Measuring Cognitive Translation Effort with Activity Units

Moritz SCHAEFFER[1], Michael CARL[2,4], Isabel LACRUZ[3], Akiko AIZAWA[4]

[1] Johannes Gutenberg University Mainz, An der Hochschule 2 76726 Germersheim, Germany
[2] Copenhagen Business School, Dalgas Have 15, 2000 Frederiksberg, Denmark
[3] Kent State University, 475 Janik Drive, Kent, Ohio 44242
[4] National Insitute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-0003, Japan

mschaeffer@uni-mainz.de, mc.ibc@cbs.dk, ilacruz@kent.edu,
aizawa@nii.ac.jp

Abstract

Despite the increased quality of Machine Translation output, human interaction will remain a crucial activity to guarantee the quality of the final translation products. Human-computer interaction in translation will likely be the more successful the more we understand the properties and complementarities of both partners. This paper traces cognitive approaches in machine translation back to the mid-1980s and argues that we now have the technologies available that will allow us to eventually arrive at an in-depth understanding of the human translation processes. It illustrates some of the research methods in empirical translation process research and suggests ngrams of Activity Units for measuring the translation process.

1. Introduction

As a reaction to the then predominant rule-based translation paradigm, Nagao (1984) suggests a cognitive approach to translation, which mimics the human translation process. He states that "[m]an does not translate a simple sentence by doing deep linguistic analysis, rather, [...] first, by properly decomposing an input sentence into certain fragmental phrases ..., then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence." Based on this model, a large number of different example-based MT (EBMT) systems have been developed, which Carl and Way (2003) classify into:

- Pure EBMT: All translation relevant processing takes place at run-time
- Generalizing EBMT: preprocessing of translation templates
- Tree-based EBMT: preprocessing of dependency and phrase-structure trees

As a simulation of the human translation process, and in line with many earlier models of the human translation process (e.g. Nida, 1964), these systems assume that translators proceed:

- Sequentially, in a sentence-by-sentence (or phrase-by-phrase) mode
- Stratificationally, by decomposing – transferring – recomposing

Figure 1 visualises sequences of activities as hypothesized by Nagao's model as successive activities of reading "certain fragmental phrases" of source text words, followed by typing of the corresponding string of TT words. Figure 1 shows such a scenario on the basis of artificial data. The translation progression graph plots the English source text ("the awareness of other hospital staff put a stop") on the left vertical axis and its Spanish translation ("la atencion de otros empleados del hospital puso fin") on the right side. The horizontal axis represents a time line on which the translator's activities, such as keystrokes (black), fixations on words in the source text (blue dots) and words in the target text (green diamonds) are plotted. Figure 1 shows an almost linear fragmented translation production with a 1-to-2 alignment (hospital → del hospital) and a 2-to-1 translation alignment (a stop → fin) and a syntactic re-ordering (hospital staff → empleados del hospital).
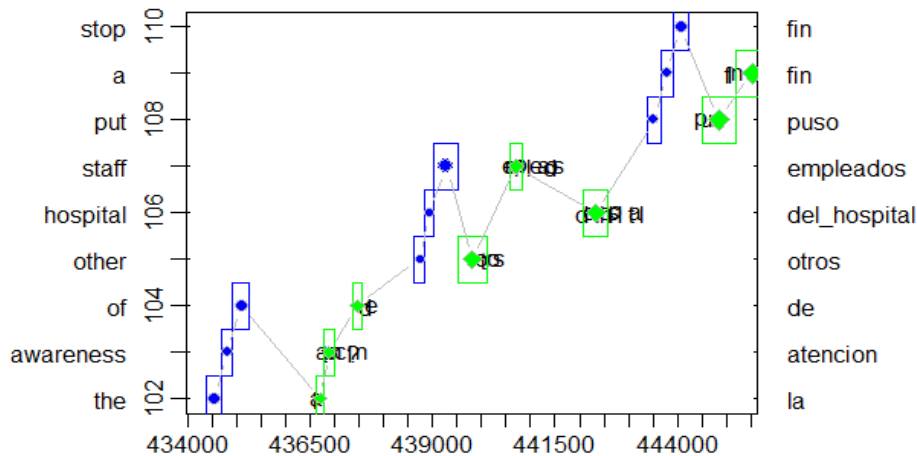

Figure 1. Visualisation of an artificial scanpath exemplifying Nida's / Nagao's sequential / stratificational model of the human translation process.

Translation process research in the past years suggests that translation processes only exceptionally take place in such a linear manner, whereas often iterative and concurrent processes can be observed. However none of the EBMT systems, and for that matter of MT systems in general, are based on actual empirical investigations of the human translation process. The sequential and stratificational nature of the (human) translation and post-editing processes is just taken for granted.

In this paper we present novel methods to analyse empirical translation process data which suggests a more complex picture of the translation process. We look at pauses during typing activity, which have been taken as indicators of cognitive effort in translation and post-editing (Jakobsen, 1998, O'Brien, 2006). On a large data set which

was collected in various studies over the past 8 years[1], we compute the pause-word ratio (PWR, Lacruz et al. 2014) which aims at measuring translation difficulties based on the pausing and typing structure. We examine 2101 English → Danish, Spanish, German, Chinese and Japanese translations segments and 1783 English → Spanish, German, Chinese and Japanese post-editing segments, which are extracted from the KTHJ08, BML12, SG12, MS12, ENJA15 and NJ12 studies in the TPR-DB which were produced by 147 different translators with different degrees of translation expertise.

We show that the PWR correlates with a gaze based translation difficult index (TDI) introduced by Mishra et al (2013). Then we suggest Activity Units as a means to analyse in more detail gazing behaviour during text production pauses.

2. The Pause-Word Ratio

A central question in translation and in MT post-editing is related to measuring the cognitive effort involved in the translation production. Lacruz et al. (2014) observe that post-editors proceed in sequences of typing activities interrupted by relatively short pauses, between 300 ms and 2,000ms. Consequently, these authors develop the pause-to-word ratio (PWR), based on the assumption that pauses between keystrokes provide information on cognitive effort, where higher PWR values are associated with more cognitive effort.
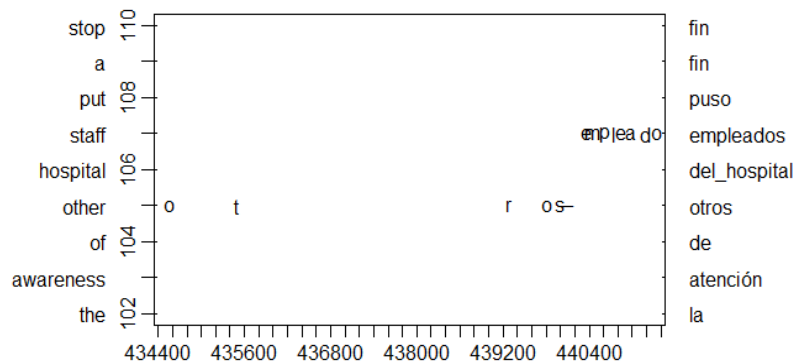

Figure 2. Translation Keystrokes in time

Figure 2 plots a translation progression graph in which a translator produces the translation of English "the awareness of other hospital staff" into Spanish "la atención de otros empleados del hospital"; it shows the distribution of the keystrokes within a

---

[1] The data can be downloaded free of charge from the CRITT TPR-DB the TPR-DB:
https://sites.google.com/site/centretranslationinnovation/tpr-db.
Carl et al (2016) provides a comprehensive introduction to the database and its features.

sequence of approximately 6.7 seconds in which the Spanish translation "otros empleados" (other staff) is typed. There are three longer pauses of approximately 940 ms between the typing of "o" and "t", a pause of 3.8 seconds between the typing of "t" and "r" and a pause of 530 ms between typing "r" and "os empleados". All other inter-keystrokes intervals amount to less than 300 ms.

Sequences of successive keystrokes can be fragmented into segments of coherent text production (i.e. typing bursts), where each production contains at least one keystroke.

The length and duration of a production P depends on the gap of time G between two successive keystrokes: a Production Unit P_i ends if the gap between the last keystroke in P_i and the first keystroke in P_{i+1} exceeds a pre-defined threshold T. The keystrokes of an entire translation session are thus fragmented into successive gaps G_i and typing bursts P_i.

The session initial first gap G_1 can be shorter than the threshold T. A production P contributes to the production of one target text segment TS, and each keystroke K_k in P contributes to exactly one word within the TS. If two successive keystrokes K_k and K_{k+1} are part of two segments TS_t and TS_{t+1}, a unit boundary is inserted so that K_k is the last keystroke in P_i and K_{k+1} is the first keystroke in P_{i+1} even if the gap G_{k+1} between K_k and K_{k+1} is shorter than the threshold T. P_i will be part for the production of segment TS_t and P_{i+1} will be considered part for the production of segment TS_{t+1}. The production of a segment TS is thus made up of p units $\{G_1 , P_1\} ... \{G_p , P_p\}$. The revision of a segment can lead to non-consecutive typing activities. For instance, a first draft of a segment TS_t may be produced with units $\{G_1 , P_1\} ... \{G_m , P_m\}$ and a revision of that segment is done at a later stage with units $\{G_{n+1} , P_{n+1}\} ... \{G_p , P_p\}$ while units $\{G_{m+1} , P_{m+1}\} ... \{G_n , P_n\}$ are part of one or more other segments. The segment TS_t is thus edited twice. The following features are thus computed for each segment TS:

- *SegInitialGap*: duration of the first gap durations G_1 in TS
- *ProductionGap:* sum of the gap durations G_2 to gap G_p
- *ProductionDur*: sum of the production durations $P_1 ... P_p$
- *Duration:* production duration of the segment (sum of P and G)
- *ProductionNum*: number 'p' of units $\{G,P\}$ in TS
- *Nedit*: number of revisions of the segment

The above definitions lead to the following equation:

$$Duration = SegInitialGap + ProductionDur + ProductionGap$$

In addition, the *ProductionDur* and the *ProductionGap* are segmented differently according to a thresholds T. Five different thresholds are currently implemented in the TPR-DB, which lead to different values of the **dur*, **gap*, and **num* features, as shown in the Table 1.

Different thresholds segment the keystroke data in different ways, but the translation duration remains the same. That is, the sum of *ProductionDuration + ProductionGap* is identical irrespectively of the thresholds. Thus:

$$Mdur+Mgap = Sdur+Sgap = Pdur+Pgap = Ldur+Lgap = Kdur+Kgap$$

Table 1. Features for different thresholds in the TPR-DB

Unit type	Threshold (T)	Production *dur, gap, num*	PWR
M-Unit	300ms	Mdur, Mgap, Mnum	PWR300 = Mgap / #ST tokens
S-Unit	500ms	Sdur, Sgap, Snum	PWR500 = Sgap / #ST tokens
P-Unit	1000ms	Pdur, Pgap, Pnum	PWR1000 = Pgap / #ST tokens
L-Unit	2000ms	Ldur, Lgap, Lnum	PWR2000 = Lgap / #ST tokens
K-Unit	5000ms	Kdur, Kgap, Knum	PWR5000 = Kgap / #ST tokens

3. The translation difficulty index (TDI)

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 TDI score of unseen data during translation on the basis of the linguistic features.
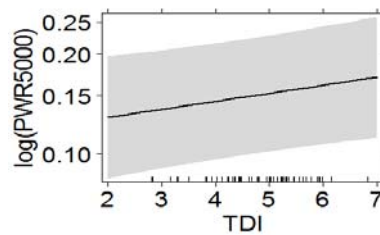
Figure 3. Plot of the effect of TDI on (log) PWR

The prediction was relatively accurate (MSE = 12.88), but to our knowledge, the TDI measure has not been tested on new data. In the next section we test the TDI score for its prediction of the PWR score.

4. Correlation of TDI and PWR

The data used in Mishra et al (2013) were 80 English source sentences translated into Spanish, Danish and Hindi. For the current purpose we use 1476 English sentences translated into Spanish, Danish, Hindi, Chinese, German and Japanese which is a subset

of the 2101 segments mentioned in the itroduction (segments with PWR=0 were excluded). These segments include the 80 sentences used by Mishra et al.

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, Christensen, and Brockhoff, 2014) was used.

The LMEM for experiment 1 had the following random variables: item, participant, text and target language. PWR was log transformed because it was not normally distributed. The predictor was the TDI score. TDI had a significant (β=0.056, SE=0.026, t=2.137, p< .05) positive effect on the PWR with a pause threshold of 5000ms (see *Table* 1).

This result validates the TDI as a good predictor of translation effort and it also validates PWR as a good measure of effort during translation and post-editing (Lacruz et al 2014). However, it does not explain what exactly happens during the pauses.

5. Activity Units

The PWR analysis does not explain what happens within the pauses and the TDI score does not differentiate when a translator is looking where on the source or target texts. As
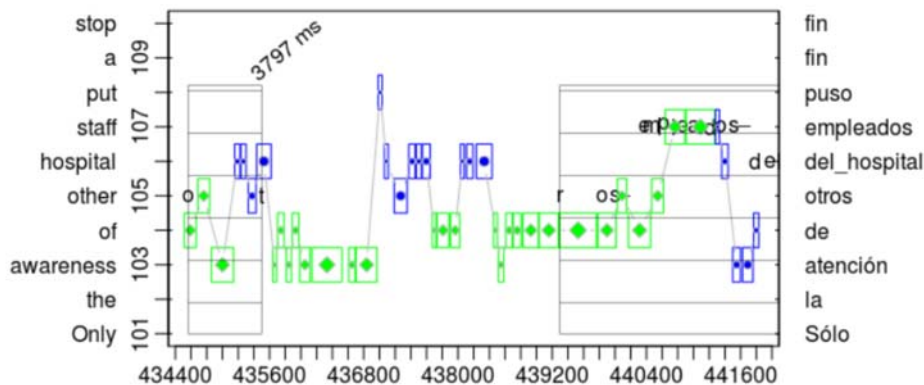

Figure 4. Translation segment of Fig. 1 with keystrokes and fixation data

eye movements are independent from finger movements, there exist several kinds of sequential or concurrent gazing and typing activities which – as we will show - carry important information. As an illustration consider Figure 4 which is a replication of Figure 2 and which shows two typing bursts "ot" and "ros empleados" (in dashed boxes, as generated by the 1000ms threshold). The two Production Units are separated by a 3.8 sec typing pause. The translation progression graph shows, in addition, the gaze behaviour as the eyes move back and forth between the source text (blue dots) and the target window (green diamonds). While the PWR takes into account the pausing

information, TDI would predict the sum of the gaze durations. However, both measures ignore how the eyes move across the texts and how the eyes and hands are coordinated: there are concurrent gaze activities on the source or the target text during text production and sequential stretches of source or target reading during the typing pauses, of different length and location. For instance, when typing "ot", the translator monitors the production of "o" in the target text and gazes at the source text while typing "t".
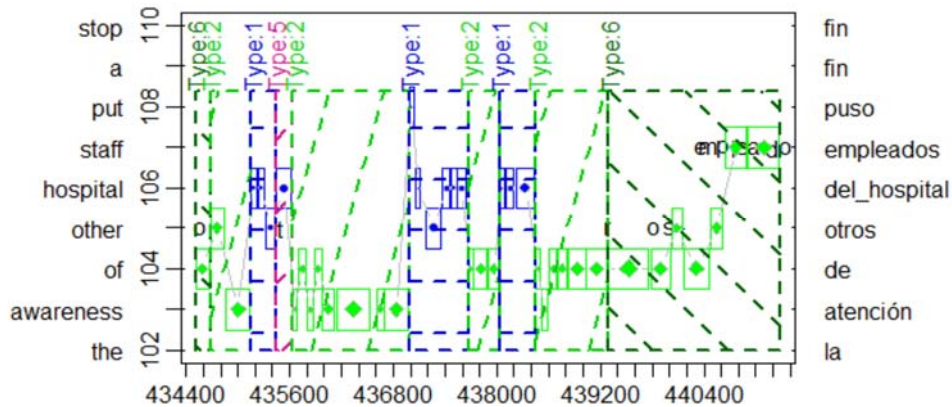

Figure 5. Translation segment of Fig. 4 segmented into ten successive Activity Units

In contrast to the PWR and TDI metrics, Activity Units take this information into account. Similar to (Carl et al, 2016) we make a distinction between 6 different basic types of activities[2]:

- type 1: ST reading
- type 2: TT reading
- type 4: translation typing (no gaze data recorded)
- type 5: ST reading and typing (touch typing)
- type 6: TT reading and typing (translation monitoring)
- type 8: no gaze or typing activity recorded for more than 5 seconds

Figure 5 shows the segmentation of the same data into Activity Units. As can be seen, taking into account the gaze activities allows for a much finer-grained fragmentation of the data into the six types of activity. The first activity in Figure 5 consists of concurrent TT reading (gaze on the translation of word 104) while typing "o", the first letter of the translation for ST word 105 (type 6). In the following typing pause of 940 ms, the eyes first hover over the beginning of the translation of ST words 105 (i.e. the just typed "o") and the translation of word 103 (atención).

---

[2] The Activity Unit of type 7, as suggested in (Carl et al, 2016), which entails concurrent type 1, 2 and 4 behaviour is not assumed here. Instead the activities were split into the six types above.

Table 2. Features which describe the ten Activity Units from Fig. 5

Time	Type	Dur	nKey	nFix	DFix	ScSpan	Turn	SD
434515	6	172	1	1	1	0	0	0
434687	2	469	0	2	2	2	0	1.41
435156	1	281	0	3	2	1	0	0.58
435437	5	203	1	1	1	0	0	0
435640	2	1344	0	8	2	1	3	0.5
436984	1	687	0	6	3	3	1	0.91
437671	2	360	0	3	1	0	0	0
438031	1	422	0	3	1	0	0	0
438453	2	843	0	6	2	1	1	0.41
439296	6	1985	11	7	3	2	2	0.85

This TT reading activity (type 2) is followed by ST reading (type 1) of word 106 (hospital) and 105 (other). The gaze then remains on the source text while "t" is typed which is an activity of type 5. After this, the long pause of 3.7 seconds is structured into five alternating Activity Units of type 2-1-2-1-2 in which the translator seems to develop

Table 3. Correlation matrix for scanpath measures

	nFix	DFix	ScSpan	Turn	SD
Dur	0.71	0.68	0.38	0.72	0.26
nFix		0.83	0.40	0.88	0.25
DFix			0.57	0.87	0.39
ScSpan				0.42	0.93
Turn					0.26

and check a translation strategy for the following noun phrase before, in an activity of concurrent typing and TT monitoring (type 6), "ros empleados" is produced without much hesitation. For each Activity Unit, a number of features are extracted, most of which relate to gaze patterns:

- *Dur*: is the duration in ms of the Activity Unit.
- *nKey*: count of the keystrokes during the Activity Unit.
- *nFix*: number of fixations in the Activity Unit.
- *DFix*: number of different words fixated.
- *ScSpan*: span of the gaze path calculated as difference between largest and smallest word
- *Turn*: moves from regression to progression or vice versa.
- *SD*: standard deviation of median word fixations.

The features *DFix*, *ScSpan*, *Turn* and *SD* describe the linearity of the sequence of fixations in an Activity Unit for which eye movements have been recorded (type 1,2,5 or 6). If *DFix* equals *nFix*, then each word within an Activity Unit has only been fixated once. A large *ScSpan* (scanpath span) occurs if long stretches of text are read, and when the distance between fixated words is large, as indexed by the sequential numbering of words. *SD* is 0 if only one word is fixated. It becomes larger the more words are fixated far away from the median fixated word in a scanpath.

Table 2 shows the extracted features for the 10 units in Figure 5. It shows the starting time and the duration of each of the Activity Units together with the six extracted features.

Table 4. Frequency distributions of bigrams

Bigram	Freq
1_2	0.45
1_5	0.34
1_6	0.20
Total	**0.99**

Correlations between *Dur*, *nFix*, *DFix*, *ScSpan*, Turn and *SD* are relatively high – the longer a coherent reading activity is (*Dur*), the more likely it is that more different words are fixated (*DFix*), resulting in a larger *ScSpan*. It is also more likely that progressions and regressions occur (higher Turn) (see Table 3).

6. Sequences of Activity Units

Martínez-Gómez et al. (2014) cluster sequences of translation events and find that concurrent reading and translation typing correlates with translator experience; i.e. experienced translators are better able to distribute their attention on ST reading and TT production. We investigate ngrams (bigrams and trigrams) of Activity Units to assess fluent translation production and the typical gazing patterns during effortful translation production.
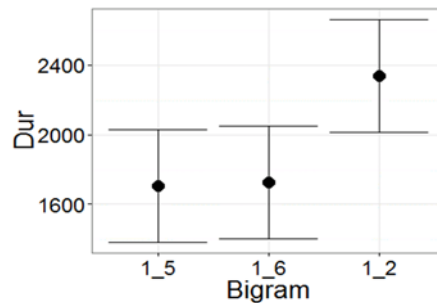
Figure 6. The effect of ST reading duration (Type 1) on subsequent activity (Type 2, 5, 6)

The same data as the one used above was used in this experiment. However, rather than using complete segments, we extract all bigrams and trigrams of Activity Units from the translation drafting phase (Jakobsen 2002) which start with type 1 (ST reading). This resulted in 30,803 Activity Unit bigrams. The three bigrams shown in Table 4 made up more than 99% of the data. Other bigrams (1_4 and 1_8) had a frequency of less than 0.01 and were excluded.

The LMEM for experiment 2 had the following random variables: participant, text and target language. The predictor was the ngram and all of the scanpath measures described above in Table 3 and as measured on the Activity Units of type 1, were entered as dependent variables of the LMEMs. Sliding contrasts were applied to the bigrams. As all scanpath measures correlate, results for the other measures were very similar in size and significance.

There is no significant difference in the duration of Activity Units type 1 followed by either type 5 or 6. However, the difference in *Dur* between bigrams 1_2 and 1_6 was highly significant.

Similarly, the standard deviation (*SD*) from the median word fixation within a scanpath in an Activity Unit of type 1 is lower if the subseqent activity consists of typing. The difference in the *SD* value of Activity Units of type 1 followed by type 5 as compared to those followed by type 6 was relatively small, but highly significant (β = 0.61, SE= 0.14, *t*= 4.28, p< .001). However, the difference between the *SD* of Activity Units of type 1 followed by those of type 2 was much larger as compared to those which were followed by Activity Units of type 6 (β = *1*.66, SE= 0.14, *t*= 11.99, p< .001).

This suggests that the longer and the less linear a scanpath on the ST is, the more likely it is that translators will read the TT immediately after reading the ST. And conversely, the shorter and the more linear a scan path on the ST the more likely it is that translators will start typing immediately after reading the ST.

7. Trigrams of Activity Units

The duration and scanpath linearity of Activity Unit 1 has only an effect on the subsequent activity but not on later units: Trigrams showed a very similar picture to those in bigrams. The frequency distribution of trigrams is shown in Table 5 (trigrams

Table 5: Frequency distributions of Activity Unit trigrams in translation

Trigram	Freq
1_2_1	0.24
1_5_1	0.21
1_2_6	0.18
1_6_2	0.12
1_5_6	0.07
Total	**0.82**

which occurred with a frequency of 4% or less were excluded). Sliding contrasts were applied to the trigrams. There was no significant difference in the SD of Activity Units type 1 for trigrams 1_5_1 and 1_5_6. The difference between trigrams 1_5_2 and 1_5_1 was not significant either. The difference between trigrams 1_6_2 and 1_5_2 was significant ($\beta = 0.83$, SE= 0.28, $t= 2.97$, p< .01). The difference between trigrams 1_2_6 and 1_6_2 was highly significant and relatively large ($\beta = 1.41$, SE = 0.19, $t=7.53$, p< .001). The difference between trigrams 1_2_1 and 1_2_6 was also highly significant, but smaller ($\beta =0.64$, SE = 0.16, $t=4.06$, p< .001). The effect on *Dur* and all other scanpath measures was similar in size and significance, but, more importantly: trigrams containing typing activity in the second place were associated with more linear scanpaths of the initial activity type 1.

Properties of Activity Units of type 1 is therefore restricted to the immediately following Activity Units but has no effect on later activities.
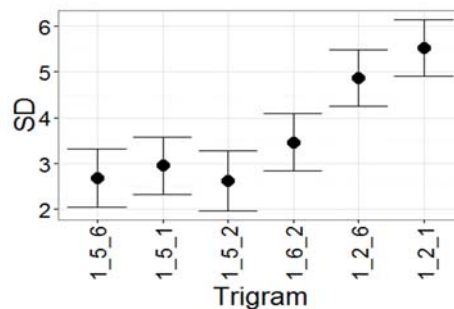

Figure 7. The effect of fixations *SD* during Type 1 activity on next tro subsequent Activity Units

8. PWR and Activity Units in Post- editing

While the previous sections looked at translational behaviour in from-scratch translation, in this section we look into post-editing behaviour. For this investigation we examine all 2101 translations segments and 1783 post-editing segments, as described in the introduction.
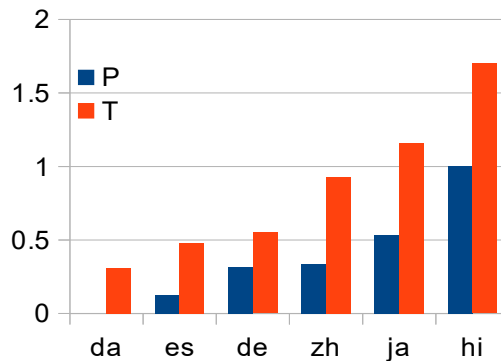

Figure 8. PWR scores for post-editing (P)
and from-scratch translation (T) for different language pairs

1. PWR scores are generally lower during post-editing than during from-scratch translation, indicating a lower cognitive load according to the PWR metric. Figure 8 plots PWR1000 scores, it suggests that the European languages (Danish, Spanish, German) are easier to translate from English than the Asian languages (Chinese, Japanese, Hindi), although the graphs do not take into account the expertise of the translators.

Table 6. Frequency distribution of Activity Unit following Type 1 for post-editing and translation

Bigram	Freq-P	Freq-T
1_2	0.81	0.45
1_5	0.10	0.34
1_6	0.08	0.20
Total	**0.99**	**0.99**

2. The transitions between Activity Units is differently distributed for translation and post-editing. As shown in Table 6, 81% of all the outgoing transitions from 2.

Activity Unit of type 1 during post-editing go to unit of type 2. That is almost twice as much as during translation (see Table 4, which is reproduced for conveniance in Table 6 as column Freq-T). Only in 18% of the cases do post-editors engage in a typing activity after ST reading, as compared to 54% in the translation condition. This suggest that much more reading back-and forth between ST and TT takes place during post-editing than during translation.
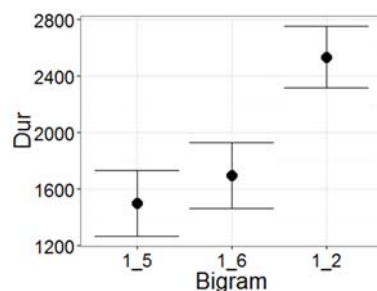

Figure 9. The effect of ST reading duration (Type 1)
on the subsequent Activity Units during post-editing

3. However, it is interesting to note that the relationship between the duration of the ST scanpath and subsequent Activity Units (typing or TT reading) is similar in both translation and post-editing. After an ST reading activity of around 2.5 seconds, translators as well as post-editors are likely to switch to TT reading (Type 2) while for shorter ST reading activities of about 1,5 to 2 seconds, both translators and post-editors are more likely to engage in TT typing.

9. Conclusion

Much of recent investigations in translation process research has sought to identify characteristics of either the ST or the TT in order to predict how difficult it is to process or produce a string of words during translation or post-editing (e.g. O'Brien 2005, Dragsted 2012, Alves and Gonçalves 2013, Schaeffer et al 2016). Lacruz et al (2014) take a different approach and associate pauses in the translation production process (i.e. gaps in typing activities) with cognitive meta-activity. In this paper we investigate the gazing patterns that take place during keystroke pauses to assess what happens during typing pauses that makes translation more or less effortful. Recorded gaze data fills the typing pauses and make it possible to "identify the specific motivation of a particular pause" (Kumpulainen, 2015: p 47). Our research shows that translation processes are much less sequential, (sentence-by-sentence, chunk-by-chunk) and much less stratificational than predicted by earlier translation models (e.g. Nagao, 1984). Rather, we observe iterative translation and revision processes and gazing patterns which indicate earlier and later translation processes (Schaeffer et al, 2016). While there is a substantially higher amount of iteration from ST reading to TT reading during post-editing than during translation, the basic behavioural pattern seems to be similar:

- after a longer and non-sequential scanpath on the ST → translators and post-editors are likely to read the TT
- after a short, sequential and linear scanpath on the ST → translators and post-editors are likely start typing
- properties of the ST reading behaviour only has an impact on the immediately following Activity Unit

References

Alves, F., and Gonçalves, J. L. (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), 107–124.
`http://doi.org/10.1075/target.25.1.09alv`

Baayen, R. H. (2013). "languageR: Data sets and Functions with "Analyzing Linguistic Data: A Practical Introduction to Statistics". Retrieved from
`http://cran.r-project.org/package=languageR`

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). {lme4}: Linear mixed-effects models using Eigen and S4. Retrieved from `http://cran.r-project.org/package=lme4`

Carl, M., Dragsted, B., Jakobsen, A.-L. (2011) A Taxonomy of Human Translation Styles. In: Translation Journal, Vol. 16, Nr. 2, 2011

Carl, M., Schaeffer M., Bangalore, S. (2016). The CRITT Translation Process Research Database. In: New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB. . ed. Michael Carl; Srinivas Bangalore; Moritz Schaeffer. Cham : Springer, p. 13-54

Carl, M., Way, A. (Eds.). (2003). Recent Advances in Example-Based Machine Translation. Text, Speech and Language Technology, Springer Netherlands

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

Immonen, S. (2006): Translation as a Writing Process: Pauses in Translation versus Monolingual Text Production. In Target 18 (2), 313-335

Jakobsen, A-L. (1998). Logging Time Delay in Translation, LSP Texts and the Translation Process. Copenhagen Working Papers. 73 − 101.

Jakobsen, A-L. (2002). "Translation Drafting by Professional Translators and by Translation Students." In Empirical Translation Studies. Process and product. Edited by Gyde Hansen, 191-204. Copenhagen: Samfundslitteratur

Kumpulainen, M. (2015). On the operationalisation of 'pauses' in translation process research. Translation and Interpreting Vol 7 No 1 (2015)
`http://trans-int.org/index.php/transint/article/download/367/183`

Kuznetsova, A., Christensen, R. H. B., Brockhoff, P. B. (2014). lmertest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (lmer Objects of lme4 Package). R package version 2.0-6.
Retrieved from `http://www.cran.rproject.org/package=lmerTest/`

Lacruz, I., Shreve, G. M. (2014). Pauses and Cognitive Effort in Post-Editing.. In Post-editing of Machine Translation: Processes and Applications. Edited by Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia

Martínez-Gómez, P., Minocha, A., Huang, J., Carl, M., Bangalore, S., and Aizawa, A. (2014). Recognition of translator expertise using sequences of fixations and keystrokes. Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '14, 299–302.
`http://doi.org/10.1145/2578153.2578201`

Mishra, A., Bhattacharyya, P., and Carl, M. (2013). Automatically predicting sentence translation difficulty. *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2, 346–351.

Nagao, M. (1984). A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle. In Proc. Of the International NATO Symposium on Artificial and Human Intelligence (pp. 173–180). New York, NY, USA: Elsevier North-Holland, Inc.

Nida, E. A. (1964). Toward a science of translating : with special reference to principles and procedures involved in Bible translating. Leiden: E.J. Brill.

O'Brien, S. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. Machine Translation, 19(1), 37–58. http://doi.org/10.1007/s10590-005-2467-1

O'Brien, S. (2006) *Pauses as indicators of cognitive effort in post-editing machine translation output.* Across Languages and Cultures, 7 (1). pp. 1-21. ISSN 1585-1923

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

Schaeffer, M., Dragsted, B., Kristian Tangsgaard Hvelplund, Laura Winther Balling, Michael Carl. *Word Translation Entropy: Evidence of Early Target Language Activation During Reading for Translation*. In: New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB. . ed. Michael Carl; Srinivas Bangalore; Moritz Schaeffer. Cham : Springer, Pages 183-210.