Experiments in Non-Coherent Post-editing

Mª Cristina Toledo Báez
University of Córdoba, Plaza del Cardenal Salazar 3, 14003 Córdoba, Spain
cristina.toledo@uco.es
Moritz Schaeffer
University of Mainz, An der Hochschule 2, 76726 Germersheim, Germany
mschaeffer@uni-mainz.de
Michael Carl
Renmin University, 59 Zhongguancun St, Haidian Qu, China, 100872
Copenhagen Business School, Dalgas Have 15, 2000 Frederiksberg, Denmark
mc.isv@cbs.dk

Abstract

Market pressure on translation productivity joined with technological innovation is likely to fragment and decontextualise translation jobs even more than is currently the case. Many different translators increasingly work on one document at different places, collaboratively working in the cloud. This paper investigates the effect of decontextualised source texts on behaviour by comparing post-editing of sequentially ordered sentences with shuffled sentences from two different texts. The findings suggest that there is little or no effect of the decontextualised source texts on behaviour.

1 Introduction

Machine Translation (MT) has made tremendous progress in the past two decades, first since the introduction of statistical approaches (Brown et al., 1988) and more recently with the emergence of neural-based approaches, also referred to as neural MT (NMT) (e.g., Wu et al., 2016). Klubička et al. (2017) have found that NMT reduces the amount of errors in the translation produced (for news content type, English-to-Croatian) by 54%, as compared to SMT.

Despite the tremendous increase of MT quality in the past years, post-editing of MT output (PEMT) remains a compulsory activity if the translation product is to be used for dissemination. However, better MT output leads to quicker post-editing cycles and increases productivity and efficiency (Specia, 2011). Thus, the even better quality of NMT output is likely to be well suited for post-editing, as it reaches an unprecedented degree of fluency (Toral, 2017)

In a typical PEMT scenario, a human post-editor corrects the translation generated by an MT system. For instance, many translation memory systems (Trados, MemoQ, OmegaT, etc.) provide access to MT systems, the output of which may be merged into the set of translation proposals that are retrieved from the translation base of the TM system.

Recently, the possibility of active learning and active interaction has emerged (Ortiz-Martínez, 2016; Martínez-Gómez et al., 2012; Peris et al., 2016), in which an MT system re-orders the source language segments to be translated and post-edited so as to maximise productivity and the learning effect. Instead of presenting a text in its original sequential order, the system sorts the segments according to a degree of confidence, so that it can learn most (quickly) from the human corrections. This comes along with novel conceptualizations of the translation workflow which link these new possibilities with innovative usage of crowdsourcing. In order to fully exploit the potential in crowdsourcing, novel ways need be found to split up coherent texts into units that can be translated and edited independently by a large number of translators at the same time. Due to the increased demand for translation productivity and shorter translation turnaround, some translation tools (Wordbee, Trados, MATECAT) offer collaborative functionalities. Some LSP companies (Unbabel[1], MotaWord[2]) are seeking possibilities to experiment with a more dynamic approach to collaborative translation that segments a document into smaller units. MotaWord, for instance, declares to be "The World's

[1] https://unbabel.com/ [17.6. 2017]

[2] https://www.motaword.com [17.6. 2017]

The Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT), pages 11–20,
Varna, Bulgaria, Sept 7, 2017
https://doi.org/10.26615/978-954-452-042-7_002

Fastest Human Translation Platform" which is based on a collaborative cloud platform "coordinated efficiently through a smart back end" in which over 9,000 translators participate. This is only possible if large documents are split into small segments and by deploying the crowd to post-edit a limited number smaller units. However, it is unclear how translators cope with a situation in which smaller segments - possibly from different parts of the same document - are presented out of context. The impact on translation behaviour has, to our knowledge, never been studied if translators translate segments in a non-sequential order.

In this paper, we investigate the translation processes of post-editors when dealing with segments in a randomized order. We observe translators' post-editing behaviour using research methods (eye tracking and key-logging) and metrics known in the research field of Translation Process Research. Dragsted (2004:32) points out that, from a prescriptive perspective, the translation unit can be considered "the most appropriate segment for establishing SL/TL equivalence" (see, among others, Vinay and Darbelnet, 1995; Catford, 1965; Bell, 1991). From a descriptive, cognitive-oriented perspective, Dragsted (2004:32) argues that the translation unit can also be described "as the segment actually processed, […] identified on the basis of cognitive processes observable (indirectly) in a set of data." What constitutes the ideal translation unit has received considerable attention: there are several proposals for a cognitively or linguistically plausible unit (e.g. Dragsted 2005, 2006; Carl and Kay 2011; Jakobsen 2011; Alves and Gonçalves 2013). However, for the purpose of this study, and practical reasons, we define a translation unit as a sentence (segment) as demarcated by full stops.

2 Experimental Setup

16 Translation students and 4 professional translators post-edited four English texts into Spanish. The texts were taken from the TPR-DB multiLing[3] corpus. Two of the texts were news texts (Text 1 and 3) and two texts were taken from a sociological encyclopaedia (Texts 5 and 6). Every source text (henceforth ST) had between 122 and 160 words (5-11 segments) and all four texts were machine translated using google translate, as of June 2016 and post-edited. One news text and one sociological text were presented in the original coherent form, and two texts were composed of mixed sentences from the two other texts. Translog-II (Carl 2012) was used as a post-editing tool. A line break separated each new segment in the source and the target side. In total, 80 post-edited texts were collected: 40 posteditions (284 segments) in the mixed-segment mode, and 40 posteditions (284 segments) in the coherent translation mode.

Table 1 shows the mean total duration per post-edited text (*Dur*) in minutes in the two conditions (P=coherent mode, Pm=mixed mode), for the orientation, the drafting and the revision phases. *TrtS* is the total time spent reading the ST and *TrtT* is the total time spent reading the target text (henceforth TT), also in minutes. Deletions and Insertions are counted in characters. It is obvious from the means that the order in which the segments are shown has little (in the case of insertions and deletions) effect or no effect on average values.

Task	**Total Duration**	**Orientation**
P	5.97 (4.21)	0.53 (0.47)
Pm	5.69 (2.70)	0.53 (0.43)
P	**Draft**	**Revision**
Pm	4.58 (3.43)	0.86 (0.94)
	4.50 (2.28)	0.65 (0.71)
	TrtS	**TrtT**
P	1.46 (1.12)	3.72 (2.87)
Pm	1.44 (0.74)	3.39 (1.86)
	Deletions	**Insertions**
P	125 (61)	134 (73)
Pm	142 (60)	144 (65)

Table 1: descriptive statistics for the data: means and standard deviation in parentheses.

3 Translation Difficulty Indicator

Mishra et al (2013) develop a Translation Difficulty Index (TDI) which aims at predicting the effort during translation, measured in terms of the sum of ST and TT reading times (TDI score). They show that the TDI score correlates with the degree of polysemy, structural complexity and length of ST segments. They train a Support Vector Machine on observed eye movement data and predicted the

[3] https://sites.google.com/site/centretranslationinnovation/tpr-db

TDI score of unseen data during translation on the basis of the linguistic features.

The segments in the mixed post-editing mode were ordered according to the Translation Difficulty Indicator (TDI) (Mishra et al., 2013), ordering from the highest to the lowest TDI. The texts which resulted from merging two texts were then split up again into two texts which were post-edited independently from each other. This resulted in one text each with an overall higher TDI score and one with an overall lower TDI score, given that the segments in the merged text had been ordered by the TDI score from high to low. Texts were presented in a pseudo-randomized order, but post-editors had to post-edit first the two coherent texts and then the two texts in the mixed mode.

Table 2 shows how the segments were order in

STseg	Text	Otext	OSTseg	TDI
1	53	5	2	4.11
2	53	3	1	4.02
3	53	5	5	3.3
4	53	3	5	2.82
5	53	5	1	3.16
6	53	5	6	3.3

Table 2: Ordering of segments in the mixed post-editing mode

the mixed mode. *STseg* is the number in the sequential order in which the ST segments were shown to post-editors. *Text* is the unique identifier for the texts. In this case, there are two merged texts: Text 35 is composed of the segments from the original texts 3 and 5 - Otext shows the text to which the segments originally belong. *OSTseg* shows the sequential numbering of the original (not mixed) texts. The segments in the mixed texts are ordered according to the TDI score (minor adjustments were made to avoid that two segments were shown in the original sequential order).

4 Aims and Method

It is the aim of the current study is to investigate whether text level coherence has an effect on production speed in general and on eye movement behaviour and cognitive effort in particular. In other words, it is the aim to find out whether presenting segments belonging to two different texts in a relatively random order has an effect on cognitive effort.

For all the analyses in the present study, R (R Development Core Team, 2014) and the lme4 (Bates et al., 2014) and languageR (Baayen 2013) packages were used to perform linear mixed-effects models (LMEMs). To test for significance, the R package lmerTest (Kuznetsova et al., 2014) was used. The R^2 for LMEMs was calculated with the MuMIn package (Bartoń 2009). Data that were more than 2.5 standard deviations below or above the participant's mean for the individual measure were excluded from analyses. All the LMEMs had participant and item as random variables.

5 The Effect of Text Level Coherence on Behaviour

For the effect of text level coherence on production duration, the scaled and centred typing duration per segment (*Dur*) was used as dependent variable. The dependent variable was scaled and centred, because the predictors were on very different scales. The variable *Dur* is defined by the keystrokes belonging to a given sentence. It does not include the time that elapses between the last keystroke of the previous sentence and the first keystroke of the current sentence, but it does include any pauses between keystrokes belonging to the same sentence. Given that participants were post-editing, a segment can have a typing duration of zero if the participant did not change anything in the MT output. All potentially relevant variables were entered as predictors in the LMEMs and those which were not significant were excluded. The final model for production duration per segment (*Dur*) had the following predictors: word translation entropy (*HTra*), the number of insertions (*ins*) in characters, Task (post-editing a text in coherent order (*P*) and in the mixed mode (*Pm*)), sequential numbering of the segments as they were shown to participants (*STseg*), the total reading time on the source and target segments, i.e. the total time a particular source segment (*TrtS*) or target segment (*TrtT*) was read, how often the typing was not in a sequential order, i.e. how often the post-editor typed successive keystrokes which were part of two or more different words (*Scatter*) and the number of times a segment has been edited (*Nedit*). Word translation entropy (*HTra*) describes the number of lexical (word translation) choices for the final TTs - the smaller the *HTra* value, the less lex-

ical choices a translator has in the final target sentence (cf. Carl et al. (2016) for a detailed description of this metric).

HTra had a relatively large and significant positive effect on *Dur* (see Table 3): more translation choices induce longer translation times. The effect of the number of insertions was expectedly large, positive and significant. As would be expected, total reading time on the source (*TrtS*) and on the target (*TrtT*) had both very large and highly significant effects. Both *Scatter* and *Nedit* had relatively large significant effects on typing duration (*Dur*).

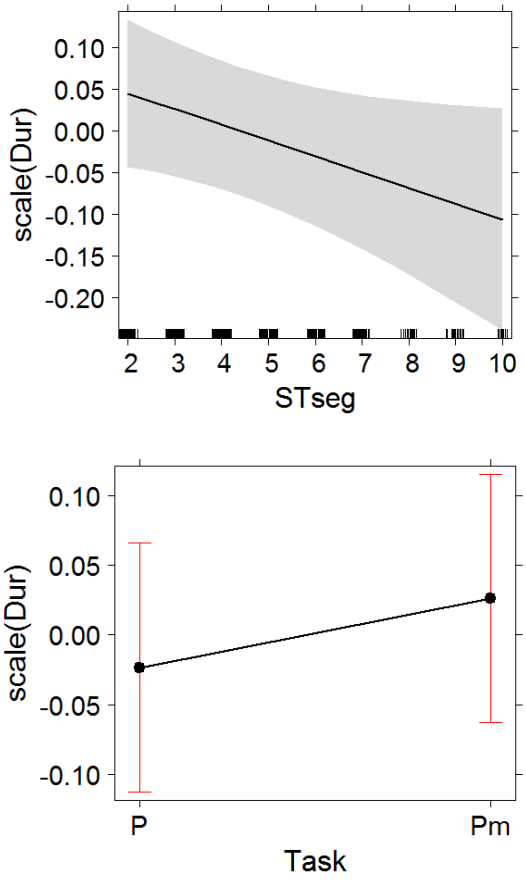

Figure 1: The effect of *Task* and sequential numbering of source segments (*STseg*) on typing duration (*Dur*).

Both *Scatter* and *Nedit* are indicators of revision behaviour suggesting faulty MT output. The Task (coherent versus mixed mode) had no significant effect on typing duration (*Dur*). This result was surprising, given that it could be expected to be more effortful to process a text where segments from two different texts are jumbled in one text. The marginally significant, negative and modest effect of the sequential numbering of segments (*STseg*) (see Figure 1) would have been expected if post-editors had only worked in the coherent mode, since it could be argued that post-editors become more familiar with the topic, the semantic fields and other aspects of the ST and target language as they progress through the text. Schaeffer et al. (2016) show that *STseg* has a facilitating effect on all relevant eye movement measures during from scratch translation and argue that translators create a text level coherence model which makes translation less effortful and the TT more predictable. In the mixed post-editing mode, it is arguably more difficult to create a text level coherence model which would facilitate the process. However, when the interaction between Task and *STseg* was not even approaching significance (β=-0.04, SE= 0.05, t= -0.76, $p < 0.450$). What these results suggest is that both the finding in Schaeffer et al. (2016) and the effect of *STseg* on behaviour is not dependent on textual coherence, but is related to a task facilitation effect - the longer the task is carried out, the easier it becomes. The model for typing duration (*Dur*) without interaction provided a very good fit (marginal $R^2 = 0.74$, conditional $R^2 = 0.77$).

In order to investigate the reading behaviour on the ST, *TrtS* (the sum total of all fixation durations on an ST segment) was used as dependent variable. *TrtS* was log-transformed because it was not normally distributed. The predictors were the number of tokens in the ST segment (*TokS*), word translation entropy (*HTra*), the number of deletions per

Predictor	*β*	SE	*t*	*p*	
HTra	0.09	0.03	2.75	0.007	**
ins	0.15	0.05	3.34	0.001	***
TaskPm	0.05	0.05	1.08	0.282	
STseg	-0.05	0.02	-1.93	0.055	.
TrtS	0.27	0.03	10.65	<2e-16	***
TrtT	0.42	0.03	13.43	<2e-16	***
Scatter	0.13	0.04	2.98	0.003	**
Nedit	0.06	0.02	2.27	0.024	*

Table 3: The LMEM for typing duration (*Dur*).

segment (*del*), and finally Task and *STseg*. No other variables had a significant effect on *TrtS*.

TokS had an expectedly large positive and highly significant effect on *TrtS* (see Table 4). *HTra* also had a relatively large, positive and highly significant effect *TrtS*. This effect has been observed previously for from-scratch translation (Schaeffer et al., 2016a) and simply shows that the less literal a translation is, the more cognitive effort is required to arrive at a TT. The number of deletions had a large, positive and significant effect on *TrtS*. Task had a modest, positive and significant effect on *TrtS*, while *STseg* had a large, negative and highly significant effect on *TrtS* (see Figure 2). Again, the fact that the effect of *STseg* was so large, highly significant and negative was surprising, given that half the texts were post-edited in the mixed mode and it could be argued that it is very difficult to develop a text level coherence model in this mode. The interaction between Task and *STseg* was not significant (see Figure 3), suggesting that the *STseg* effect is a task facilitation effect in both tasks and that this effect is very similar in both tasks. However, text level coherence did have an effect on *TrtS*, suggesting that the lack of coherence requires a modest amount of additional effort when reading the ST. The model for total reading time on the ST (*TrtS*) without interaction provided a relatively good fit (marginal $R^2 = 0.32$, conditional $R^2 = 0.63$).

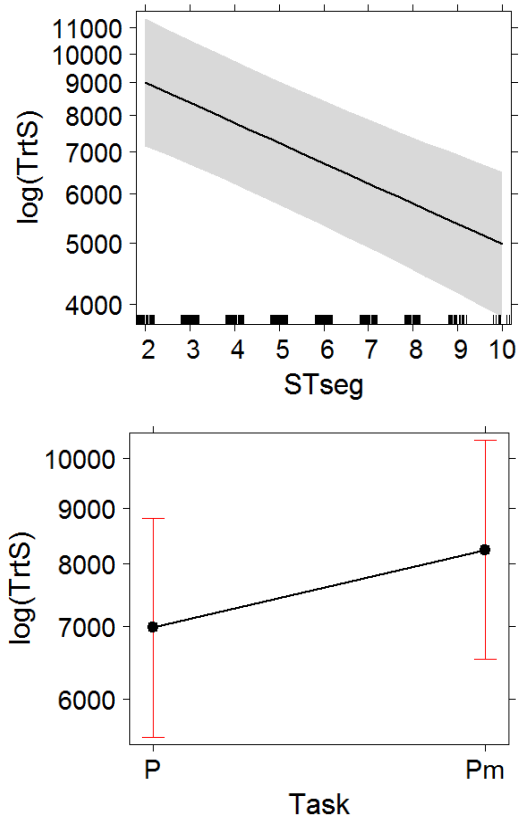

Figure 2:The effect of *Task* and sequential numbering of source segments (*STseg*) on total reading time on the ST (*TrtS*).

Rather than looking only at the absolute time participants spent reading the TT, we also investigated the effect of text level coherence on the percentage of the total reading time participants (*TrtS* + *TrtT*) spent reading the TT (*Perc_TrtT*). Results were broadly similar (the model with the absolute values was more complex and Task had no significant effect on *TrtT*), but the proportional aspect seemed more informative. The final model for *Perc_TrtT* had the following predictors: the number of deletions (*del*), Task, *STseg* and *Nedit*.

Predictor	β	SE	t	p	
TokS	0.44	0.05	8.57	4.12E-10	***
HTra	0.16	0.05	3.50	0.001	***
del	0.11	0.04	3.11	0.002	**
TaskPm	0.16	0.06	2.85	0.005	**
STseg	-0.18	0.03	-5.81	1.31E-08	***

Table 4: The LMEM for total reading time on the ST (*TrtS*)

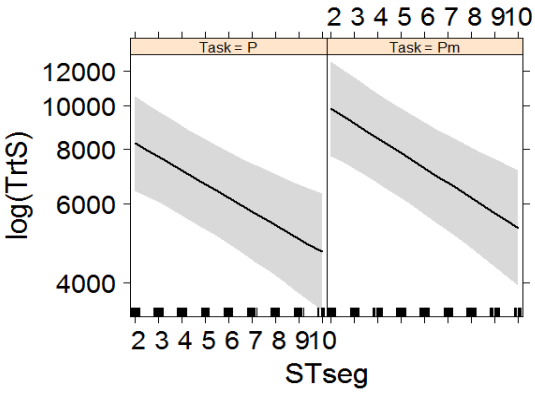

Figure 3: Interaction between Task and *STseg* for total reading time on the ST (*TrtS*)

The number of deletions per segment (*del*) had a relatively large, positive and highly significant effect (see Table 5), as did *Nedit*. Task had a small negative effect on *Perc_TrtT*, such that in the mixed mode participants spent slightly less time reading the TT and more time reading the ST - in proportion (see Figure 4). *STseg* had a relatively large, negative and highly significant effect on

Perc_TrtT. The interaction between Task and *Perc_TrtT* was not significant (see Figure 5).

The effect of STseg on *Perc_TrtT* suggests that the cognitively effortful activity of divided attention between languages (source and target) be-

Predictor	β	SE	t	p	
del	0.13	0.03	3.85	1.55E-04	***
Task Pm	-2.13	1.03	-2.06	0.039	*
STseg	2.05	0.25	8.33	3.55E-15	***
Nedit	1.46	0.59	2.47	0.014	*

Table 5: The LMEM for *Perc_TrtT*

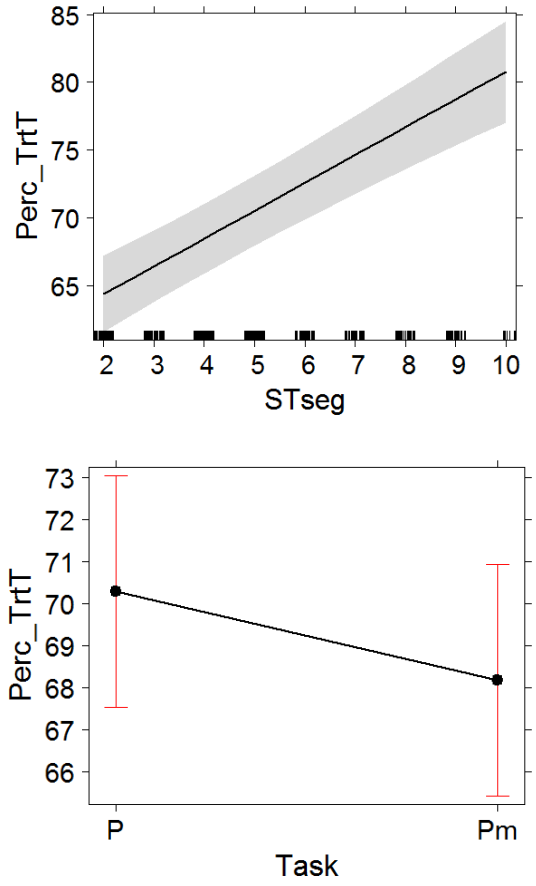

Figure 4: The effect of *Task* and sequential numbering of source segments (*STseg*) on the percentage participants spend reading the TT (*Perc_TrtT*).

comes more centred on the TT (proportionally) as participants progress in the task. This effect is the same in the two modes. Again, what this shows is that the order of the segments in the text and text level coherence more generally plays a negligible role in post-editing - regarding the time spent on the ST and the TT (proportionally). The model for (*Perc_TrtT*) without interaction provided a modest fit (marginal $R^2 = 0.18$, conditional $R^2 = 0.34$).

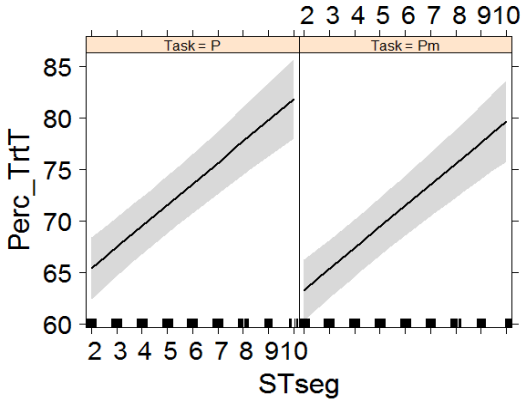

Figure 5:Interaction between Task and *STseg* for (*Perc_TrtT*)

Carl et al. (2016) and similarly Schaeffer et al (2016b) use Activity Units to describe the behaviour during post-editing and from-scratch translation. An Activity Unit slices the data stream of eye movements and keystrokes into 6 types of activity: Either participants read the ST (Type 1), or they read the TT (Type 2), or they produce keystrokes while no eye movements are recorded (Type 4). However, these activities can co-occur: participants may read the ST while (touch) typing (Type 5) or they are reading the TT while typing (Type 6). Finally, if no activity is recorded for more than 2.5 seconds, this is then Type 8. This classification exhaustively slices up the data stream into Activity Units of a certain duration. The duration of Activity Units can be an indicator of how effortful the process is: the longer these activities are, the more effort is required for the particular task such as ST reading (Type 1), TT reading (Type 2) or no recorded activity (Type 8).

The model had the (log transformed) duration of the Activity Unit (*Dur*) as dependent variable and the following predictors Task, Activity Unit Type, the number of fixated words (*PathNmbr*). For Activity Units Type 4 and 8, *PathNmbr* was set to 1, given that these Activity Units do not include any recorded eye movements. And finally the sequential numbering of Activity Units as they occurred (*Id*). *Id* is similar to *STseg* in the previous analyses,

in that it can show whether there was a task facilitation effect. The random variable was Participant. In addition, we tested for the interaction between Task and Activity Unit Type and Task and *Id*. The reference level for Activity Unit Type was Type 1.

There was a small, but highly significant effect of Task on the duration of Activity units, such that, in the mixed mode, Activity Units were overall about 100ms shorter than in the coherent mode (see Figure 6). This is not a large effect, but it does indicate that the mixed mode was slightly easier for participants than the coherent mode. *PathNmbr* had an expectedly large, positive and highly significant effect. *Id* also showed a task facilitation effect for the duration of Activity Units - it was relatively large, negative and significant. Neither of the interactions were significant. The model without interactions provided a relatively good fit (marginal $R^2 = 0.46$, conditional $R^2 = 0.47$).

Predictor	β	SE	t	p	
TaskPm	-0.06	0.01	-4.02	5.95E-05	***
PathNmbr	0.11	0.00	73.28	2.00E-16	***
Type2	0.37	0.02	22.38	2.00E-16	***
Type4	0.81	0.17	4.89	1.02E-06	***
Type5	0.04	0.04	0.95	0.344	
Type6	0.57	0.02	26.01	2.00E-16	***
Type8	1.68	0.10	16.43	2.00E-16	***
Id	-0.0004	0.0001	-3.19	0.001	**

Table 6: LMEM for the duration of Activity Units.

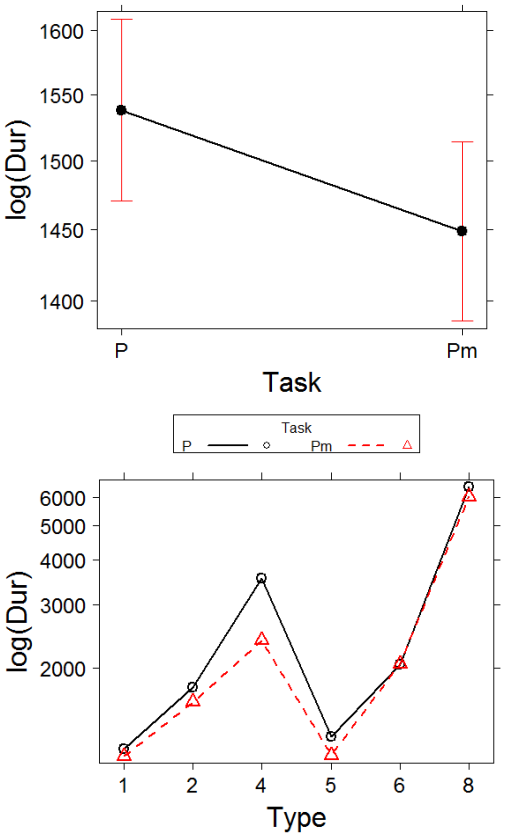

Figure 6:The effect of Task on Activity Unit Duration and the interaction between Activity Unit Type and Task.

In sum, we can say that, in terms of the duration of Activity Units, participants behaved generally in a similar way and if at all, the mixed mode was slightly easier for participants than the coherent mode.

For example, Dragsted (2005) and Alves and Vale (2009) argue that longer stretches of continuous activity are actually indicative of less effortful behaviour. However, in both studies, the definition of a unit occurs on the basis of a pause threshold which defines uninterrupted typing of between 1 and 2 (Dragsted 2005) and 1 and 5 seconds (Alves and Vale 2009). However, the above studies cannot describe what happens during the continuous typing activity, which might actually not be continuous, according to our definition of Activity Units. The effect we report here is much smaller (~100ms) and might well fall within the pauses or typing activities and would thus not be captured by the metrics proposed by the above studies.

6 Discussion

This paper reports from an experiment in which participants post-edited two kinds of texts: in the coherent mode, participants post-edited 2 short texts which were presented as a whole each and in which the order of segments was unaltered. In the mixed mode, participants saw 2 texts which had their segments mixed up and in addition, the order of the segments was jumbled. In the mixed mode, it would have been arguably difficult to generate a text level coherence model, given that the order was jumbled and two texts with rather different topics were jumbled. Surprisingly, this had little or no effect on behaviour.

Maybe most surprisingly, the sequential numbering of segments and of Activity Units had a negative effect on typing duration for both modes. The same was true for the reading times on the ST: participants spent less time on reading the ST the

closer they came towards the end of the text - irrespective of whether the segments were presented unaltered or in the mixed mode. Participants spent more time overall reading the ST in the mixed mode than in the coherent, unaltered mode. This was the only instance of an effect which it could be interpreted as a negative consequence of the lack of text level coherence. However, the effect was small (about 1 second per segment). The proportion of time participants read the TT increased as they progressed in the task and this was again true irrespective of whether segments were coherent or jumbled. Activity Units describe minimal types of activity: ST reading, TT reading, TT typing, a combination of the latter and pauses (no recorded eye movements or keystrokes when participants maybe look away from the screen). The duration of Activity Units can be seen as an indicator of cognitive effort - the longer they last, the higher the cognitive effort. Interestingly, participants had overall slightly shorter Activity Units (about 100ms) in the mixed mode. What all these results suggest is that the mixed mode is not detrimental or cognitively more demanding and rather beneficial or equivalent to coherent mode. These results are promising given that presenting segments in an order which differs from how the ST presents the segments makes it possible to (also) present the ST in a different order from the original one, rather than (only) as a coherent text and this, in turn, makes it possible to fully exploit the potential in crowdsourcing by splitting up coherent texts into units that can be translated and edited independently by a large number of translators at the same time.

However, it has to be borne in mind, that the texts used in this study were very small, due to the limitations as dictated by the recording instruments. Professional translators typically translate texts which are much longer, more specialized and with a whole range of tools. A further limitation to our study is that we did not (yet) examine the quality of the TTs before and after post-editing in the two modes. This is a crucial aspect with important ramifications. Despite these limitations it is appropriate to find these findings encouraging - given the sensitivity of the metrics and the broadly positive results: they are positive against the backdrop of arguments brought forward by those who argue against decontextualization, such as Pym (2011) who points out that technology disrupts linearity in texts, because they are segmented and broken into smaller units. The disruption of text's linearity is inherent to both translation memories (TMs) and machine translation (MT) as TMs and MT segments tend to be sentences or sentence-like units. Consequently, working with a text at a sentence level makes it very complicated to provide "an accurate and fluent translation that adheres to the cohesive and contextual norms of the target language, where, for instance, common linguistic devices of cohesion such as anaphora and cataphora typically function at the paragraph and document level" (Doherty, 2016: 954). Although the longer reading times on the source text in the mixed mode may be indicative of the search for coherence in the ST it has to be borne in mind that this effect was very modest in size (~ 1sec per sentence). Serious disruption would have left a much stronger trace in the behavioral data. Participants spent more time on the TT (proportionally) as they progressed in the task (irrespective of mode), i.e., a shift of attention away from the ST to the TT occurred. This suggests a shift from comprehension to production. The opposite would be a clear indicator of disruption. This was not the case.

It remains to be seen what happens if texts are broken down into sub-sentence units and how this affects behavior, quality and productivity. Again, the results presented here are not discouraging.

References

Fabio Alves and José Luiz Gonçalves. 2013. Investigating the conceptual-procedural distinction in the translation process. A relevance-theoretic analysis of micro and macro translation units. *Target: International Journal on Translation Studies*, 25(1), pages 107–124. https://doi.org/10.1075/target.25.1.09alv

Fabio Alves and Daniel Couto Vale. 2009. Probing the unit of translation in time: Aspects of the design and development of a web application for storing, annotating, and querying translation process data. *Across Languages and Cultures: A Multidisciplinary Journal for Translation and Interpreting Studies*, 10(2), pages 251–273. https://doi.org/10.1556/Acr.10.2009.2.5

R. Harald Baayen. 2013. languageR: Data sets and Functions with R. Analyzing Linguistic Data: A Practical Introduction to Statistics. *R package version 1.4.1*

Kamil Bartoń. 2009. MuMIn: Multi-Model Inference. *R Package version 1.15.6*.

Douglas M. Bates, Martin Maechler, Ben Bolker, and Steven Walker. 2014. {lme4}: Linear mixed-effects models using Eigen and S4. *R package version 1.0-6*

Peter F. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1988. A Statistical Approach to Language Translation. In *Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*. pages 71–76.

Michael Carl and Martin Kay. 2011. Gazing and Typing Activities during Translation: A Comparative Study of Translation Units of Professional and Student Translators. *Meta: Translators' Journal* 56 (4), pages 952–75. https://doi.org/10.7202/1011262ar.

Michael Carl. 2012. Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research. In *The Eighth International Conference on Language Resources and Evaluation. 21-27 May 2012, Istanbul, Tyrkiet*. Department of International Language Studies and Computational Linguistics, pages 2–6

Michael Carl, Moritz J. Schaeffer and Srinivas Bangalore. 2016. The CRITT Translation Process Research Database. In Michael Carl, Srinivas Bangalore and Moritz J. Schaeffer (eds.), *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. Springer International Publishing, Cham, Heidelberg, New York, Dordrecht, London, pages 13–54.

John C. Catford. 1965. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*, Oxford University Press, Oxford.

Stephen Doherty. 2016. The Impact of Translation Technologies on the Process and Product of Translation. *International Journal of Communication* 10, pages 947–69.

Barbara Dragsted. 2005. Segmentation in Translation: Differences across Levels of Expertise and Difficulty. *Target: International Journal on Translation Studies* 17 (1), pages 49–70. https://doi.org/10.1075/target.17.1.04dra

Barbara Dragsted. 2006: Computer-aided translation as a distributed cognitive task. *Pragmatics & Cognition* 14(2), pages 443-464. https://doi.org/10.1075/pc.14.2.17dra

Barbara Dragsted. 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*. Copenhagen Business School, Copenhagen.

Arnt L. Jakobsen. 2011. Tracking Translators' Keystrokes and Eye Movements with Translog. In Cecilia Alvstad, Adelina Hild, and Elisabet Tiselius (eds.), *Methods and Strategies of Process Research. Integrative Approaches in Translation Studies*. John Benjamins, Amsterdam and Philadelphia, pages 37-55.

Filip Klubička, Antonio M. Toral, and Víctor Sánchez-Cartagenac. 2017. Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, (108), pages 121–132, https://doi.org/10.1515/pralin-2017-0014

Alexandra Kuznetsova, Rune Haubo Bojesen Christensen and Per Bruun Brockhoff. 2014. lmertest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (lmer Objects of lme4 Package). *R package version 2.0-6*.

Pascual Martínez-Gómez, German Sanchis-Trilles, and Francisco Casacuberta. 2012. Online adaptation strategies for statistical machine translation in post-editing scenarios. *Pattern Recognition*, 45(9), pages 3193–3203. https://doi.org/10.1016/j.patcog.2012.01.011

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2, pages 346–351. http://www.aclweb.org/anthology/P13-2062

Daniel Ortiz-Martínez. 2016. Online Learning for Statistical Machine Translation. *Computational Linguistics*, 42(1), pages 121–161. http://dx.doi.org/10.1162/COLI_a_00244

Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. 2016. Interactive neural machine translation. *Computer Speech & Language*. http://dx.doi.org/10.1016/j.csl.2016.12.003

Anthony Pym. 2011. What Technology Does to Translating. *Translation and Interpreting Research* 3 (1), pages 1–9.

R Development Core Team, 2014. *R: A language and environment for statistical computing*, Vienna, Austria.

Lucia Specia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the 15th conference of the European Association for Machine Translation*. Leuven, Belgium, pages 73–80.

Antonio M. Toral. 2017. Neural and Phrase-based Machine Translation Quality for Literary Texts. (Book chapter currently under review).

Jean-Paul Vinay and Jean Darbelnet. 1995. *Comparative stylistics of French and English: a methodology for translation*, John Benjamins, Amsterdam and Philadelphia.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv e-prints*, pages1-23. http://arxiv.org/abs/1609.08144