous translations so that when translating a new segment, the user can automatically retrieve from a database its equivalent in the target language, avoiding having to translate a segment already recorded in the database (Simard, 2020). Previously, the system automatically splits the source text into segments or units and searches for a similar translation by matching of several levels. The segments are usually sentences beginning with a capital letter and ending with a full stop through the segmentation by punctuation (Oliver, 2016). Among these matching, the fuzzy matches retrieve segments from the TM that are almost similar to their equivalents in the target language and allow a translation with less post-editing. The percentage of these matches is usually calculated using an algorithm based on edit distance or Levenshtein distance (Levenshtein, 1966). In addition Tezcan, Bulté, & Vanroy (2021) reported that fuzzy matching techniques use different approaches to estimate the degree of similarity between two sentences by calculating: the percentage of tokens (or characters) that appear in both segments potentially allowing for synonyms and paraphrase, the length of the longest matching sequence of tokens, or n-gram matching, the edit distance between segments, the most commonly used metric in CAT tools, automated MT evaluation metrics such as translation edit rate (TER), the amount of overlap in syntactic parse trees, or a more recently proposed method, the distance between continuous sentence representations.

However, different authors consider different percentages for the fuzzy matches: 70% and 95% (Ranashingue, Orasan, & Mitkov, 2020) or between 1% and 99% (Bowker & Corpas Pastor, 2015). Fuzzy matches do not have an exact definition since they depend on personal settings of each user and TMS. The most important aspect is to have a high percentage of matching. To this end, several studies are being carried out in order to increase the matching percentage and improve the data retrieval.

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 44–50,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_007

The rest of the article is structured as follows: Section 2 presents previous related work, Section 3 describes the methodology of this research, the survey design and data collection, Section 4 presents, analyses and discusses the qualitative and quantitative data divided in three subsections: participants profile (4.1), use of TMS (4.2), and human evaluations of semantic and syntactic transformations (4.3). Finally, the findings, the conclusions and future work will be presented in Section 5.

2 Related work

Recent research conducted by Ranasinghe et al., (2021), Ranasinghe, Orasan & Mitkov (2020) propose a new approach of new generation translation memories using deep learning techniques. The study conducted by members of the Research Group in Computational Linguistics at the University Wolverhampton showed how to improve the performance of these systems. The authors introduced sentence encoders to improve TMS matching and retrieving processes as an alternative to conventional algorithms. Other related work by Djabri (2020) presents a comparative study for Spanish, French and Arabic applying ten semantic and syntactic transformations to calculate and analyze 1500 original and transformed segments in three pair languages: Spanish-French (ES-FR), French-Spanish (FR-ES), and Arabic-Spanish (AR-ES) with two TMS: SDL Trados and MemoQ in order to identify the TMS deficiencies and propose solutions to improve the systems. The analysis of empirical data indicates that the matching scores decrease between language pairs of the same language family such as ES-FR/FR-ES because of the transformations, but it also shows that Arabic as a source language faces other difficulties when it comes to translating into Spanish, where several segments have no matching. Arabic also registered lower matching percentage with MemoQ, which seems to have more difficulties with the word order transformation than SDL Trados despite that Modern Standard Arabic has a rich and flexible morphology in terms of word order (Bassam & al, 2017) and there are several possibilities of syntactic typology: Verb, Subject, Object (VSO), Subject, Verb, Object (SVO) and Verb, Object, Subject (VOS). For Spanish, the first results show that MemoQ has considerable difficulties when it comes to transform the active/passive voice and the substitution of a word by a synonym. SDL Trados faces difficulties with the substitution of two words by their synonyms and the change of the word order, sentence and/or clause order. As for French, these systems record lower matching when transforming the active/passive voice and replacing a word by a personal pronoun despite the nature of the French language with a frequent use of the passive voice unlike the Spanish language where the use of the passive voice is limited (Weber, 2014). These results indicate that TMS need to integrate more linguistic processing to improve the data retrieval.

Similarly, others researches (Silvestre Baquero & Mitkov, 2017) demonstrated the importance of integrating more language processing in TMS after the calculation and analysis of fuzzy matches. The authors suggested lexical, semantical and syntactical transformations for English-Spanish/Spanish-English and it was observed that Spanish as a target language has more lexical and syntactical difficulties due to the syntactic complexity of Spanish. Consequently, the research demonstrated the shortcomings and the linguistics limitations of TMS.

However, improving TMS retrieving processes still needs human assessment by professional translators, not only from a linguistic and computational point of view. This paper, which presents the second phase of our research on different possibilities to improve the TMS matching process, discuss the results of questionnaires conducted with native speakers and professional translators of Spanish, French and Arabic. The objective is to continue to evaluate the empirical data obtained on the first phase and analyze all the qualitative and quantitative data in order to draw conclusions about different aspects, in particular on how to apply these linguistic transformations to improve the TMS efficiency.

3 Survey Design and Data Collection

The objective of this research is to explore the possibilities to improve the TMS matching process through the evaluation of ten linguistic transformations for Spanish, French and Arabic: 1) Change active to passive voice, 2) Change passive to active voice, 3) Change the word, sentence or clause order, 4) Replace one word with its synonym, 5) Replace two words with their synonyms, 6) Replace two words with their synonyms and change the word, phrase or clause order, 7) Replace one word into a personal pronoun, 8) Replace one word into a personal pronoun and change the word, sentence or clause order, 9) Change active to passive voice and replace one word with its synonym, and 10) Change active to passive voice and replace one word into a personal pronoun. For the survey[1], we selected ten original segments with their corresponding transformation as mentioned previously for each language pair: ES-FR, FR-ES, and AR-ES. We collected and built multilingual corpora from the United Nations General Assembly, Internal Regulation and share it with the participants in Excel files.[2]

The questionnaire was designed with Google Forms in three languages (Spanish, French and Arabic) according to the participants' profile. The choice of these three languages is based on two main reasons i) addressing the participants with their native language and/or working language allows us to have clear assessments and avoid communication ambiguities and ii) the three languages are the language pairs analyzed in the first phase of our research. Thus, it is a logical continuation.

Regarding the data collection, the participants were contacted in early January 2021 and the questionnaires were shared online by emailing the participants and explaining the stages of the research with a guideline[3] in three languages so every participant had to choose their language. Once accepted, the participants completed the questionnaires between 17 and 27 January 2021.

Furthermore, all participants are graduated, post graduated or professional translators in different fields, including two participants who are PhD students in translation. The participants come from i) Algeria and Egypt for the Arabic language, ii) Spain for the Spanish language, and finally iii) from France for French language (see 4.1).

4 Analysis, discussion of quantitative and qualitative data

In this section, we present and analyze the results of the questionnaires. First, we present the collected data with tables, graphs and/or descriptive statistics depending on the nature of the questions: open and closed ones (Saldanha & O'Brien, 2014); then, we analyze the participants answers, question by question according to the questionnaire order. We will present and discuss the human assessment covering three subsections i) participants profile, ii) use of TMS, and finally iii) human evaluation of the linguistic transformation.

4.1 Participants' profile

With regard to the participants' profile (15, 5 for each language), the first question related to their experience in translation obtained fifteen affirmative responses, i.e. 100% of participants are professional translators for Spanish, French or Arabic which is a significant advantage since all the participants are part of the studied discipline and work in the field.

The second question is presented to define the years of professional experience each participant (P) has. For Spanish participants (ES), the number of years of experience are between three and ten years. With regard to French participants (FR), the translators have experience from three to eight years. Finally, Arabic participants (AR) have a professional experience between three and ten years (see Figure 1).

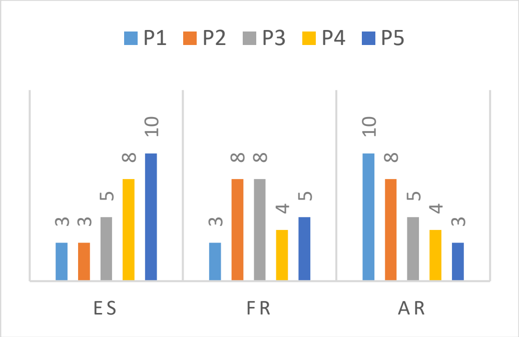

Figure 1: Years of professional experience of the participants.

The third question defines the field or specialty of participants. This question supports the results that indicates that most translators are specialized in technical translation and TMS are generally intended to translate technical and repetitive texts

[1] The full survey is available in the following links for Spanish, French, Arabic with an English translation.

[2] The original and transformed segments for each language pair are available at the following link.

[3] The questionnaires guidelines are available at the following link.

(Timonera & Mitkov, 2015) or texts with specific typology, even if the TMS are used for all types of texts: general, administrative, technical and specialized (Leblanc, 2016). Therefore, determining the specialty is of importance. The results indicate that four Spanish translators are specialized in legal and administrative translation. Other specialties are related to economy, construction, transport, and translation of medical texts. For French, it was found that three of the five translators are specialized in legal and administrative translation; the fourth participant indicated other specialties: translation in the field of Information Technology, Human Rights and Marketing; the fifth participant is specialized in translation related to the field of transport. As for the Arabic translators, three participants are specialized in legal translation; the other two responses indicate medical and economical translation respectively.

4.2 Use of TMS

As for the TMS use (question 4), 60% of Spanish, French and Arabic translators report using TMS. In other words, nine out of fifteen translators use TMS.

On the other hand (question 5), when participants are asked about which TMS they usually use (MemoQ, SDL or another TMS), 20% of Spanish translators use MemoQ and 80% use another TMS. However, no translators from French use MemoQ or SDL, the French translators reported using another TMS apart from these two. Finally, only one translator for Arabic language (20%) uses SDL and 80% use a different TMS.

4.3 Human Evaluation and Linguistic Transformation

The third subsection of the questionnaire corresponds to the analysis of the human evaluation of the ten transformations applied in order to analyze the possibilities of improving the matching process for the three pair languages and to study how to integrate them by analyzing different related aspects from a translation point of view. Five questions are dedicated to this group. To do this, participants were invited to read the ten examples corresponding to their language presented in Excel spreadsheets (see footnote 2). Each Excel spreadsheet consists of two columns: the first called "original segments" which includes the segments without any transformation and the second column called "transformed segments" where each segment is transformed by applying the ten semantic or syntactic modifications.

The spreadsheet is prepared for each language to obtain an unambiguous assessment where each participant chooses one of the three Excel files according to their language.

In the next step, transformations are evaluated by comparing them with the original segments The participants were asked to give their assessment on an increasing linear scale (from 1 to 5 or from ambiguous to clear transformations). Spanish translators give evaluations that are between three and five (see Figure 2): 40% have an average evaluation (moderately clear transformations), 20% indicate that the transformations are almost clear, and 40% indicate that the transformations are clear by granting a five.

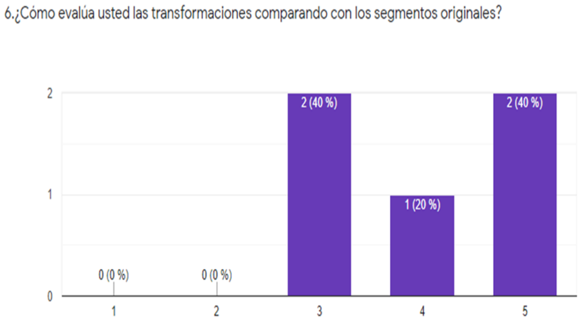

Figure 2: Linear scale for Spanish transformations.

As for the French language, the evaluations are between two and five (see Figure 3): 2 participants gave a 5 (40%), however, the other participants selected 2, 3 and 4 each one (20% for each category) for the semantical and syntactical transformations; they considered them almost ambiguous, moderately clear and almost clear.

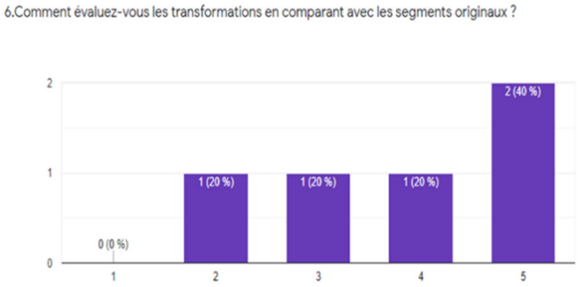

Figure 3: Linear scale for French transformations.

For Arabic (see Figure 4), the assessments range from two to five: 20% say the transformations are almost ambiguous, 40% think they are almost clear, and 40% think they are clear.

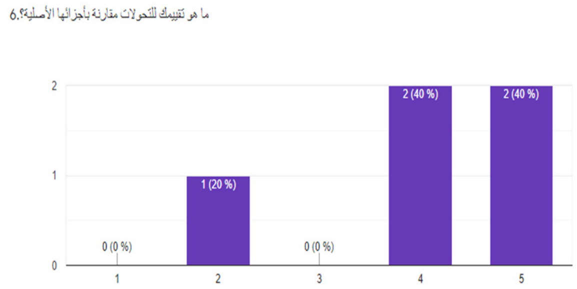

Figure 4: Linear scale for Arabic transformations.

After analyzing the three linear scales, it is observed that all the participants do not consider any of the transformations ambiguous since they have rated the transformations from 2 to 5.

Following that, the seventh question aims at whether translators will integrate these ten transformations into their future translations, the results indicate that all the answers for Spanish and Arabic are affirmative, that is, 100% of the participants will integrate these transformations. As for French, 80% of the participants said they wanted to integrate these transformations into their future translations and 20% did not want to do so. These evaluations confirm the possibilities of integrating more language processing into the TMS given that not only translators have given evaluations that are mostly favorable but also wish to add them to their work for Spanish, French and Arabic.

The participants then indicate what type of transformation they could apply in their translation in order to define the linguistic changes to be proposed to improve the matching process. We present below the transformation with the highest score for each language (see Table 1). For Spanish, most are related to transformation four, seven and nine, i.e. 80% to i) replace a word with it synonym, ii) replace a word with a personal pronoun, iii) change active to passive voice and replace a word with its synonym.

With regard to French, transformations four and five are the ones that obtain the highest percentage with 80%; participants indicate that they are able to replace a word or two words with their respective synonyms without changing the word order. In addition, participants could apply transformations one, two, three and seven, i.e. change active to passive voice, change passive to active voice, change the word, sentences and/or clauses order as well as replace a word with a personal pronoun. For Arabic, transformation four gets a total of 100%, Arabic translators prefer the substitution of a word by its synonym.

Spanish	French	Arabic
Replace a word with its synonym.	Replace a word with its synonym.	Replace a word with its synonym.
Replace a word with a personal pronoun.	Replace two words with their respective synonyms.	
Change active to passive voice and replace a word with its synonym.	Change active to passive voice.	
	Change passive to active voice	
	Change the words, sentences and/or clauses order	
	Replace a word with a personal pronoun.	

Table 1: Linguistic transformations translators would apply.

The second to last question aims to add other language transformations where translators are asked which transformations they wish to add in addition to our ten transformations (see Table 2). There are two proposals for Spanish: omission and addition. For French, it was found that the reformulation, the addition of relative pronouns and the change of masculine/feminine form. Finally, Arabic translators also add the reformulation of the original segments to the ten transformations already proposed.

Spanish	French	Arabic
Omission	Reformulation	Reformulation
Addition	Change of masculine/feminine form	
	Relative pronouns	

Table 2: Translators' suggestions.

Finally, we asked about the translation process itself. For this purpose, we have defined three categories of this process: pre-editing process (Pre-E), editing during the assisted translation (AT) and post-editing process (Post-E). The Pre-E process is the preparation prior to the computer-assisted translation where translators modify the source text by correcting possible errors, changing the word order, remove the use of passive voice or set an appropriate terminology (Arenas, 2019). For us, this process is elaborated before uploading the file to the TMS while the second process (AT) is performed during the translation after uploading the source text into the TMS since we consider that in this phase the translators could set the segmentation of the imported document (for instance, split or combine segments or adjust the alignment of the parallel documents). As for the third process (Post-E), it refers to the revision and correction of the translation provided by the TMS.

Therefore, the surveyed translators would include these transformations during the translation with 60% of the answers for Spanish, French and Arabic (see Figure 5), that is to say that the participants indicate to integrate these transformations after uploading the document in the system while 20% include these same transformations before the AT process or during the Pre-E. On the other hand, 20% include transformations at the end of the AT process (Post-E process). These results clearly demonstrate that translators not only employ linguistic transformations such as synonym substitution, change active-passive voice etc., but also that the pre and post-editing process is not as significant compared to the process of adding transformations during translation and after uploading the document into the system.

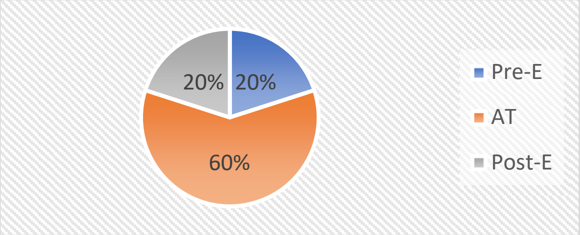

Figure 5: The translation process.

5 Conclusions and Future Research

In this paper, we have presented the results of a questionnaire survey conducted among professional translators specializing in different fields for the Spanish, French and Arabic languages. The objective of this survey, which is the second phase of our research, is to find out how to improve the matching and data retrieval process by applying semantical and syntactical transformations as well as to study the different ways of integrating these transformations and other related proposals. The findings related to the participants profile as well as the use of TMS indicate that 60% of participants use TMS to translate mainly legal and/or administrative texts with 67% of the calculated specialties.

The analysis of the quantitative and qualitative data related to the third part of the questionnaire, i.e. the semantical and syntactical transformations, showed that translators from Spanish apply transformations by synonymy, by personal pronoun, change active/passive voice and prefer transformations with one or two modifications without having to change the words order, sentences and/or clauses order. As for the translators of French, they also apply transformation by synonymy, change active/passive, passive/active voice and substitution by a personal pronoun. All translators from Arabic indicate that they prefer substitution with a synonym without changing the word order; they only make one change in the original segment.

Similarly, our ten initial transformations receive a favorable evaluation from professional translators, which supports our approach to the possibility of integrating them through linguistic processing adapted to each language. All these transformations and other proposals can be integrated during the translation process with 60% of the answers for Spanish, French and Arabic. Thus, it is important to reflect that in the translation process itself the participants prefer editing at the same time of translating, instead of the process of pre-editing the source text before it is uploaded into the system and post-editing which only represent 20% of the participants.

In addition, the participants propose to integrate other types of transformations that we consider to be techniques or strategies used by translators to improve their work, such as reformulation, addition, omission, as well as the change of grammatical gender, or relative pronoun. These proposals will be studied exhaustively in order to consider how we can integrate translation strategies or techniques to improve the efficiency of the TMS providing ideally an editing tool or an automatic paraphrasing process integrated in the TMS.

Finally, these results will be further developed and improved as we plan to continue this research and propose new approaches, especially regarding the translation process itself and how to use translation techniques and strategies in that process.

Acknowledgements

We would like to thank Professor Doctor Ruslan Mitkov (University of Wolverhampton, United Kingdom) and Professor Doctor Gloria Corpas Pastor (University of Malaga, Spain) for their valuable feedback.
Also, we are grateful for the support of the participants and the time they devoted to help our research.

References

Arenas, A. G. (2019). Pre-editing and post-editing. In *The Bloomsbury Companion to Language Industry Studies*. https://doi.org/10.5040/9781350024960.0019

Bassam, H., Asm, M., Nadim, O., & Abeer, T. (2017). *Formal Description of Arabic Syntactic Structure in the Framework of the Government and Binding Theory*. *18*(3), 611–625. https://doi.org/10.13053/C

Bowker, L., & Corpas Pastor, G. (2015). Translation Technology. *Oxford Handbooks Online*, (November), 1–26. https://doi.org/10.1093/oxfordhb/9780199573691.013.007

Djabri, S. (2020). Análisis contrastivo de las coincidencias parciales en español, francés y árabe mediante transformaciones lingüísticas. *Skopos*, *11*, 103–120.

Leblanc, M. (2016). *La traduction spécialisée à l ' ère des nouvelles technologies : quel effet sur le texte de spécialité ? The impact of new translation technologies on specialized texts*. *1*, 77–92. https://doi.org/10.14746/strop.2016.425.006

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Oliver, A. (2016). *La traducción asistida por ordenador*. Barcelona: Oberta UOC Publishing, SL.

Saldanha, G., & O'Brien, S. (2014). Research Methodologies in Translation Studies. In *Research Methodologies in Translation Studies*. https://doi.org/10.4324/9781315760100

Silvestre Baquero, A., & Mitkov, R. (2017). *Translation Memory Systems Have a Long Way to Go*. 44–51. https://doi.org/10.26615/978-954-452-042-7_006

Simard, M. (2020). Building and using parallel text for translation. In M. O'hagan (Ed.), *The Routledge Handbook of Translation and Technology* (2020th ed., pp. 78–90). https://doi.org/10.4324/9781315311258

Tezcan, A., Bulté, B., & Vanroy, B. (2021). Towards a better integration of fuzzy matches in neural machine translation through data augmentation. *Informatics*, *8*(1), 2–27. https://doi.org/10.3390/informatics8010007

Ranasinghe, T., Orasan, C., & Mitkov, R. (2020). Intelligent Translation Memory Matching and Retrieval with Sentence Encoders. arXiv preprint arXiv:2004.12894.

Ranashinge, T., Mitkov, R., Orasan, C., and Caro Quintana, R., (2021) "Semantic Textual Similarity based on Deep Learning: Can it improve matching and retrieval for Translation Memory tools?" Parallel Corpora: Creation and Applications. John Benjamins.

Timonera, K., & Mitkov, R. (2015). Improving Translation Memory Matching through Clause Splitting. *Proceedings of the Workshop on Natural Language Processing for Translation Memories (NLP4TM)*, 17–23.

Weber, E. (2014). La traducción de la voz pasiva francesa al español: ¿cuestión de lengua o cuestión de traducción? *Mutatis Mutandis*, *7*(2), 368–385.