Knowledge of Provenance and its Effects on Translation Performance in an Integrated TM/MT Environment

Carlos Teixeira

Intercultural Studies Group, Universitat Rovira i Virgili,
Avda. Catalunya 35 – 43002 Tarragona, Spain
carlostx@linguanativa.com.br

Abstract. The integration of machine translation (MT) and translation-memory (TM) systems in professional translation settings has turned pre-translation + post-editing into an attractive alternative in terms of productivity for all parties involved in the translation process. In some cases, source files are pre-translated using a combination of customised MT and TM before reaching the translators, who then become reviewers, or post-editors. But how does this actually affect productivity and how do translators feel when performing this new activity? In order to look for answers to those questions, I ran a pilot experiment comparing two different environments. The basic difference between the two is the availability of information on the provenance of the suggested translation for a particular segment (whether it comes from MT, TM, and at which match percentage). Data were collected using screen recording, keystroke logging and post-performance interviews.

Keywords: translation technology, translation memory, machine translation, process research, speed, productivity, performance, provenance, trust.

1 Introduction

Until recently, machine translation (MT) and translation memories (TM) were seen as totally different approaches to using technology in translation. While the first approach was largely restricted to end users interested in grasping the general idea of a text written in a language they could not understand (usually while browsing the Internet), the second was addressed to professionals in the translation industry, such as translators, translation agencies or translation departments in large companies.

However, this scenario has been changing at a rapid pace in the last few years, mainly due to quality improvements and the general availability of statistical machine-translation systems, based on large amounts of human-produced bilingual text. This has allowed MT to be progressively integrated into TM tools in professional translation environments, bringing new possibilities as well as new challenges.

The potential productivity gains derived from this integration of machine translation (MT) and translation memories (TM) are calling for new work methods in the translation market. As an example, some translation agencies pre-translate their source files using a combination of TM and customised MT before sending them out to translators, who then become reviewers, or post-editors. In this scenario, translators review each segment without knowing its provenance, i.e. whether it came from a translation memory (and at which match percentage) or from a machine-translation engine. Could this missing information have an impact on the way translators perform their tasks, compared to a more traditional environment, where translators would know where each translation suggestion comes from? In other words, how does the 'knowledge of provenance' of translation suggestions affect translators' behaviour in environments that integrate TM and MT?

2 Previous Research in the Field

None of the published studies on translation technology that we are aware of seems to take into account this specific aspect that distinguishes translation memory systems from machine translation systems: TM systems show translators the 'provenance' and the 'quality' of the translation suggestions coming from the memory, whereas MT systems display the 'best translation suggestion possible' without any indication of its origin or degree of confidence. It is our assumption that this missing distinction might be one of the reasons for discrepant results in some studies that compare translation speed when (post-)editing MT and TM suggestions.

As an example, [2] compares the performance of TM vs. MT when translators work in a 'traditional' TM system, i.e. when they know the provenance of the translation suggestions they are working with. One of her findings is that "cognitive load [and processing speed] for machine translation matches is close to fuzzy matches of between 80-90% value" (p.185). For fuzzy matches above 90%, including exact matches, TM processing is faster and requires a lower cognitive load, whereas the opposite happens for fuzzy matches below 80%.

In a different study, [1] reproduces an actual scenario that can be found in industry. The author analyses time and quality when editing translation-memory suggestions vs. machine-translation suggestions, in an environment where translators could not tell the provenance of each suggestion. Under this particular condition, her "findings suggest that translators have higher productivity and quality when using machine-translated output than when processing fuzzy matches [at any percentage level] from translation memories" (p.11).

3 Research Question and Hypotheses

Inspired by these two studies and their rather contradictory findings (at least for high-percentage fuzzy matches), I set out to investigate whether the fact of knowing the provenance of the segments could provide and explanation for this discrepancy. What are the differences (if any) in the translation process between a situation where translators know the provenance of the translation suggestions they are editing and a situation where this information is not available?

In order to answer this question, I compared two translation environments. In the first environment, translators do not know the provenance of translation suggestions, whereas in the second environment translators do have access to this information. These are my working hypotheses:

Hypothesis 1 (H1): The *translation speed* is higher when *provenance information* is available.

Hypothesis 2 (H2): There is no significant difference in the *quality* level when *provenance information* is available.

Some definitions are necessary in order to operationalize the variables we want to test:

- *Translation speed* is measured as words per hour. There are separate counts for the first rendition (drafting) and second rendition (self-revising).
- *Provenance information* of translation suggestions is indicated by showing their origin (TM or MT) and, in the case of TM, by displaying its fuzzy-match percentage and highlighting the differences between the actual segment and the matching segment in the TM, as is usually the case in most TM systems.
- *Quality* is measured as a score given by two reviewers, who process all resulting translations according to predefined criteria (see section 4.3).

4 Pilot Experiment

In order to test those hypotheses and fine-tune the methodology for my future doctoral research, I ran a pilot experiment with two translators from English to Spanish. Each of them translated two similar source texts of around 500 words each, in the two different environments described below.

Environment B presents the source-text segments on the left-hand side of the screen and a pre-translated version of the source text (obtained through the pre-processing of the file with TM and MT) on the right-hand side. In this case, all no-matches were replaced with MT suggestions, and the whole text was presented as a sequence of pre-translated segments. Translators could edit the pre-translated segments as if they were revising a translated file and they had no information on the origin of each of the pre-translated segments (i.e. whether they came from a TM segment or an MT segment). For mnemonic recall, let us call this environment B (as in ‘blind’). This environment tries to reproduce as close as possible the environment described in [1].[1]

Environment V is similar to the previous one in that translators also had access to the source-text segments on the left-hand side of the screen and an editing space on the right-hand side. However, its difference consists in that, instead of working by ‘blindly’ editing pre-translated segments, translators could see where the default

[1] Our environment B presents all the pre-translated text at once, while the environment used in [1] displays each pre-translated segment at a time and does not allow for a revising phase. In order to make both environments closer (and environment B closer to environment V), we are planning to change the settings of our Trados project in the future to have it display each segment at a time. There are no plans to restrict the revising phase in our study.

translation suggestion was coming from (either from the translation memory or from the MT engine). Additionally, in the case of TM suggestions, translators could see the highlighted differences between their actual source segment and the TM source segment. For mnemonic recall, let us call this environment V (as in ‘visual’). This environment tries to reproduce as close as possible the environment described in [2].

4.1 Subjects

Both subjects are male and native speakers of Spanish. Subject1 has formal training in translation and 4 years of professional experience in several fields, especially audio-visual translation. Subject2 also has formal training in translation and around 8 years of professional experience in various fields, mainly in localisation and technical translation. Both are familiar with many different translation memory systems.

For my main experiment, I plan to have ten English-to-Spanish translators who are native speakers of Iberian Spanish: five male and five female subjects. They will be selected by means of a questionnaire and will have 5+ years of professional experience working with translation-memory systems on technical or marketing texts. Formal training in translation will not be a prerequisite.

All ten translators will translate both texts in both environments. Five translators will start working in B and the other five in V, in order to account for potential differences related to the order of the tasks

4.2 Materials

Our source texts were taken from an article in a technical magazine and deal with composite materials in car manufacturing. The main reason for choosing this kind of material was a wish to use text outside of the ‘software localisation’ domain – which is the object of most research studies in the field – still with (marketing) stylistic features that make it more demanding for translation. The specific article was chosen mainly because of its topic (technical, while still somehow interesting for translators) and length – allowing for the extraction of two excerpts of around 500 words.

The main article had a total of 1310 words, corresponding to 55 source segments, or 23.8 words per segment in average. In order to have two source texts of around 500 words, I used 21 segments for each of them. As a result, SourceText1 has 512 words, and SourceText2 has 510 words.

A translation memory was created by aligning the English source text with the Spanish target text (final version revised by a copy-editor and approved by the client) using SDL Trados WinAlign + manual verification of each segment. A decision was made to use the following fuzzy match distribution in the experiment:

- 7 ‘no matches’ (replaced by MT feeds);
- 5 exact matches;
- 9 fuzzy matches, of which
    - 3 matches within the 70%-79% range,
    - 3 matches within the 80%-89% range, and
    - 3 matches within the 90%-99% range.

The order of presentation of match types during translation was defined by a random number generator and it was different for each of the environments. Then I edited the aligned memory to obtain two memories with the characteristics above, one for each environment. Segments set to have an 'exact match' suggestion were left untouched. Segments corresponding to a 'no match' were replaced through SDL Trados Studio with translation suggestions provided by the public, freely available Google Translate machine-translation engine. Finally, for creating the fuzzy matches I resorted to the following strategies: delete parts of the source and target segments, include or replace some words in the source and target segments, or edit the source text.

4.3 Data Collection

The two translation environments were created within SDL Trados Studio 2009 Freelance. The main methods for collecting data were screen recording and keystroke logging through BB Flashback Express 2. Retrospective interviews were also used to try to obtain some insight of translators' feelings and satisfaction in both tasks. Think-aloud protocols were not used as they are known to slow down the translation process and we were precisely trying to measure translation speed in a natural(istic) environment. For testing quality, all texts were rated by two reviewers.

Time was measured by watching each of the translators' performances in BB FlashBack Player and manually noting down the start and end times for each individual task. Time was counted when translators were typing, thinking, hesitating, or looking at the source text (except when they read the full source text before starting the translation, as we cannot make a correspondence between the time and specific segments). Time was not counted when translators switched to another window to look up terminology, tried to find a specific function in the tool, or spoke with the researcher. The time counter was paused when the subject started moving the mouse to go to another application (usually a web browser) outside of the translation environment. It was also paused when the subject moved to the source segment to copy text to be pasted in the browser. Time count was resumed when the subject returned to the translation environment. Time spent on searches within the translation environment (mainly with the Concordance function) was considered as translation time.

For assessing quality, all texts were rated by two reviewers, based on an error-count system. The quality level of a translation was measured through a score, which starts at 10 and decreases according to the grid shown in Table 1.

4.4 Running the Experiment

Both subjects chose to use their own laptop computers during the experiment. Before they started, we made sure they had the required version of SDL Trados and BB FlashBack installed and configured. The aim was to have translators work in an environment as close as possible to their natural work environment, meaning that they could keep their preferred configuration in terms of keyboard, screen and mouse (either built-in or external), operating system (within the Microsoft Windows family), browser favourites, dictionaries, etc. They also had access to the Internet during the experiment.

Table 1. Quality assessment grid

Type of error	Deduction
1 misspelled word	.25
1 grammar mistake (morphology, syntax)	.25
1 use of incorrect or inconsistent terminology	.25
1 general readability (understanding) issue	.25
1 sentence structuring issue (style, register)	.25
1 instance of omitted information	.25
1 instance of incorrect or inaccurate meaning rendition	.25
1 localisation error (numerical formats, units)	.25
Other deductions	.25 each

At the beginning of the experiment, a digital voice recorder was turned on. The initial tasks subjects had to perform were: (a) copy a short passage in Spanish, and (b) translate a short passage from English to Spanish. In both tasks, the source texts were printed out and translators had to type their target texts in Microsoft Word. The purpose of these two initial tasks was twofold: to measure their baseline typing speed (and eventually assess whether this has an influence on their editing strategies) and to serve as a warm-up (and stress-down) activity. This came from a suspicion that the typing ability of each individual translator might have an influence on their performance with each kind of translation suggestion.

Next, the translators were given instructions in Spanish on how to perform the main tasks for the experiment.[2] In general terms, the instructions told the subjects that the memory they would be provided was created based on a client-approved final version of the Spanish magazine, that it contained five different kinds of matches, and that machine translation was used to replace ‘no match’ segments. The translation ‘brief’ mentioned the translators would be paid the same amount per word (no fuzzy-match discounts), so implying that they were supposed to revise all segments, including exact matches. The instructions also made it clear that their translations were going to be assessed and graded for quality by a professional reviewer, thus also implying that the translators should try to achieve maximum quality in both environments. A time limit of 1.5 hours was set for each of the texts.

During the translation of the texts in both environments, BB FlashBack was set to record screen activity; keystrokes; mouse position, movements and clicks; translators’ faces; and sound (voices, keyboard, etc).

[2] Subject1 received the instructions orally, but then the researcher decided to give similar instructions in written format to Subject2, in order to eliminate potential variations due to his oral performance. In the main experiment, all subjects should receive the same instructions in written format.

5 Preliminary Results

5.1 Subject1

Tables 2 and 3 show the average speed results for Subject1.

Table 2. Average translation speed per type of segment in environment V for Subject1.

	SOURCE WORDS	TIME (sec) 1st rendition	SPEED (words/hr) 1st rend.	TIME (sec) 2nd rendition	SPEED (words/hr) Combined
EXACT (100%) MATCHES	131	155	3036	94	1895
90-99% MATCHES	91	234	1397	101	977
80-89% MATCHES	51	153	1197	27	1019
70-79% MATCHES	87	401	781	88	577
NO MATCHES (MT FEEDS)	150	783	690	132	591
	510	**1727**	**1063**	**441**	**847**

Table 3. Average translation speed per type of segment in environment B for Subject1.

	SOURCE WORDS	TIME (sec) 1st rendition	SPEED (words/hr) 1st rend.	TIME (sec) 2nd rendition	SPEED (words/hr) Combined
EXACT (100%) MATCHES	128	566	815	49	749
90-99% MATCHES	65	369	635	53	555
80-89% MATCHES	77	210	1321	30	1155
70-79% MATCHES	77	273	1014	20	946
NO MATCHES (MT FEEDS)	165	592	1004	129	825
	512	**2009**	**917**	**281**	**805**

If we look at the average results for the first rendition (drafting), we see that translation speed is higher in V (1063 words/hr) than in B (917 words/hr), a difference of 15.9 percent. If we look at the results for the first and second renditions (drafting + self-revising) combined, translation speed is still slightly higher in V (847 words/hr) than in B (805 words/hr), but the difference is reduced to 5.2 percent. Due to the dispersion of the data and the reduced number of segments in the texts, the detected difference in the overall speed between V and B is not statistically significant.

However, if we look at the different fuzzy-match levels, differences in speed are more pronounced. In environment V, it is possible to identify three groups of speed levels: exact matches are translated the fastest, fuzzy matches between 80-99% are translated at around half that speed, and lower fuzzy matches (below 80%) and MT output are translated the slowest. This is in accordance with intuitive expectation and with the results obtained by [2].

In environment B, there is a dramatic reduction in speed (from 1895 to 749 words/hr) for exact matches, suggesting that provenance information has a high impact on this kind of translation suggestion. Matches in the 90-99% range also show a dramatic reduction in speed (from 977 to 555 words/hr), again indicating that provenance information has a significant impact in this case. Matches in the 80-89% range did not show a significant variation. For lower fuzzy matches and MT feeds, it is worth noting that there was an *increase* in speed.

5.2 Subject2

Tables 4 and 5 show the average speed results for Subject2.

Table 4. Average translation speed per type of segment in environment V for Subject2.

	SOURCE WORDS	**TIME (sec) 1st rendition**	**SPEED (words/hr) 1st rend.**	**TIME (sec) 2nd rendition**	**SPEED (words/hr) Combined**
EXACT (100%) MATCHES	131	236	2000	121	1323
90-99% MATCHES	91	354	925	160	637
80-89% MATCHES	51	225	814	77	606
70-79% MATCHES	87	456	687	135	530
NO MATCHES (MT FEEDS)	150	475	1138	214	784
	510	1746	1052	708	748

Table 5. Average translation speed per type of segment in environment B for Subject2.

	SOURCE WORDS	TIME (sec) 1st rendition	SPEED (words/hr) 1st rend.	TIME (sec) 2nd rendition	SPEED (words/hr) Combined
EXACT (100%) MATCHES	128	445	1035	95	854
90-99% MATCHES	65	275	852	52	716
80-89% MATCHES	77	226	1229	143	752
70-79% MATCHES	77	289	961	79	755
NO MATCHES (MT FEEDS)	165	568	1045	161	814
	512	1802	1023	530	790

For this translator, the average results for the first rendition (drafting) show that translation speed is also higher in V (1052 words/hr) than in B (1023 words/hr), but the difference is much smaller than for Subject1, at only 2.8 percent. The combined results for the first and second renditions (drafting + self-revising) show that translation speed is now higher in B (790 words/hr) than in V (748 words/hr), with a difference of 5.6 percent. As was the case with the data for Subject1, this difference is not statistically significant.

Now let us look again at the speed differences according to the various fuzzy-match levels. Roughly speaking, the data for environment V indicate that Subject2 processed translation suggestions coming from exact matches two times faster than suggestions coming from fuzzy matches (1323 vs. 591 words/hr in average), and he translated suggestions coming from machine translation around 33 percent faster than the average speed for fuzzy matches. The faster speed for exact matches is still in accordance with our expectations, but the reasons for machine-translation suggestions being translated faster than high-percentage fuzzy matches should be investigated further.

In environment B, similarly to what happened with Subject1, the data for Subject2 indicate a dramatic reduction in the average translation speed (from 1323 to 854 words/hr) for suggestions coming from TM exact matches. All other kinds of translation suggestions had an increase in speed, with fuzzy matches in the 80-89% range showing the largest increase (42.5 percent). It is interesting to note that differences in translation speeds tend to disappear in the blind environment: exact matches were translated slightly faster, at 854 words/hr, followed by machine-translation suggestions, at 814 words/hr, with translation-memory fuzzy matches being translated a little more slowly, between 716 and 755 words/hr. If the statistical errors are taken into account, differences between the five types of translation suggestions are actually not significant.

5.3 Quality

Two revisers assessed the quality of the four translations (two per subject) using the grid provided in section 4.3. Revisers were then told to compare the two translations from the same subject and decide which one was better, if any, and to give their final grade from 0 (worst) to 10 (best). This means each reviser scored the translations twice – once according to the grid, then again holistically. The results are shown in Table 6.

Table 6. Translation quality levels for both subjects.

	Subject1		*Subject2*	
	Text 1 (environment V)	Text 2 (environment B)	Text 1 (environment V)	Text 2 (environment B)
Reviser 1	8.5	7.0	8.5	9.0
Reviser 2	7.5	7.0	8.0	8.5
Average	***8.0***	***7.0***	***8.25***	***8.75***

According to the two evaluators, Subject1 performed better in environment V, while Subject2 performed slightly better in environment B. From the evaluators' feedback, we think that quality assessment has not been done properly and the above grades need to be revised again before we can make any definite conclusions. Furthermore, we think the rating instructions need to be made clearer and a greater number of revisers shall be used.

6 Discussion

We took pains to control most of the factors that might affect our results (type of text, length of text, source language, target language, translator's experience, translation tool, etc.) and we tried to have only our main independent variable (knowledge of provenance) act on our two dependent variables (speed and quality). However, we are aware that many potential extraneous variables (confounds) were also present and had not been properly considered.

Data from our pilot experiment do not allow us to draw a definite conclusion on our first hypothesis (on speed) if we take the whole texts as a reference. Subject1 was slightly faster (5.2 percent) in environment V, while Subject2 was slightly faster (5.6 percent) in environment B. However, we can assume that the overall speed, besides individual-specific differences, depends on the distribution of different types of translation suggestions in the texts, as both subjects were faster with certain types of suggestions. For example, if our texts contained only exact matches and machine translation feeds, our results for the entire texts would probably be different.

Although our aim was to be able to draw some conclusions from a pair of intra-subject studies (if the current pilot experiment can described in this way), the translators' personal styles (and the revisers' preferences) played a more prominent role than we had originally expected. The retrospective interviews are still being processed and a deeper analysis of this material might help shed light on some of the results. For example, Subject1 seems to have 'respected' the suggestions from the translation memory more often than Subject2, and one of the evaluators did not like the solutions present in the original memory (the version approved by the client). This fact might explain why the quality results for this translator are slightly lower.

In any case, our second hypothesis needs further verification, as the data we have on quality for each subject are not sufficient to determine even whether the translations produced in the two environments can be considered of same quality for each individual subject.

6.1 Limitations

Below is a list of known limitations of the pilot experiment and, wherever possible, some solutions to overcome them in the main experiment.

Small number of subjects. This is a common problem in translation process research, and we hope the inclusion of more subjects (10) in the future will make data statistically more relevant.

Experience increases over time. The subjects' experience (thus their speed and quality) in working in both environments (B and V) can increase over time, at least as far as post-editing MT is concerned. Therefore, the data we are gathering might be representative of performance at the beginning of a learning curve. One solution would be to train translators for some period and measure their performance after some time.

Few segments. The text chosen as source text had long segments, which obliged us to use only a few segments per type of suggestion. Since we do not want to increase the total word volume of the source texts, we will probably need to choose another article or even another text type.

Irregular segments. The shortest segment had six words, while the longest had 44, which makes them hardly comparable, as MT is known to work better with segments containing a 'single idea' and worse with long sentences. Same solution as above.

Terminology. The distribution of terms in the source text should be reconsidered for the main experiment. Even though the time used for terminology search was discounted, the time spent within the translation tool was higher when terms were more complicated. This was partly compensated for by the fact that the type of suggestion for each segment was defined randomly, but in order to eliminate extraneous variations, we will try to remove problematic terms or provide a glossary for them.

Segment identification. Sometimes it was difficult to identify which segment translators were focusing on, especially in the self-revising phase. Eye tracking is an additional data-collection method that is being considered to help solve this issue.

7 Conclusion

We set out to investigate whether provenance information about translation suggestions in translation environments that integrate TM and MT has an impact on speed and quality. We ran a pilot experiment with two subjects that translated two 500-word texts in two different environments. Through screen recording and keystroke logging, we measured the time spent for each of five different types of translation suggestions. The final translated texts were assessed for quality by human reviewers. Retrospective interviews completed the data gathering methodology with an aim at obtaining general impressions from the subject translators.

Our data show that the overall speed was not significantly different in the two scenarios and the quality was of comparable level. If we look into individual types of suggestions, data on speed also show that translators spent much longer translating (post-editing) exact matches when they did not know the provenance of the suggestions.

Although inconclusive, the results of the current study indicate that ‘provenance information’ is relevant for translators working with translation suggestions from TM and MT, and that this information should be taken into account when analysing and comparing the results of different experiments.

We expect this study will help increase knowledge on translation and post-editing processes, which can be beneficial for all parties involved in the translation scene, including independent translators, translation agencies, translation-tool developers and, ultimately, translation customers, as the results can contribute to devise optimal workflows and best practices.

Besides the potential impacts on earnings (and savings), the search for optimal processes can increase the volume of text that can be processed. Even more important, it is our concern to try to optimise the translation process in ways that will help increase job satisfaction among translation professionals. Finally, I hope the results will also be of intellectual importance, as we are trying to demonstrate that the impact of technology is not just in what it does, but also in what the stakeholders know about what it does.

References

1. Guerberof, A. Productivity and quality in the post-editing of outputs from translation memories and machine translation. Localisation Focus – The International Journal of Localisation 7/1: 11--21 (2009)
2. O’Brien, S. Eye-tracking and translation memory matches. Perspectives: Studies in Translatology 14/3: 185--205 (2006)