Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories

Carlos S. C. Teixeira carlos.teixeira@urv.cat
Translation Studies Research Unit, KU Leuven, Belgium
Intercultural Studies Group, Universitat Rovira i Virgili, Spain
Avda. Catalunya 35, Tarragona, 43002, Spain

Abstract

This paper investigates the behaviour of ten professional translators when performing translation tasks with and without translation suggestions, and with and without translation metadata. The measured performances are then compared with the translators' perceptions of their performances. The variables that are taken into consideration are time, edits and errors. Keystroke logging and screen recording are used to measure time and edits, an error score system is used to identify errors and post-performance interviews are used to assess participants' perceptions. The study looks at the correlations between the translators' perceptions and their actual performances, and tries to understand the reasons behind any discrepancies. Translators are found to prefer an environment with translation suggestions and translation metadata to an environment without metadata. This preference, however, does not always correlate with an improved performance. Task familiarity seems to be the most prominent factor responsible for the positive perceptions, rather than any intrinsic characteristics in the tasks. A certain prejudice against MT is also present in some of the comments.

1. Introduction

Translating as editing of translation memory (TM) matches, on one hand, or as post-editing of machine translation (MT) suggestions, on the other, had traditionally been studied as two separate tasks. However, in recent years research interests have moved to include the language industry's trend of combining translation suggestions from machine translation and translation memories in the same text.

As one would expect, empirical studies with a focus on translation memories (Colominas, 2008; Dragsted, 2004; Garcia, 2007; Moorkens, 2012; Webb, 1998) have reported on the use of typical translation memory systems. These are tools that offer one or more translation suggestions as the user activates a segment and that always display metadata about those suggestions, i.e. they indicate where the suggested translations come from, how similar to the reference source segment the current source segment is (fuzzy match level) and where the textual differences lie. In contrast, studies on pure machine translation post-editing (Allen, 2003; Almeida, 2013; Garcia, 2011; Guerra Martínez, 2003; Krings, 2001; Plitt & Masselot, 2010) have often resorted to editing environments that offer pre-translated text with no associated metadata, as this is the typical setup for such tools. Yet the scenario for post-editing is starting to change with the development of post-editing environments that can display confidence estimates for machine translation suggestions, such as PET (Aziz, Sousa, & Specia, 2012) and CASMACAT (2014). Those estimates are believed to represent useful metadata for repairing MT suggestions.

Some studies have compared unaided translation with TM-assisted translation or with MT-assisted translation. A recent example of the latter is Green, Heer, and Manning (2013), in

which the authors take into account the translators' perceptions by means of questionnaires, like we do in the current paper. However, only a few studies have analysed scenarios in which machine translation and translation memories are combined in the same workflow. These studies either use existing TM systems (O'Brien, 2006; Skadiņš, Puriņš, Skadiņa, & Vasiljevs, 2011; Yamada, 2011) or they resort to a purpose-built post-editing environment (Guerberof, 2009; He, Ma, Roturier, Way, & van Genabith, 2010), as there seem to be no established tools for post-editing. One question that arises from this dichotomy is how to compare the performance of TM suggestions against MT suggestions in an environment that has not been conceived with their integration in mind. On the one hand, in a post-editing tool TM matches are analysed without the associated metadata, which are an important feature of translation memory systems (Anastasiou & Morado Vázquez, 2010; Karamanis, Luz, & Doherty, 2011; Morado Vázquez, 2012; Teixeira, 2014) but are not present in post-editing tools. Metadata allow translators not only to make choices among different types of suggestions, but also to decide how to approach a suggestion when repairing it. On the other hand, in a traditional TM system, MT suggestions have to be manually inserted in the active segment and are presented surrounded by much more information than is typical in a post-editing tool, maybe decreasing the translation speed for this suggestion type and increasing the post-editor's cognitive load. Therefore, comparing the performances of TM vs. MT suggestions is not an easy task, as the general tendency is to assess one of the suggestion types in an environment for which it was not originally intended to be used.

The current paper seeks to consider this issue while investigating certain aspects of TM/MT integration. It focuses on metadata and pre-translation as control variables, and analyses how they affect translators' performances and perceptions. The study reported on here uses a traditional TM system, but the system is set up using different configurations, in an attempt to "favour" one suggestion type at a time: one task reproduces an environment that is more typical of TM systems – interactive translation (Wallis, 2006) with metadata –, while the other task is more typical of MT post-editing tools – pre-translation with no metadata.

The participants' performances are measured in terms of time, edits and errors. Time and edits are measured using keystroke logging tools, while the errors are assessed by two professional reviewers using an error-score system. This measured data is triangulated with perception data obtained from interviews done with each translator immediately after the translation tasks. The goal of this triangulation is to analyse how the presence or absence of priming elements such as suggested translations and metadata affect translators and to determine whether those factors could be the main determinants for any differences found in performance.

2. Experiment description

An experiment was run with ten professional translators working from English into Spanish, who performed three different tasks within the same tool. One task presented no translation suggestions (translation from Scratch); another task presented translation suggestions from both TM and MT, and metadata about the suggestions (Visual task); and another task presented pre-translated text also from TM and MT but no metadata about the suggestions (Blind task).

2.1. Participants

The ten translators who took part in the experiment were native speakers of Spanish, with some of them being bilingual Spanish/Catalan speakers. There were five men and five women, with ages ranging from 24 to 51. They had been working for 1.5 to 18 years as full-time translators for a small translation company in Barcelona, where they had been translating IBM material

and using IBM TranslationManager[1], the translation memory system used in the experiment. They all had experience post-editing machine translated texts for IBM and/or other customers for 0.5 to 3 years. As a compensation for performing the tasks in the experiment, they were paid their regular hourly rates. Table 1 shows the demographics of the experiment participants.

Participant	Gender	Age	Years working as a translator	Years working with IBM TM/2	Years working with MT post-editing
P01	F	30	7	6	0.5
P02	M	37	14	13	0.5
P03	F	32	3.5	3	0.5
P04	M	26	2.5	2	2.0
P05	F	26	3	3	0.3
P06	M	29	2.5	2	0.5
P07	F	24	1.5	1.5	1.0
P08	M	51	18	18	0.8
P09	F	43	10	10	3.0
P10	M	47	15	14	0.5

Table 1: Demographic data about participant translators

2.2. Translation tasks

For the sake of ecological validity, the experiment was conducted with translators working with their computers of habitual use in their regular office space, and the project was configured in a way as similar as possible to their normal IBM assignments. Each translator was asked to perform the following three tasks:

a) Translation from Scratch: To translate a short text (118 words, 5 segments) from English into Spanish in IBM TranslationManager, without any help from translation memories or machine translation.

b) Translation in a Visual setting: To translate a longer text (505-542 words, 28 segments) from English into Spanish in IBM TranslationManager, with one translation suggestion per segment and metadata about the translation suggestions.

c) Translation in a Blind setting: To translate a longer text (505-542 words, 28 segments) from English into Spanish in IBM TranslationManager, with pre-translated segments but no metadata about the translation suggestions.

Task *a* (Scratch) was always the initial task, while Tasks *b* (Visual) and *c* (Blind) were performed in different orders depending on the participants, in order to have an even distribution of task orders. Task *a* was always first because it served rather as a warm-up activity and was not the focus of the study. Two different texts were used for Tasks *b* and *c* and distributed evenly between the two tasks. Table 2 shows the distribution of task and text orders among the participants.

The source texts used for the three translation tasks were excerpts from the *Troubleshooting Guide* for the IBM Tivoli Monitoring software. In the task where translators had to type from scratch, a text with 118 words and 5 segments was used, and no translation suggestions were provided. Translators were instructed to open a previously configured folder (project) in IBM TranslationManager and to translate the only file it contained.

[1] Also known as TM/2

	1st Task		2nd Task		3rd Task	
Participant	Configuration	Text	Configuration	Text	Configuration	Text
P01	Scratch	0	Blind	1	Visual	2
P02	Scratch	0	Blind	2	Visual	1
P03	Scratch	0	Visual	2	Blind	1
P04	Scratch	0	Visual	2	Blind	1
P05	Scratch	0	Blind	2	Visual	1
P06	Scratch	0	Blind	2	Visual	1
P07	Scratch	0	Blind	1	Visual	2
P08	Scratch	0	Blind	1	Visual	2
P09	Scratch	0	Visual	1	Blind	2
P10	Scratch	0	Visual	1	Blind	2

Table 2: Distribution of task and text orders among the participants

For the Visual and Blind tasks, each of the 28 segments in the texts was randomly assigned one of four possible types of translation suggestions – exact matches, fuzzy matches in the 70-84% range, fuzzy matches in the 85-99% range and machine translation feeds – resulting in seven translation suggestions of each type per text. An authentic IBM translation memory was used as a reference for producing the exact and fuzzy matches, without any special tricks being inserted intentionally. The machine translation feeds came from a commercial Moses (Koehn et al., 2007) statistical engine that had been trained with product-specific terminology and was used in production for regular IBM projects in the company.

In the Visual task, one translation suggestion was provided for each segment, and the translators had to actively insert it in the editing area and edit it if they considered it to be a usable suggestion, or they could type their translation either from scratch or on top of the source text. The most common way for the translators to insert translation suggestions was by using a keyboard shortcut, although in some cases they preferred to copy and paste either the whole or parts of the suggestions. In this task, translation suggestions were provided with metadata, which in IBM TranslationManager are indicated by means of a letter placed to the left of the suggestion: blank for exact matches, "f" for fuzzy matches and "m" for machine translation feeds. Additionally, in the case of fuzzy matches, the tool highlights the text portions that differ between the source text in the active segment and the source segment in the translation memory.

In the Blind task, there was also one translation suggestion per segment, but the suggestion had been previously inserted in the segment, so the file displayed as pre-translated text to be edited, instead of source text to be replaced with a translation suggestion. The application panes where the translation suggestions are usually displayed were empty, so no translation metadata were displayed.

2.3. Interviews

The interviews were conducted immediately after the translation tasks, both as semi-structured dialogues and as retrospection with replay (see Hansen, 2008). The base questions asked during the dialogues were:

1) Do you think you translated *faster* in any of the environments? If so, in which one?
2) Do you think the quality of your final translation was *better* in any of them? If so, in which one?
3) In which environment did you feel more *comfortable* working?

During the retrospection, the translators watched selected passages from their performance recordings and commented on certain aspects of the translation tasks based on prompts

from the researcher. For two participants it was not possible to carry out the retrospection, because of technical reasons (P05) and because one participant refused to do it (P08).

3. Data collection and processing

The translation processes were recorded with BB FlashBack and Inputlog (Leijten & van Waes, 2013). This made it possible to measure the total time spent and the total number of characters typed by each translator in each task. All translations were then assessed for quality by two reviewers, who had been revising this type of material for 12 and 19 years in the company. The reviewers revised the translations as Word documents by highlighting their corrections with the *Track Changes* feature. The severity of errors had been previously identified through a series of interviews with project managers in the company, based on their common practice for this type of translation project. Errors related to misinterpretation of the original, missing or added information, tag corruption and misspelt brand names scored two points. Errors such as inconsistencies, misspellings, wrong grammar and punctuation scored one point. Other text issues such as those related to style and fluency were not taken into account. The researcher acted as a third reviewer, making small adjustments to the scores when the two reviewers had too different opinions and marking any obvious errors that had not been detected by the reviewers.

As for the qualitative data, the interviews were recorded then transcribed and coded. In order to better visualise the results, tables were created for each subject, where the verbal data was organised according to the three tasks (Scratch, Visual, Blind) and the three main variables: time (verbalised as 'speed'), effort (verbalised as 'comfortable') and quality.

A third and last data analysis step was necessary to make the qualitative and quantitative data comparable. The approach used here was to rank each variable in each of the tasks for each subject, both as measured and as perceived, and then to compare the rankings. The next section explains this method and presents the results.

4. Results and analysis

4.1. Quantitative data

Table 3 shows the measured results for all ten subjects. *Time* is indicated as seconds per 100 source words. *Edits* is a percent ratio between the total number of relevant key presses and the total number of characters in the final target text, including spaces. *Errors* is the total number of weighted errors (as explained in the previous section) per 100 source words.

	TIME			EDITS			ERRORS		
Participant	Scratch	Visual	Blind	Scratch	Visual	Blind	Scratch	Visual	Blind
P01	257	191	200	102	14.0	11.9	3.8	1.0	1.0
P02	235	167	229	97	22.8	15.7	1.3	1.9	1.4
P03	324	215	193	103	50.0	13.3	5.5	4.3	4.5
P04	566	223	266	103	16.8	15.4	3.8	3.1	3.6
P05	259	121	157	106	12.4	11.5	5.1	4.3	4.2
P06	296	143	162	102	12.0	12.1	4.2	4.8	5.0
P07	613	232	334	109	13.0	14.5	3.8	3.3	3.1
P08	777	497	343	132	29.8	18.6	3.0	1.2	1.9
P09	344	139	139	153	22.6	6.37	8.9	5.4	5.0
P10	240	139	120	108	16.1	9.32	3.0	4.3	5.2

Table 3: Measured times (seconds/100 words), edits (%) and errors (weighted errors/100 words) per participant in the three translation tasks

In Table 4, the values shown in Table 3 are converted into score levels. Thus, for each particular subject and for each variable in Table 3, the task with the lowest number is assigned level 1 in Table 4, the task with the highest number is assigned level 3 and the intermediary task is assigned level 2. When the difference between two tasks is not relevant, considering a deviation of ± 5 percent, the same level is assigned to more than one task, giving preference to the extreme levels 1 and 3. The reason for preferring the extremes is that it corresponds better to human perception and to the types of answers available from the interviews (e.g. the fastest task vs. the slowest task).

	TIME			EDITS			ERRORS		
Participant	Scratch	Visual	Blind	Scratch	Visual	Blind	Scratch	Visual	Blind
P01	3	1	1	3	2	1	3	1	1
P02	3	1	3	3	2	1	1	3	1
P03	3	2	1	3	2	1	3	1	1
P04	3	1	2	3	2	1	3	1	3
P05	3	1	2	3	1	1	3	1	1
P06	3	1	2	3	1	1	1	3	3
P07	3	1	2	3	1	2	3	1	1
P08	3	2	1	3	2	1	3	1	2
P09	3	1	1	3	2	1	3	1	1
P10	3	2	1	3	2	1	1	2	3

Table 4: Measured times, edits and errors as a score level in the three translation tasks

Table 4 indicates that all translators spent the most time and made the most edits (represented by the number 3) when translating from Scratch. The same cannot be said about the errors, since three of the translators made the fewest errors when translating from Scratch. The table also shows that most translators performed the fewest edits in the Blind task, except for one translator, who typed less in the Visual task. More will be said about the results in this table when comparing them with the translators' perceptions.

4.2. Qualitative data

Table 5 shows how the translators perceived their performance after the translation tasks, as a result of coding the interview data.

	TIME			"EFFORT"			ERRORS		
Participant	Scratch	Visual	Blind	Scratch	Visual	Blind	Scratch	Visual	Blind
P01	3	3	1		1		1	1	1
P02	3	1	1		1	2		1	1
P03		1		3	1	1	3	1	1
P04		1			1			1	3
P05		1			1			1	
P06	2	1	3	2	1	3	2	1	3
P07		1		2	1	3	2	1	3
P08	3	2	1		1	1		1	2
P09		1		2	1	3			3
P10		1	2		1			1	1

Table 5: Perceived time, effort and errors as a score level in the three translation tasks

The blank cells in the table represent data for which no clear answer was given in the interview. As a general observation, the table shows that all participants thought they made

fewer errors and invested less effort in the Visual task than in any of the two other translation tasks, and that most of them considered they spent the least time on the Visual task. In the following sections, we will compare the measured and perceived data in detail for each of the dependent variables.

4.3. Comparison between quantitative and qualitative data

The time measured per 100 words was consistently higher when translating from scratch for all ten participants. This is in accordance with their perception, except for one translator, who thought he spent less time translating from scratch than he did in the Blind task. For the seven translators who thought they were faster in the Visual task than in the Blind task (P03, P04, P05, P06, P07, P09, P10), all but two (P03 and P10) were indeed faster. For the two translators who thought they were faster in the Blind task than in the Visual task (P01, P08), their perception corresponded to their measured times. The only participant who thought he was as fast in the Visual as in the Blind task (P02) was actually much faster in the Visual task. P01 thought she spent the least time on the Blind task, whereas she actually spent less time on the Visual task.

Seventy percent of translators made the most errors when translating from scratch, which might indicate their reliance on translation suggestions, after many years of practice working with translation memories. There was no clear advantage between the Visual and the Blind tasks in terms of error rates, although all the translators thought they made the fewest errors in the Visual task, except for one translator, who did not distinguish explicitly between the Visual task and translating from scratch. Their perception corresponded with the reviewers' quality assessment in 70 percent of the cases, whereas two translators actually made the most errors in the Visual task and one translator made more errors in the Visual task than when translating from scratch.

As indicated in Table 4, the Blind task was the condition in which the translators typed the least, except for one translator, who typed less in the Visual task. Two translators typed as much in the Blind task as in the Visual task. A simple comparison of the middle columns in Table 4 and Table 5 reveals no coincidence between the measured edits and the perceived "effort" while performing the task. This could be attributed to any of the factors mentioned in Section 4.4, but in this case, the discrepancies in the results are probably due to a poorly formulated question. The quantitative variable being measured as an indication of effort was the amount of editing, which is a simple measurement of physical effort, while in the interviews the translators were asked about the task in which they felt more "comfortable". It turns out that typing effort and the feeling of "comfort" while performing a task are not directly comparable. This is in accordance with the conclusions of other studies, such as Koponen, Aziz, Ramos, and Specia (2012), who suggest that "keystrokes, while very useful as a way to understand how translators work, may not be an appropriate measure to estimate cognitive effort" (p. 20).

4.4. Additional information from the interviews

A major goal of the interviews was to let participants express their priorities. This was achieved through a relatively free dialogue format, which was responsible for some missing data in Table 5, but also allowed other factors to come into play that had not been included as the main variables in the study.

Translation vs. revision vs. post-editing

The interviews indicate a clear difference in the way translators perceived the two main translation tasks. All participants except one made a clear distinction between "translate", for the Visual task, and "revise" or "proofread" ("revisar", in Spanish) or "post-edit", for the Blind

task.[2] The quantitative data support this perception, as they show many more iterations per segment in the Visual environment, as if the translators were first translating, then self-revising. In the Blind environment, which they considered to be revising or post-editing, they completed the task in a single round. This difference made seven of the translators feel that they had performed a regular revision (on text that had been translated or proofread by another translator) when working in the Blind task (my translations here and throughout):

> P10: I'm very much used to working the first way, to translate. I had never done the other task before actually, to find everything at 100% and to revise it.

> P02: The other one was already done, we just had to revise.

> P01: Post-editing, a revision that had already been done and that I had to revise.

For these participants, the text they were "revising" was in principle better than the text they had in the Visual task:

> P08: We assume that in theory it should be better.

> P04: There was a lot of [translation] memory and it was quite good compared with other folders.

> P07: We could notice some segments had been leveraged from the memory... they were better, I didn't have to change much.

Only one participant felt she was "translating" when performing the Blind task: she actually talked about both tasks in terms of the presence or absence of metadata on the translation suggestions (P05).

The role of translation suggestions

Seven translators acknowledged the usefulness of translation suggestions (as opposed to translating from scratch):

> P02: Because [when you translate from scratch] you have to think more.

> P03: It always helps to have pre-translated stuff or when there is something previous that is useful, because if you translate everything from scratch, you always make mistakes, [it's a little] more difficult. Having something as a basis is always welcome.

> P06: When you have a suggestion from the memory, you insert it and if you change a word, maybe you go faster too, with some memory. [Pause] Translating 500 words with memory suggestions is faster than from scratch…

> P07: Because you have an external aid from previous memories and machine translation [...] you always go faster. [...] it is always better to have some help.

One of those participants (P09), however, pondered that it might be easier to translate from scratch:

[2] In the current state of play, with MT and TM suggestions being presented together, it is not surprising that no clear distinction is made between post-editing and revising.

P09: It is easier to translate from scratch, because I don't have to look at anything. And I don't need to check if what is suggested is correct or not, or if it's in the right order or in the wrong order.

Along the same lines, P01 said:

P01: I don't think it is especially faster having the memory, because when you translate from scratch, one advantage I can see is the vocabulary, but the other is that there is no suggestion to look at, no differences to check for between one sentence and the other. [...] I think I compensate what I use—the help from the memory—with the time I spend checking the passage, checking for differences.

The role of metadata

Even if the translators did not consider the metadata to be the main distinction between the Visual and the Blind tasks in their comments, they demonstrated awareness of how translation metadata could help them:

P01: If I see a fuzzy match, the first thing I'll look at is the Source of Proposal. For me it's easier with a memory, with fuzzy matches, with information on whether it comes from MT or from fuzzy or whatever, because it allows me to look at it in one way or another.

P02: If you see that it's 100%, that it's not machine translation, then, in principle, in an everyday translation, when you go fast, you don't even look at it. You assume it's correct or that you translated it yourself before [...] A fuzzy match, if I see that everything is translated and there is only one word that changes, I change that word, I don't even look at the rest.

P04: Because you can't see below where it comes from... [when there is no metadata]

P05: TM/2 indicates the fuzzy matches... it highlights what is missing, what is extra, what has changed.

P06: You always look at what has changed and you change there. [...] You didn't even need to read the sentence, you just had to change a word that was highlighted and that's it.

P08: When it's pre-translated you don't have... you don't know the quality of the suggestion; in contrast, when you have the memory, you know if it's an MT suggestion or if it comes from a... from another publication. TM/2 indicates if it's an Exact Match or if it's an MT suggestion or if it's a fuzzy match... [...] Sometimes you just look at what has changed. On the other hand, when you have it pre-translated, I don't know where it comes from... I would prefer to know... the environment where you see the suggestion, if it's machine translation, if it's... or if it comes from another publication that has been checked by somebody else. I think it's better to have the information, because it tells you what has changed; so if you know what's changed, you focus more on what's changed. Your natural tendency is to trust more what appears as unchanged.

P09: The second one [Visual] had several fuzzies at 95%, 85%, so it's very easy to detect where the small changes are, and it's very useful. [...] If you look at the suggestion, since it tells you exactly what the changes are, it's easier to detect. [...] For me it's much easier to upload or to edit.

Morado Vázquez (2012) obtained similar feedback from the translators in her study: "In terms of participants' attitude towards the metadata received, most of the participants did not find it distracting, and the majority of them would prefer a translation memory which contained metadata." It is worth noting, however, that one of my translators stated, "the environment that gives you more information is, at the same time, more complex" (P08).

The perception of machine translation

In general, the participants had mixed feelings about machine translation. Although in some cases they criticised it as being poor, they also recognised that some machine-translated segments were "almost perfect" and that MT helped them increase productivity.

Two translators felt the text in the Blind task contained more machine-translated segments than the text in the Visual task, although the translators were told that both texts actually had the same distribution of suggestion types, and only 25% of the suggestions were actually machine translation feeds (see Section 2.2). Therefore, in their comments the translators made statements about the (presumably lower) quality of the translation suggestions based on their assumption that the suggestions came from machine translation:

> P06: In the revision task, since they come from machine, they are always faulty.

> P09: [The Blind task] is mostly machine, so it takes me longer to think about what changes [...]. I do have to keep thinking what the core of the segment is and to change it.

He et al. (2010) and Guerberof (2013, pp. 87–88) also show evidence that translators tend to trust fuzzy matches more than they trust machine translations and that in many cases subjects are not able to tell TM suggestions from MT suggestions.

Task familiarity

Eight out of the 10 participants (P01, P02, P04, P05, P06, P07, P09, P10) reported being more comfortable tackling the Visual task, even when some believed the Blind task could be faster. The other two participants (P03 and P08) were equally comfortable working in the Blind task. P08 found the Blind task "more simple":

> P08: You look at the English, the Spanish and that's it. [...] In the other one, you have to look at the English, the Spanish, and sometimes choose among five suggestions – not the case in this experiment though, where you had only one suggestion.

The main reason given by the translators (mentioned by 7 out of 10) for feeling more comfortable and actually preferring the Visual task was that they were very "used to" or "more familiar with" (in Spanish: "acostumbrado", "familiarizado", "habituado") the Visual task, while the Blind task was new to them. Another reason given by the translators (3 out of 10) for preferring the Visual task was that they felt more confident in this environment. It is unclear in some statements whether this feeling of confidence is only related to task familiarity or also to the metadata or to any other characteristics present in the Visual task.

> P01: I prefer to translate with a memory. [...] For me it's more comfortable, it makes me feel more confident.

> P04: Surely because this is what I've been doing for IBM lately, [I feel] more confident, maybe more familiar with it.

P08: If you know it's a fuzzy match, since you know it has been checked by a human translator, it gives you more confidence.

Different strategies

Since all the participant translators were used to doing revisions in IBM TranslationManager, where the text to be revised comes pre-translated (but with metadata on the provenance of existing translations), their feeling of unfamiliarity or lack of confidence with the Blind task can probably be explained by the absence of metadata in this task. This suspicion is reinforced by several statements in which translators explain that they use different strategies for exact matches, fuzzy matches and machine translation:

P01: If I see it's an "m" [machine translation], I read the sentences from A to Z, or I go and check for some things or I look for some things or for other things. If I see a fuzzy match, the first thing I'll look at is the Source of Proposal. For me it's easier with a memory, with fuzzy matches, with information on whether it comes from MT or from fuzzy or whatever, because it allows me to look at it in one way or the other. If I see a fuzzy match, I look at the Source of Proposal; if I see an MT, that is, if I see an "m", and it gives me the impression that the sentence is more or less correct, then I insert it and, depending on the case, I fix it, because sometimes the sentence is almost entirely perfect.

P02: If you know it's... you look at it differently. If you see that it's 100%, that it's not machine translation, then, in principle, in an everyday translation, when you go fast, you don't even look at it.

P08: [...]if you know it's MT, you look at it with more... respect. Conversely, if you know it's a fuzzy match, since you know it has been checked by a human translator, it gives you more confidence. Sometimes you just look at what has changed.

These testimonials are in accordance with feedback provided by participants in other studies (O'Brien, 2006, p. 198), as different types of translation tasks seem to activate different translation strategies and to require different allocation of cognitive resources (Carl, Kay, & Jensen, 2010; Dragsted, 2012; House, 2000; Hvelplund, 2011; Jääskeläinen, 1993; Lörscher, 1991). The fact of knowing which type of suggestion is being dealt with when processing a segment could reduce cognitive load and account for the reported feeling of comfort.

5. Discussion

Although the quantitative results between the three environments do not show a clear advantage when translating a specific task, participants preferred to work on the more traditional Visual task, with translation suggestions and metadata. This might be explained by a feeling of increased performance in some cases, as they tended to over-rate the Visual task, but also by task familiarity and the increased level of confidence resulting therefrom.

The metadata factor (present in the Visual task, absent in the Blind task) did not correlate with a consistent increase in performance according to the measured data. A more in-depth analysis of the experiment results has shown that this factor does have a positive effect on performance indicators for certain types of translation suggestions, namely high fuzzy matches and exact matches. The results presented here indicate that metadata are also a relevant factor to increase confidence and reduce cognitive load, by giving translators a hint on how to initially approach a suggestion, as they reportedly use different strategies for different kinds of suggestions.

In the current experiment, only one translation suggestion was presented for each segment, so the study only allowed us to analyse how metadata can help translators use the one suggestion provided. Since translating with CAT tools usually involves a dual process of selection + repairing of suggestions, it would be interesting to complement the current study with a follow-up experiment including multiple suggestions, to investigate how metadata can also help translators choose among different proposals. Likewise, the experiment could be extended by isolating the "pre-translation" and "metadata" factors, as in the current study both those variables were playing a role: one task had pre-translation and no metadata and the other one had "regular translation" and metadata.

The pre-translation factor has also proved to affect translators psychologically in the way they approached the text and the trust they attributed to the proposals – having being previously translated by an (assumedly reliable) human translator.

In the interviews, the question "In which environment did you feel more comfortable?" assumed that "comfortable" (Spanish "cómodo") might inversely correlate with typing effort. This proved to be a very naive assumption, as comfort seems to correlate more with long-time experiential factors than with momentary task characteristics. If a similar experiment is reproduced, the question to be asked should be simply "In which environment do you think you typed more?". Alternatively, a different measurement for cognitive effort should be used.

Still regarding the interviews, a better strategy should be found to elicit answers for the variables in all tasks, in order to have all cells completed in Table 5, while still making sure the answers are not influenced by the researcher's prompts. The interview data in this study are admittedly incomplete, but they have still provided enough information to draw relevant conclusions about the translators' perceptions.

6. Conclusion

The goal of this paper was two-fold: first, to propose a translation environment where suggestions coming from a translation memory and from machine translation could be compared on a fair basis; second, to compare the measured performances and perceived performances of professional translators when exposed to different translation conditions.

The first goal was pursued by setting up two tasks in the same tool, one that emulated a typical TM-assisted workflow and another one that was more typical of post-editing environments. Some problems were found and the task setups should still be improved in future studies to bring both tasks closer to real scenarios.

The second goal was pursued by ranking the measured performances, ranking the perceived performances and comparing both rankings. Not all expected answers could be elicited during the interviews, but the missing data did not prevent us from making conclusive observations. The main conclusion is that translators' perceptions about their performances do not always correlate with their actual performances. The interviews also provided additional information on topics such as task familiarity and translation strategies, indicated that translators tend to associate pre-translated text with revision and post-editing, and gave hints on the translators' opinions about machine translation.

The study found that the measured performances were positively affected by the presence of translation suggestions, but not so much by the presence of translation metadata. However, the interviews indicate that translators preferred the task with translation metadata, even when it did not correlate with an improved performance. Most of the participants felt more comfortable handling this task and had the impression it allowed them to work faster and to make fewer errors. The main reason identified for the positive perception of the Visual task was task familiarity.

A general correlation between being familiar with a task and preferring to do that task is not a particularly surprising result. Indeed, it seems to follow a general trend related to the

adoption of new technologies, as previously reported by studies such as Dillon and Fraser (2006). However, it might suggest that practice is a major factor to improve job satisfaction, even if it does not always imply increased performance.

Acknowledgments

I would like to thank Anthony Pym and three anonymous reviewers for their comments on earlier versions of this manuscript. I would also like to acknowledge the funding to my doctoral research, provided through the European Commission's TIME Marie Curie fellowship (FP7-PEOPLE-2010-ITN-263954).

References

Allen, J. H. (2003). Post-editing. In H. L. Somers (Ed.), *Computers and translation. A translator's guide* (pp. 297–317). Amsterdam, Philadelphia: John Benjamins Pub. Co.

Almeida, G. de. (2013). *Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages* (Doctoral thesis). Dublin City University, Dublin. Retrieved from http://doras.dcu.ie/17732/

Anastasiou, D., & Morado Vázquez, L. (2010). Localisation Standards and Metadata. In S. Sánchez-Alonso & I. N. Athanasiadis (Eds.), *Communications in Computer and Information Science. Metadata and Semantic Research* (pp. 255–274). Berlin, Heidelberg: Springer Berlin Heidelberg.

Aziz, W., Sousa, S., & Specia, L. (2012). PET: A Tool for Post-editing and Assessing Machine Translation. In : *LREC, Eighth International Conference on Language Resources and Evaluation* (pp. 3982–3987). Instanbul, Turkey. Retrieved from http://www.mt-archive.info/LREC-2012-Aziz.pdf

Carl, M., Kay, M., & Jensen, K. T. (2010). *Long Distance Revisions in Drafting and Post-editing: Paper presented at CICLing-2010, Iaşi, Romania*. Retrieved from http://research.cbs.dk/portal/en/publications/long-distance-revisions-in-drafting-and-postediting%28fd3ffefc-6ea1-4362-9fc4-be80fef79af7%29/export.html

CASMACAT. (2014). *Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation*. Retrieved from http://www.casmacat.eu/

Colominas, C. (2008). Towards chunk-based translation memories. *Babel*, *54*(4), 343–354. doi:10.1075/babel.54.4.03col

Dillon, S., & Fraser, J. (2006). Translators and TM: An investigation of translators' perceptions of translation memory adoption. *Machine Translation*, *20*(2), 67–79. doi:10.1007/s10590-006-9004-8

Dragsted, B. (2004). *Segmentation in Translation and Translation Memory Systems: An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process* (Doctoral thesis). Copenhagen Business School, Frederiksberg.

Dragsted, B. (2012). Indicators of difficulty in translation — Correlating product and process data. *Across Languages and Cultures*, *13*(1), 81–98. doi:10.1556/Acr.13.2012.1.5

Garcia, I. (2007). Power shifts in web-based translation memory. *Machine Translation*, *21*(1), 55–68. doi:10.1007/s10590-008-9033-6

Garcia, I. (2011). Translating by post-editing: is it the way forward? *Machine Translation*, *25*(3), 217–237. doi:10.1007/s10590-011-9115-8

Green, S., Heer, J., & Manning, C. D. (2013). The efficacy of human post-editing for language translation. In W. E. Mackay, S. Brewster, & S. Bødker (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 439–448).

Guerberof, A. (2009). 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.

Guerberof, A. (2013). What do professional translators think about post-editing? *The Journal of Specialised Translation*, (19), 75–95. Retrieved from http://www.jostrans.org/issue19/art_guerberof.php

Guerra Martínez, L. (2003). *Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output* (Master's thesis). Dublin City University, Dublin.

Hansen, G. (2008). The dialogue in translation process research. In *Translation and Cultural Diversity. Selected Proceedings of the XVIII FIT World Congress 2008* . Shanghai, China: Foreign Languages Press. Retrieved from http://www.translationconcepts.org/pdf/Hansen_ArticleMethods.pdf

He, Y., Ma, Y., Roturier, J., Way, A., & van Genabith, J. (2010). Improving the Post-Editing Experience using Translation Recommendation: A User Study. In *AMTA 2010 - 9th Conference of the Association for Machine Translation in the Americas* . Retrieved from http://amta2010.amta-web.org/AMTA/papers/2-27-HeMaEtal.pdf

House, J. (2000). Consciousness and the Strategic Use of Aids in Translation. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and mapping the processes of translation and interpreting. Outlooks on empirical research* (pp. 149–162). Amsterdam/Philadelphia: John Benjamins.

Hvelplund, K. T. (2011). *Allocation of cognitive resources in translation. An eye-tracking and key-logging study* (Doctoral thesis). Copenhagen Business School, Frederiksberg.

Jääskeläinen, R. (1993). Investigating translation strategies. In S. Tirkkonen-Condit & J. Laffling (Eds.), *Kielitieteellisiä tutkimuksia, Studies in languages: Vol. 28. Recent trends in empirical translation research* (pp. 99–120). Joensuu: Joensuu University. Retrieved from http://scholar.google.com/scholar?cluster=8725738219265795995&hl=en&as_sdt=0,22

Karamanis, N., Luz, S., & Doherty, G. (2011). Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation*, *25*(1), 35–52. doi:10.1007/s10590-011-9093-x

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., . . . Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P07-2045

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. In : *WPTP, AMTA 2012 Workshop on Post-Editing Technology and Practice* (pp. 11–20). San Diego, USA. Retrieved from http://www.mt-archive.info/AMTA-2012-Koponen.pdf

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes.* (Koby, G. S., Ed.). Kent, Ohio: Kent State University Pr.

Leijten, M., & van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, *30*(3), 358–392. doi:10.1177/0741088313491692

Lörscher, W. (1991). *Translation performance, translation process and translation strategies: A psycholinguistic investigation*. Tübingen: Gunter Narr.

Moorkens, J. (2012). *Measuring Consistency in Translation Memories. A Mixed-Methods Case Study* (Doctoral thesis). Dublin City University, Dublin.

Morado Vázquez, L. (2012). *An empirical study on the influence of translation suggestions' provenance metadata* (Doctoral thesis). University of Limerick, Limerick.

O'Brien, S. (2006). Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, *14*(3), 185–205. doi:10.1080/09076760708669037

Plitt, M., & Masselot, F. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, *93*, 7–16. doi:10.2478/v10108-010-0010-x

Skadiņš, R., Puriņš, M., Skadiņa, I., & Vasiļjevs, A. (2011). Evaluation of SMT in localization to under-resourced inflected language. In M. L. Forcada, H. Depraetere, & V. Vandeghinste (Eds.), *Proceedings of the 15th conference of the European Association for Machine Translation* (pp. 35–40).

Teixeira, C. S. C. (2014). The handling of translation metadata in translation tools. In S. O'Brien, L. Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of machine translation. Processes and applications* (pp. 109–125). Newcastle upon Tyne: Cambridge Scholars Publishing.

Wallis, J. (2006). *Interactive Translation vs Pre-translation in the Context of Translation Memory Systems. Investigating the effects of translation method on productivity, quality and translator satisfaction* (Master's thesis). University of Ottawa, Ottawa. Retrieved from http://www.localisation.ie/resources/Awards/Theses/Thesis%20-%20Julian%20Wallis.pdf

Webb, L. E. (1998). *Advantages and Disadvantages of Translation Memory. A Cost-Benefit Analysis* (Master's thesis). Monterey Institute of International Studies, Monterey.

Yamada, M. (2011). *Revising text: An empirical investigation of revision and the effects of integrating a TM and MT system into the translation process* (Doctoral thesis). Rikkyo University.