Comparing spoken and written translation with post-editing in the ENJA15 English → Japanese Translation Corpus

Michael Carl[1] Isabel Lacruz[2] Masaru Yamada[3] Akiko Aizawa[4]
1. National Institute of Informatics and Copenhagen Business School
2 Kent State University
3 Kansai University
4 University of Tokyo and National Institute of Informatics

1 Introduction

Spoken language applications are becoming increasingly operational and are used in many computer applications today. Translation dictation is a mode of translation by which a translator reads a source text and speaks out its translation, instead of typing it. Translation dictation is thus a method of translation situated in between interpretation, where the interpreter hears a text and speaks out the translation (e.g., during conference interpreting) and conventional translation by which a written source text is translated mainly using the keyboard. It is close to sight translation. Translation Dictation was a technique used in some translation bureaus in the 1960s and 1970s (Gingold, 1978) but it has been used less frequently since the mid-80s, as professional translators started using micro-computers (Zapata and Kirkedal, 2015).

Already, the ALPAC report (Pierce et al., 1966) mentioned that "productivity of human translators might be as much as four times higher when dictating" as compared to writing, and with today´s increasing quality of voice recognition this mode of translation is experiencing a come-back. The usage of Automatic Speech Recognition (ASR) systems provides an efficient means to produce texts, and our experiments suggest that for some translators and types of text translations it might become even more efficient than post-editing of machine translation.

In this paper we describe the ENJA15 translation study and corpus. The ENJA15 corpus is a collection of translation process data that was collected in a collaborative effort by CRITT and NII. The ENJA15 data is part of a bigger data set which will enable us to compare human translation production processes across different languages, different translation modes, including from-scratch translation, machine translation post-editing and translation dictation.

2 The TPR-DB multilingual translation corpus

The ENJA15 study is part of a multilingual translation corpus in which six short English texts are translated under various different conditions into a number of different target languages, which so far also include Chinese, Danish, German, Hindi, and Spanish. The goal of the multilingual translation corpus is to gather translators' activity data (text perception and production behaviour, as recorded by keystroke loggers, eye-trackers, etc.) in order to investigate variations in the human translation process across different translator profiles, translation modes and different target languages.

To date, experimental data has been collected from more than 150 different translators in more than 760 translation sessions, which accumulate to more than 110 hours of translation data. Some knowledge has been generated from this corpus which is, among other outlets, reported in an edited volume (Carl et al., 2016).

During each translation experiment, every translator translated 6 short English source texts of approximately 110 - 160 words under different translation conditions, including from-scratch translation and post-editing of machine translation (mostly from google translate) and, in the ENJA15 study, also translation dictation. User activity data (keystrokes, gaze data, spoken translation) were recorded and post-processed as described in Carl et al. (2016). Each translation experiment (consisting of the six text translation sessions) typically takes between 2 and 3 hours. Four of the texts are taken from a news domain and two from a sociology encyclopedia. Translators were advised to produce a 'good enough' translation for publication without spending too much time on terminological or stylistic subtleties (Mesa-Lao, 2014). Translators were told not to use external help (lexica, concordance tool, etc.) during their translations and instead to concentrate only on the screen, since otherwise we would have lost track of their gaze. Translators were also asked to fill out a meta-data form to keep track of their translation experiences (years of formal training, years as active translator, attitude and experience in post-editing, etc.). The translation process was recorded with Translog-II (Carl, 2012), and with an eyetracker. The collected data was anonymized and processed, and is publicly available under a creative commons license in the CRITT TPR-DB.

3 The ENJA15 study

The ENJA15 translation study extends the multilingual translation corpus, adding data for the language pair English → Japanese. As a novelty in ENJA15 experiment, translators spoke their translations in one of the conditions, using an automatic speech recognition (ASR) system, Nuance Naturally Speaking. The ENJA15 translation experiment consists, thus, of three different conditions:1. from-scratch translation (T), 2. translation dictation (D), 3. MT post-editing (P)

Participants translated two texts in each of these conditions, first two from-scratch translations (T), then two texts with translation dictation (D) and finally two texts using post-editing (P). The order of the translation modes remained identical, but the texts were permuted, with the goal of obtaining an equal number of translations of each text in each translation mode. The time needed to complete the translation of six texts was not restricted but usually took between 2 to 3 hours. Participants were remunerated between 4000 and 6000 yen (approx. 30€ and 45€), depending on their experience. Participants were made familiar with the goals of the translation experiment, and they signed a form in which they agreed that their translation data would be made publicly available under a creative commons license. They also out filled two questionnaires, one before starting the translation session and another after having finished. Questionnaire 1 contained questions concerning expertise of the participant, years of translation experience, Questionnaire 2 was to be filled after the experiment and contained questions concerning satisfaction with the three translation modes, and an estimation of the effort used in each of the translation modes.

All translation sessions were recorded with Translog-II and gaze data was recorded with an SMI mobile eyetracker at 60 and 120Hz. For the dictation sessions, the ASR system was trained by each translator prior to performing the translation dictation task. Training took approx. 10 minutes.

As with the other data sets in the CRITT TPR-DB (see Carl et al. 2016), the English → Japanese data collected from the ENJA15 experiment was post-processed: tokenized, sentence aligned, keystrokes were mapped on target words and summary tables were produced, which included: offline gaze-to-word mapping by running the Translog-II replay tool, manual word alignment of the EN → JA using YAWAT browser (Germann, 2008), automatic generation of the TPR-DB

The data was anonymized and added to the publicly available TPR-DB under a creative commons license which can be downloaded free of charge from the CRITT TPR-DB .

4 Participants

All participants had Japanese as their first and English as their second language and reported between 0 and 20 years translator experience. The distribution of translation experience was very uneven: four participants had 15 or more years, while 16 participants had two or fewer years of professional exposure. (We expect to be able to attract more balanced translators during the remaining translation experiments.) Two participants reported using a speech recognition system every day, although not for translation dictation; all others said they had never used one. Approximately half of the participants reported using machine translation for post-editing, with a level of satisfaction on a 5-point Likert scale range from “highly dissatisfied” to “highly satisfied”.

The four participants with long translation experience were either translation teachers or freelancers, while the other participants were third or fourth year language students from Kansai or Kobe University with one or two years of translation training.

5 Preliminary evaluation

A preliminary evaluation of the data was conducted with respect to the productivity of the three translation modes where we found that translation dictation and post-editing are quicker than from-scratch translation. We also investigate the pause structure and properties of the text production units where it was observed that post-editing produces the most scattered typing behavior while during dictation the translations are generated in the most coherent manner. Eye tracking data reveals that the different translation modes imply quite different gazing patterns.

5.1. Translation productivity

Figure 1 shows average translation production times in ms per ST word for the three translation conditions. It shows that from-scratch translation almost always takes more time than both translation dictation and post-editing, on average 7.2 seconds for from-scratch translation (T) and 5.2 seconds for D and P. Only for the fastest from-scratch translators do D and P modes not increase translation speed.

	P	D	T
Deletions	323	256	210
Insertions	293	842	823

However, the translation time does not correlate with the number of produced insertions or deletions: Table 1 shows that in the P condition the number of insertions produced is on average less than half of the number produced in dictation or from-scratch translation, while the number of deletions is highest in post-editing.

Figure 2 shows the number of insertions per participant. The smallest number of 74 insertions was produced by participant P11; the highest number of 652 insertions by participant P15. The variation in insertion numbers in the other translation modes is not as high. However, a large variation in the number of deletions was observed in all three translation modes.

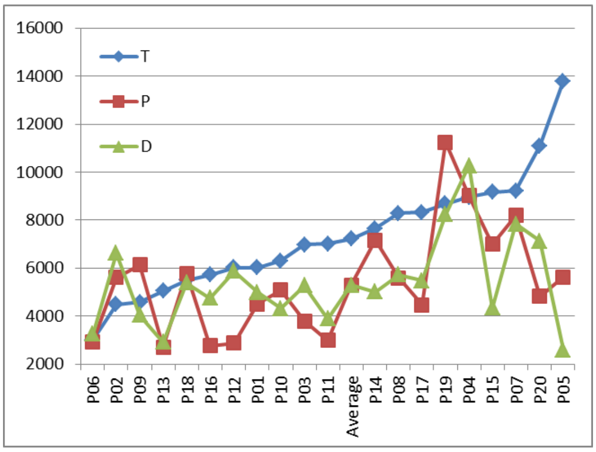

Figure 1: Average translation durations per ST word for each participant (P01 .. P20) for the three translation modes dictation (D), post-editing (P) and from-scratch translation (T)

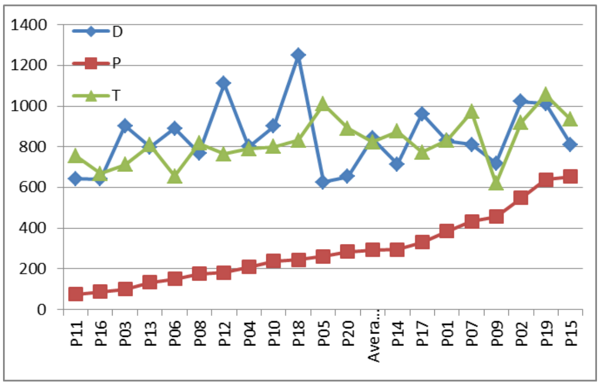

Figure 2: Total number of insertions per participant for the three translation modes dictation (D), post-editing (P) and from-scratch translation (T)

5.2. Pause Analysis

The meaning of pauses in the human translation production flow has been a topic of investigation for many years (e.g., Schilperoord, 1996, O'Brien, 2006). Interruptions in the translators' typing activies have been analyzed as indicators of cognitive processing effort (Kumpulainen, 2015) and some measures to determine sequences of coherent typing have been suggested. In line with Immonen (2006) we find (see Figure 3) that inter-key pauses are shortest within words (0), longer between the last character of a word (or white space character) and the first character of the next word (1), and the longest inter-key pauses are observed before the typing of the first character of a new sentence (3). Figure 3 shows this pattern is common to all three translation modes. It is interesting to note that inner word keystrokes in the P mode are shorter than in the D and T modes, but longer in word- or sentence initial positions.

As a reverse analysis to pausing structure we also examine the sequences of text production activities. There is some discussion as how to define the length of inter keystroke pauses; we take it – with O'Brien (2006) - that "1 second is appropriate for observing delays in a text production event", which is also the measure adopted in the TPR-DB for the definition of production units (Carl & Kay, 2011). Production units, defined in this manner, consist of one or more keystroke. Figure 4 shows their length (in terms of produced characters) is different for the three translation modes. As can be expected, the production units are smallest when post-editing (P) and longest in the D mode, with an average of 3.15, 3.65 and 3.99 characters for the P, T and D modes, respectively. In line with these findings, some participants reported that they translated longer chunks in the dictation mode than during from scratch translation, which most found an interesting effect but also cognitively more effortful. In a discussion after the experiment, one translator said:

"my brain seems to work in a different mode during translation dictation. I have the feeling I would need to better understand the source text before starting dictation so as to produce an 80% correct translation, whereas when typing I can already read ahead in the source text and delete or rearrange the translation more easily. In this sense I find translation dictation more effortful than from-scratch translation".

Figure 5 shows the average number of deletions per production units per participant. Here there is also a clear difference between the three translation modes, with an average of 3.0, 4.1 and 4.6 deleted characters for T, D and P production units, respectively. It was to be expected that longer deletions should be observed in the P mode, since entire words or phrases might be more often replaced during post-editing than during translation where the translator is in control of producing the first draft. The relatively high number of deletions in the D mode might be explained by speech recognition errors.

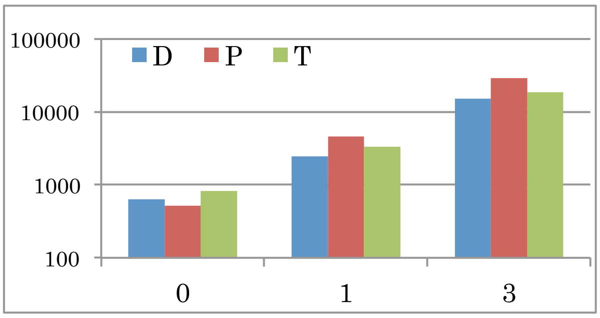

Figure 3: Average length of inter keystroke pauses. 0: inside a word 1: first char of a word, 3: first char in a sentence

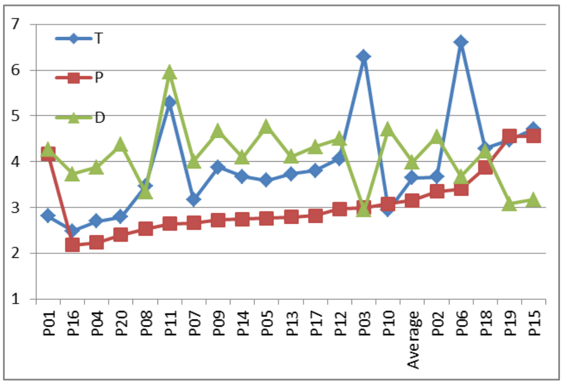

Figure 4: Average number of insertions per production unit

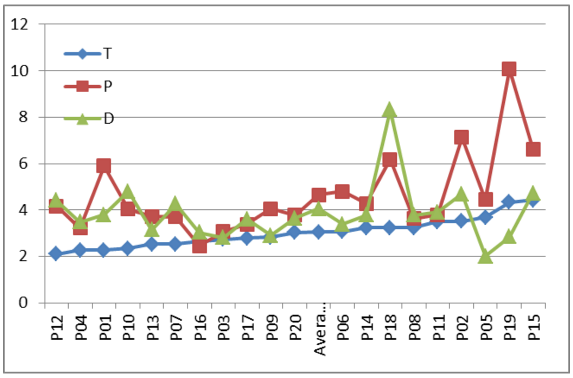

Figure 5: Average number of deletions per production unit

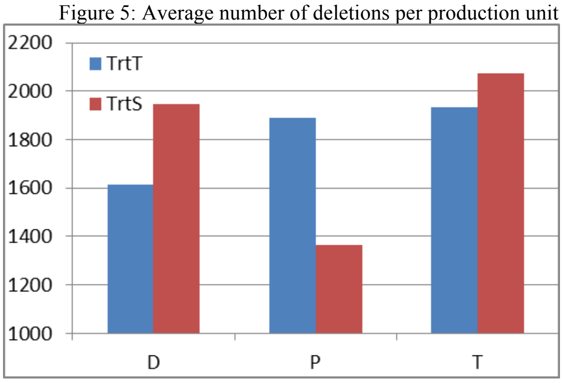

Figure 6: Average total reading times of ST words (TrtS) and target text words (TrtT) for the three translation modes.

5.3. Gaze Data Analysis

Previous findings showed that during from-scratch translations total reading times in the target text are longer than in the source texts (Balling 2014). This finding could not be reproduced with the ENJA15 data collected so far. However, as in previous studies, gaze durations during post-editing are also much longer in the target text than in the source text. It is interesting to observe that, despite the fact that approximately the same number of deletions was produced during P and D (see Table 1), the gazing behavior is quite different in these translation modes. When dictating, the gaze seems mostly fixated on the source text, while during post-editing it is more often on the target text.

6 Conclusion

Our findings confirm those of a previous study of Mees et al. (2015) who find that speaking “translations will encourage [students] to deal with larger units, and thus translate the overall meaning instead of individual words”.

References

Balling, Laura Winther; Carl, Michael. 2014. Production Time Across Languages and Tasks: A Large-Scale Analysis Using the Critt Translation Process Database. The Development of Translation Competence: Theories and Methodologies from Psycholinguistics and Cognitive Science. ed. / John W. Schwieter; Aline Ferreira. Newcastle upon Tyne : Cambridge Scholars Publishing, 2014. p. 239-268

Carl, Michael. 2012. Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)}, 2012, Istanbul, Turkey, Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Asuncion Moreno and Jan Odijk and Stelios Piperidis, European Language Resources Association (ELRA)

Carl, Michael; Moritz Schaeffer; Srinivas Bangalore. 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

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

Schilperoord, J. 1996. It’s About Time: Temporal Aspects of Cognitive Processes in Text Production. Amsterdam: Rodopi.

Minna Kumpulainen. 2015. On the operationalisation of ‘pauses’ in translation process research. Translation & Interpreting Vol 7 No 1 (2015) http://trans-int.org/index.php/transint/article/download/367/183

Mesa-Lao, Bartolomé. 2014. Gaze Behaviour on Source Texts: An Exploratory Study comparing Translation and 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, Cambridge Scholars Publishing pp 219 - 246

Gingold, Kurt. 1978. “The Use of Dictation Equipment in Translation.” In La traduction, une profession. Actes du VIIIe Congrès mondial de la fédération internationale des traducteurs, edited by Paul A. Horguelin, 444–448. Ottawa: Conseil des traducteurs et interprètes du Canada

Germann, Ulrich. Yawat. 2008. Yet Another Word Alignment Tool Proceedings of the ACL-08: HLT Demo Session (Companion Volume) , pages 20–23. Association for Computational Linguistics http://www.aclweb.org/anthology/P08-4006

Mees, Inger, Barbara Dragsted, Inge Gorm Hansen, and Arnt Lykke Jakobsen. Sound effects in Translation. In Interdisciplinarity in interpretation and translation Process Research, Ehrensberger-Dow, Goepferich and O`Brien (eds). 2015, Benjamins Current Topics,

Julián Zapata and Andreas Søeborg Kirkedal. 2015. Assessing the Performance of Automatic Speech Recognition Systems When Used by Native and Non-Native Speakers of Three Major Languages in Dictation Workflows. Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)