# Correlating Translation Product and Translation Process Data of Professional and Student Translators

**Michael Carl and Matthias Buch-Kromann**
Dept. of International Languages Studies & Computational Linguistics,
Copenhagen Business School, 2000 Frederiksberg, Denmark
{mc.isv;mbk.isv}@cbs.dk

## Abstract

The paper presents an exploratory study of the translation processes for 12 student and 12 professional translators. We relate properties of the translators' process data (eye movements and keystrokes) with the quality of the produced translations, using BLEU scores and human evaluation scores for fluency and accuracy to assess translation quality. We also investigate how BLEU scores correlate with human scores, and how BLEU scores depend on the number of reference translations. We segment the translation process into skimming, drafting and post-editing phases, and show that the translation behavior of student and professional translators differ with respect to how they use the translation phases. We also show that students and professionals differ mainly with respect to produced translation fluency.

## 1 Introduction

Although machine translation quality has increased over the past years, current state-of-the-art general-purpose MT rarely meets high quality standards without human intervention. A number of tools for translation assistance have been proposed, such as Translation Memories, MT post-editing tools, and interactive MT. While TMs are widely adopted in the translation industry, they do not include all the possible translation aids that can be provided by a computer today, and the full potential for the utilization of MT in human translation has not been reached yet. However, to provide better MT-based support for human translators, we need a better understanding of human translation processes in different groups of human translators, the bottlenecks experienced by these translators, and how the bottlenecks can be mitigated by automated assistance.

In this paper, we are particularly interested in the differences between professional translators and student translators. For instance, do professionals produce measurably better (or different) translations than students? Do the two groups differ with respect to their working styles? More generally, are there kinds of automated translation assistance that are most helpful in a crowd translation context, and others that are better targeted towards professional translators, or do we need the same tools in both cases?

To approach these questions, we analyze the user activity data (eye movements and keystrokes) of 12 student and 12 professional translators translating two small English texts into Danish. The human translations, as well as a machine translation from produced by Google Translate, are evaluated and compared, both automatically with BLEU and manually with human scores for fluency and accuracy. We also analyze the translation process data and correlate them with the translation quality.

In Section 2, we investigate and quantify different translation phases (skimming, drafting, revision) of student and professional translators and analyze differences in their translation behavior. In Section 3, we look at the quality of the produced translations in terms of BLEU score, accuracy and fluency, and translation time. We compare human translations with Google's MT output. While there is no notable difference for our data in BLEU score between human and machine translation, the human and the machine translation differ significantly in fluency and accuracy.

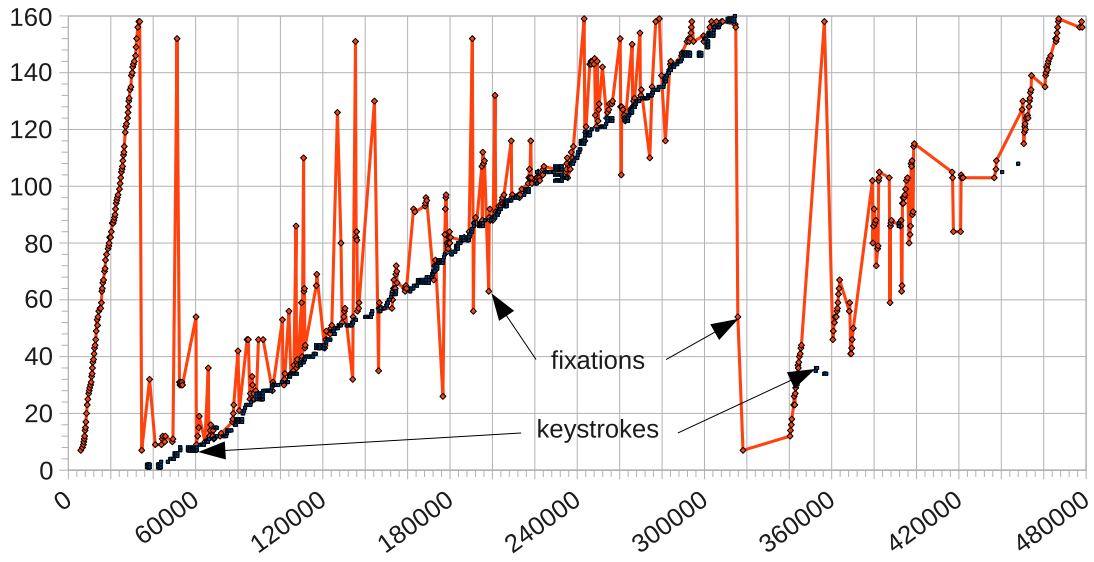


Figure 1: Translation progression graph of translator S17, plotting time (in milli-seconds) against word positions in the source text. Keystrokes and eye movements show a clear separation into skimming, drafting, and post-editing phases.

## 2 Translation Phases

Human translators are usually trained to proceed in three phases: *skimming*, *drafting* and *post-editing*.[1] However, in practice, translators vary greatly with respect to how they produce translations. In the skimming (or orientation) phase, the translator gets acquainted with the material, discovers the meaning of the source text, detects difficult terms, and researches possible translations; in the drafting phase, the actual translation is produced; and in the post-editing phase, the draft is checked and revised. Depending on the size and type of the translation job, further revision cycles may be required, but one revision cycle tends to suffice in small-scale translations, as in the current experiment.

### 2.1 Experimental Design

We conducted a translation experiment (Jensen, 2009) in which 12 professional and 12 student translators translated two texts from English into Danish using the Translog software.[2] Translog presents the source text (ST) in the upper part of the computer screen, and lets the translator type the target text (TT) in the lower part of the screen. When the start button is pressed, the program displays the source text and records the translator's eye movements and keystrokes. After completing the translation, the translator must press a stop button. The program then stores the user activity data (UAD), i.e. the translation as well as the translation process data (eye movements and keystrokes) in a log file.

The translators were asked to translate two texts (A and B) from English into Danish. The texts were articles on current topics which appeared in British newspapers in 2008 and contained approximately 160 words each. Both articles were manipulated so as to vary in their level of complexity, while being comparable with respect to their total character length. The English source texts are shown in the Appendix.

The levels of complexity of the experimental texts were established using three quantitative indicators (Jensen, 2009): readability indices, word frequency calculations, and the number of occurrences of non-literal expressions such as idioms, metaphors, and metonyms. All three indicators showed an increase in the level of complexity from text A to text B. The U.S. grade level indices revealed that 7.8 years of schooling were needed to successfully comprehend text A, while 17.3 years of schooling were needed to successfully comprehend text B. Word frequency in text A was found to contain few low-frequency words (10.7%), while text B contained 28.1% low-frequency words, and the number of non-literal expressions in text A was 1 against 15 non-literal expressions in text B. A complex text is not necessarily difficult to trans-

[1] While the existence of these phases is generally acknowledged, several terms are used to describe them. (Göpferich, 2009), for instance, uses *orientation* or *pre-phase*, *translation* or *main-phase* and *revision* or *post-phase*.

[2] The software can be obtained from `www.translog.dk`

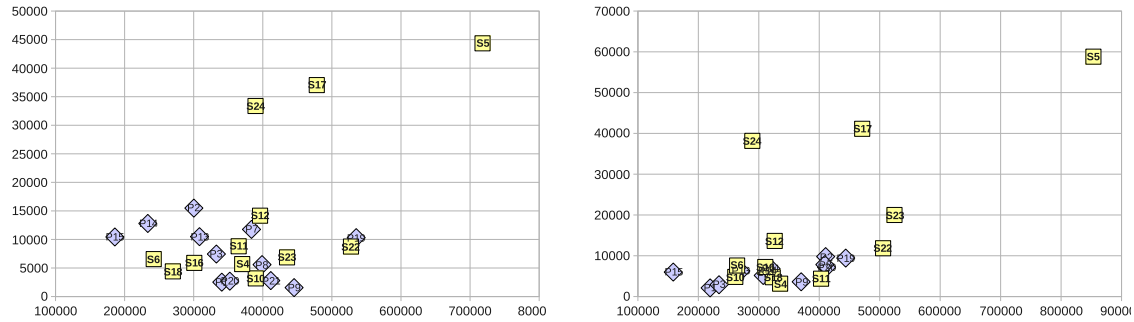


Figure 2: The relationship between drafting time (horizontal) and skimming time (vertical) for the two texts. The left figure represents A data, the right figure B data. Rectangles represent student translators, diamonds represent professionals. On average, students have longer skimming phases than professionals.

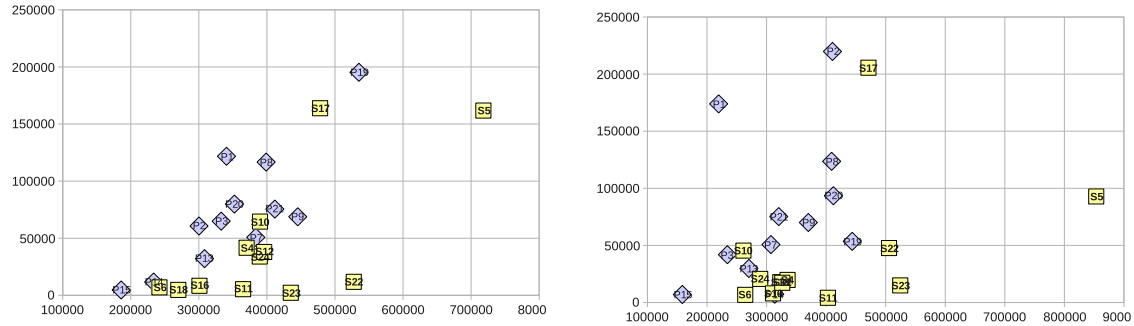


Figure 3: The relationship between drafting time (horizontal) and post-editing time (vertical). The left figure represents A data, the right figure B data. Rectangles represent students, diamonds represent professionals. Professionals tend to have longer post-editing times than students.

late — this depends very much on the experience, skill, and specialization of the translator. However, since all indicators pointed in the same direction, it may be expected that more effort is involved in translating the more complex text B than the less complex text A.

## 2.2 Translation Progression Graphs

The user activity data (UAD) can be represented in so-called translation progression graphs (Perrin, 2003). Figure 1 shows the translation progression graph of student S17. The horizontal axis represents translation time in milliseconds, the vertical axis represents source-language words from the beginning of the text (bottom) to the end (top). As described in (Carl, 2009), keystrokes in TT words are mapped onto their corresponding ST words, ie, all keystrokes that contribute to the translation of the $i$th source word are represented as single dots in the $i$th line from the bottom of the graph. The red (grey) line plots the gaze activities on the source text words. Individual eye fixations are marked with a dot on the fixation line.[3]

[3]Notice that only fixations on the source text are represented in the graph. Our software was not able to compute and map fixations on the emerging target text words.

The progression graph of subject S17 displays a clear distinction between *skimming*, *drafting* and *post-editing* phases. Subject S17 spends almost 40 seconds on getting acquainted with the text, and the graph shows a progression of fixations in which the ST is apparently read from beginning to end.

The drafting phase, which results in an initial translation, takes place between the 40th and 320th seconds. Eye movements can be observed where the translator moves back and forth between the ST, the TT, and the keyboard. In one instance around the 360th second, the recorded eye movements seem to have resulted in a spurious fixation on an ST position far from the current TT position.

The drafting phase is followed by a post-editing phase, from the 320th to the 480th second. Translator S17 seems to re-read much of the ST during post-editing, but only few keystrokes occur, around the 360th and the 440th seconds.

## 2.3 Skimming, Drafting and Post-editing

The analyses of the two texts show a number of similarities. For students, there is a clear tendency towards longer skimming and shorter post-

editing phases, whereas professional translators tend to have shorter skimming and longer post-editing phases. Figure 2 shows the relationship between drafting time and skimming time, and Figure 3 the relationship between drafting time and post-editing time. 9 out of 12 professional translators have some kind of post-editing, while 7 out of 12 student translators do not show any post-editing at all. The inverse observation can be made with respect to skimming: 3 students and no professional translator show skimming times of more than 20 seconds.

| | Text A | | | Text B | | |
|---|---|---|---|---|---|---|
| | TT | ST | PT | TT | ST | PT |
| stud. | 406 | 14 | 60 | 435 | 17 | 41 |
| prof. | 352 | 7 | 81 | 383 | 6 | 78 |

Table 1: Average Translation Time (TT), Skimming Time (ST) and Post-editing Time (PT) in seconds, for students and professional translators for the A and B texts

Overall translation time is approximately 15% longer for students than professional translators. Students spent twice as much time on skimming as professional translators, but only half as much time on post-editing. The average translation, skimming and post-editing time is given in Table 1.

## 3 Correlating Process and Product Data

In this section, we measure the quality of the 24 human translations and compare with a machine translation produced by Google Translate. We study the impact of the number of reference translations on the BLEU score and the correlation of BLEU with accuracy, fluency, and translation time.

### 3.1 BLEU Evaluation

The BLEU score is a metric to evaluate machine translation quality, and is widely used to tune the development of MT systems (Lin and Och, 2004). Based on the assumption that a good translation will share more lexical items with a set of (human generated) references than a bad translation, BLEU compares a test translation with a number of reference translations.

In order to estimate the impact of the number of the reference translations on the BLEU scores, we translated the A and B texts with Google Translate into Danish and evaluated the translations on $5 \times 24$ different subsets of the 24 reference translations. Table 2 shows the results of this experiment where the column #RS indicates the number of used reference translations, the max and min columns the best and worst BLEU scores, and the ratio column the ratio between the max and min scores. That is, the first line in Table 2 gives the minimum and maximum BLEU score for the 24 evaluations obtained when using one reference translation. Line 2 shows the min and max BLEU scores when using any two different reference translations, line 3 the same for 4 reference translations etc.

| | Text B BLEU scores | | | Text A BLEU scores | | |
|---|---|---|---|---|---|---|
| #RS | min | max | ratio | min | max | ratio |
| 1 | 10.68 | 38.99 | 3.65 | 22.90 | 44.53 | 1.94 |
| 2 | 22.79 | 44.50 | 1.95 | 38.84 | 53.05 | 1.36 |
| 4 | 37.00 | 50.71 | 1.37 | 47.71 | 61.18 | 1.28 |
| 8 | 47.18 | 55.47 | 1.17 | 57.57 | 66.02 | 1.14 |
| 23 | 59.12 | 60.25 | 1.09 | 67.08 | 68.14 | 1.01 |

Table 2: Impact of the number of reference translations (#RS) on the BLEU scores for the same google translations when using different subsets of the same 24 reference translations

Table 2 shows that the worst BLEU results are obtained when the reference set contains only one reference translation, while the best results are obtained when the reference set includes 23 reference translations. Looking at the scores for single reference translations, there is a ratio of 3.65 and 1.94 between the worst and the best scores for texts B and A respectively. The ratio of max/min BLEU scores of the same translation based on 4 reference translations decreases to 1.37 and 1.28, and with 23 reference translations it is only 1.09 and 1.01 for the two texts.

That is, the larger the set of reference translations, the more stable we can expect the BLEU score to be, and second, adding more reference translations to a set of existing references will in general increase the value of the score. The table also shows that the Google translation of the A text has better BLEU scores than the B text.

### 3.2 BLEU Evaluation of Human Translations

We have also used BLEU to evaluate the quality of the 24 human translations. Given that more reference translations provide more reliable results, we have evaluated each of the 24 translations by taking the other 23 translations as reference. Although the reference sets are different in each evaluation, with the results discussed in Section 3.1, we suspect that the obtained scores are nevertheless comparable (with an error margin of 1.01 and 1.09 for the A and B texts). The resulting BLEU

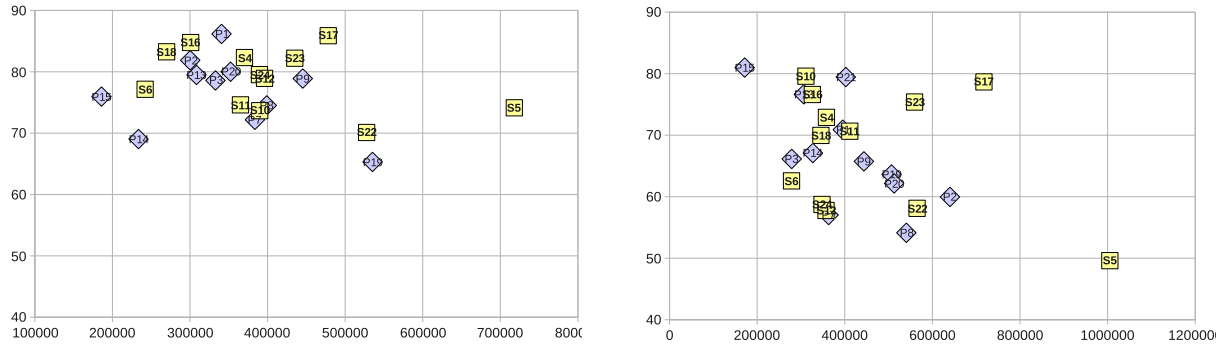


Figure 4: Relation between BLEU score (vertical) and translation time (horizontal) for the A and B texts. The difficult text B (right) has lower BLEU scores than the easier text A (left). Each BLEU score was computed by taking the other 23 translations as reference. The graph shows lack of correlation between BLEU score and translation time, and lack of correlation between BLEU score and translator expertise for the easy text and a negative correlation for the more difficult text (rectangles represent students, diamonds represent professional translators).

| Prof. | B | A | Stud. | B | A |
|---|---|---|---|---|---|
| P15 | 80.99 | 75.96 | S10 | 79.59 | 73.66 |
| P21 | 79.45 | 93.18 | S17 | 78.66 | 85.88 |
| P13 | 76.65 | 79.48 | S16 | 76.61 | 84.77 |
| P1 | 70.89 | 86.17 | S23 | 75.39 | 82.21 |
| P14 | 67.1 | 69.04 | S4 | 72.9 | 82.32 |
| P9 | 65.74 | 78.9 | S11 | 70.65 | 74.61 |
| P3 | 66.14 | 78.59 | S18 | 69.96 | 83.26 |
| P19 | 63.61 | 65.31 | S6 | 62.57 | 77.15 |
| P20 | 62.16 | 80.01 | S22 | 58.11 | 70.11 |
| P2 | 59.96 | 81.87 | S24 | 58.74 | 79.54 |
| P7 | 57.02 | 72.18 | S12 | 57.76 | 78.93 |
| P8 | 54.13 | 74.52 | S5 | 49.6 | 74.13 |

Table 3: BLEU scores for students and professional translators for texts A and B.

| | st+pr. | stud. | prof. |
|---|---|---|---|
| A text average | 78.41 | 78.88 | 77.93 |
| A text median | 78.92 | 79.24 | 78.75 |
| B text average | 67.27 | 67.55 | 66.99 |
| B text median | 66.62 | 70.31 | 65.94 |

Table 4: Average and median BLEU scores for 24 translations, texts A and B: slightly better scores for student translators are not statistically significant.

scores for professional and student translators are shown in Table 3. The worst comparable MT BLEU score is 59.12% and 67.08%, respectively, which is better than the worst human translation, even when taking the error margin into account.

Table 4 shows that text A gives better BLEU scores on average than text B. It also shows that student translators obtain a better BLEU score on average than professional translators, however the difference is not statistically significant, p=0.36 and p=0.44 for the A and B text respectively.

### 3.3 BLEU Score and Translation Time

Figure 4 shows the correlation between translation expertise (students and professional translators), translation time, and the obtained BLEU score. While students need longer than professional translators, no correlation (r=-0.19) can be seen between needed translation time for text A and BLEU score. The B data in Figure 4 give the impression that longer translation time might lead to worse BLEU scores (r=0.44), suggesting that translators who run into problems produce worse translations, even though they spend more time on the translation.

### 3.4 Accuracy and Fluency

The 25 translations were evaluated manually with scores for accuracy and fluency. The A-text consists of 11 segments while the B-text, due to its longer sentences, consists of only 9 segments. Each segment was evaluated independently and blindly by a native speaker of Danish, and assigned a score between 0 and 5 for accuracy and fluency. The scale is inspired by (White, 1992) and reproduced in the appendix. Because one translator (S12) did not translate the entire text B, we added a category 0 for non-translated segments. We computed the average accuracy and fluency score for each translator, only taking into account the fully translated segments.

Table 5 shows the average accuracy and fluency scores for both texts. Students and professional translators obtain approximately the same degree of accuracy for the text A, but professionals per-

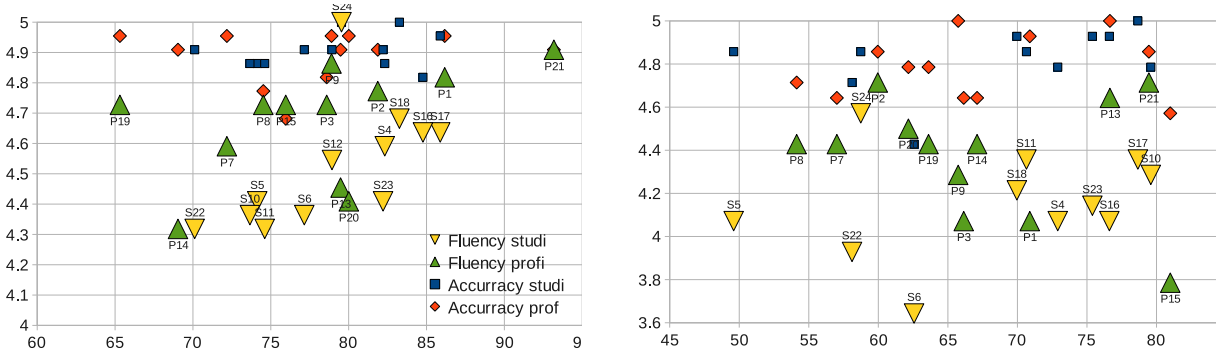


Figure 5: Relation between accuracy and fluency (vertical) and BLEU score (horizontal) for text A (left) and text B (right). The graphs have different symbols for students and for professional translators, with Translator IDs shown for fluency scores.

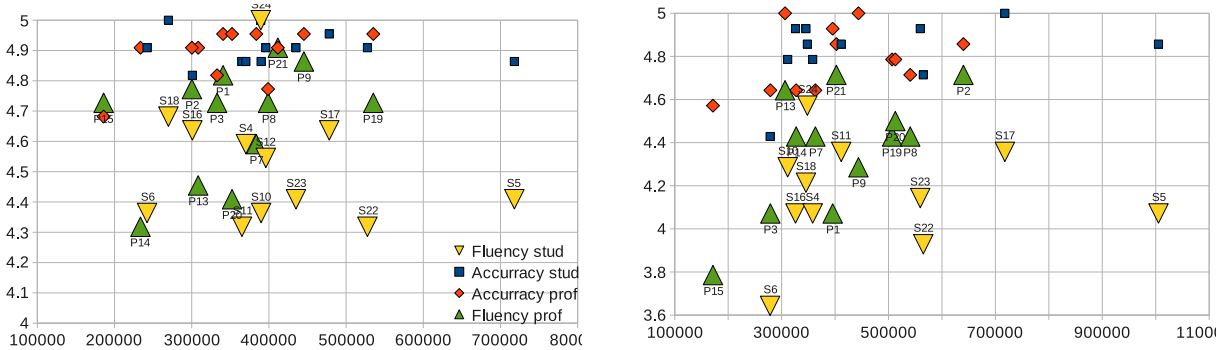


Figure 6: Relation between accuracy and fluency (vertical) and translation time (horizontal) for text A (left) and text B (right). Student and professional Translators' ID is provided with the fluency scores.

form slightly better in the more complicated text B. Professional translators have better fluency scores on average than students, and both groups outperform the Google translation significantly in terms of both accuracy and fluency.

There is one outlier (S12) who translated only 2/3 of text B, but who nevertheless reaches a BLEU score of 57.76, with average accuracy and fluency scores of 2.71 an 2.6, respectively. Not taking this person into account, the average fluency of the B text becomes 4.16 and the average accuracy 4.82 (these figures are in bold in Table 5).

| | Text A | | | Text B | | |
|---|---|---|---|---|---|---|
| | stud. | prof. | gt | stud. | prof. | gt |
| accuracy | 4.91 | 4.89 | 3.89 | **4.65** | 4.79 | 3.21 |
| fluency | 4.52 | 4.67 | 3.14 | **3.96** | 4.38 | 2.71 |

Table 5: Accuracy and fluency for student and professional translators and for the Google translation.

### 3.5 BLEU, Accuracy and Fluency

Figure 5 shows the relation between the BLEU score and the accuracy and fluency of the translations. The BLEU score is not a particularly good predictor of the accuracy of human high-quality translations (r=0.13 and r=0.3 for the A and B texts respectively). This could be expected, given that BLEU exclusively uses the reference translations to determine a score (ie, it only compares target language sentences).

The data, however, show that fluency is correlated with BLEU score (r=0.44) for the easier text A, for students even more so than for professionals. This suggests that translators tend to agree on how to render the easy translations. The situation is different in the more difficult text B (Figure 5, right) where there seems to be no correlation (r=0.14) between the BLEU score and the fluency of the translation. The average fluency score is lower, but the unrelated BLEU score indicates that there are more ways to render a complex translation fluently (and accurately), and that translators seem to diverge on the formulation of the translation.

We used a one-tailed two-sample t-test to test whether professionals were better than students with respect to fluency and accuracy in the two texts. The tests showed that professionals were slightly better than students with respect to fluency (p=0.038 and p=0.025 in texts A and B, respectively), but that there was no difference with respect to accuracy (p=0.31 and p=0.23, respectively) While accuracy of the translations can, thus,

be reached for students and professionals to a high degree, professional translators seem to be better able to produce more fluent texts.

### 3.6 Accuracy, Fluency and Translation Time

Figure 6 does not suggest a notable correlation between accuracy and translation time (r=0.22 and r=0.2 for A and B texts respectively). Translators seem to be able to quite accurately transfer the meaning into the target language independently of the time they actually use to produce the translation. The graphs in Figure 6 show, however, that professional translators are better capable of turning longer translation times into more fluent translations, (r=0.4 and r=0.64) while this is not so the case for the students (r=-0.22 and r=0.15). One possible explanation is that although translators could be expected to produce a better translation as they spend more time on it, this effect is counter-weighed by the effect that good translators tend to be faster than poor translator, an effect which is particularly strong for student translators.

## 4 Conclusion

The paper compares the translation behavior of student and professional translators and correlates it with the produced translation quality. Reading and text production activities are registered and analyzed based on eye-tracking and keyboard-logging data. The translation processes can be divided into three phases, a skimming phase in which the translator obtains a rough idea of the text, a drafting phase in which a first version of the translation is drafted, and a post-editing phase in which the draft is revised. We have manually and automatically evaluated the translations and related it to the analyzes of the process data. Our investigation suggests the following conclusions:[4]

- Student translators use more time for skimming than professional translators (Figure 2)
- Professional translators use more time for post-editing than student translators (Figure 3)
- For difficult texts, BLEU scores may correlate negatively with the total translation time. (Figure 4).
- Students and professionals produce equally accurate translations (Figures 5 and 6).
- Professional translators produce more fluent texts more quickly than students (Figure 6).
- For easy texts, BLEU scores correlate with translation fluency (Figure 5).

Our study shows that for the texts in the experiments, non-professional translators (bilingual students and translation students) are able to reproduce the source text meaning in their native target language just as well as professionals. They need approximately 15% more time than professional translators, but do not reach the same degree of fluency. Professionals work in a more structured manner, postponing revisions to a post-editing phase, while student translators revise their translations during the drafting phase.

These findings suggest that different tools are needed to assist different types of translators, who have different degrees of expertise and training, during the different translation phases. Laypersons may profit from skimming tools, since, unlike professional translators, they seem to need better access to the ST. Skimming support tools might resume parts of the ST, point to frequent terms or collocations, and suggest translations of those passages. Untrained translators might also profit from tools that help to increase the fluency of the target language production. Such considerations could, for instance, be taken into account when designing Wiki translation tools.

[4]These results would have to be taken with caution because of the nature of the texts (short newspaper articles) and the translation setting (the translations were performed on a voluntary basis in an academic context).

## References


Carl, Michael. 2009. Triangulating product and process data: quantifying alignment units with keystroke data. *Copenhagen Studies in Language*, 38:225–247.

Göpferich, S. 2009. The translation of instructive texts from a cognitive perspective: novices and professionals compared. In Göpferich, Susanne, Fabio Alves, and Inger M. Mees, editors, *New Approaches in Translation Process Research*, pages 5–55, Copenhagen. Copenhagen: Samfundslitteratur.

Hansen, Gyde, editor. 1999. *Probing the process in translation: methods and results*, volume 24 of *Copenhagen Studies in Language*. Copenhagen: Samfundslitteratur.

Jensen, Kristian T.H. 2009. Distribution of attention between source text and target text during translation. In *IATIS*.

Lin, C. and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*. `http://www.mt-archive.info/ACL-2004-Lin.pdf`.

Perrin, Daniel. 2003. Progression analysis (PA): investigating writing strategies at the workplace. *Pragmatics*, 35:907–921.

White, J. S. 1992. The DARPA Machine Translation Evaluation: Implications for Methodological Extensibility. In *Association for Machine Translation of the Americas*, San Diego.

## Appendix

### Source Text A

*Killer nurse receives four life sentences*
Hospital Nurse Colin Norris was imprisoned for life today for the killing of four of his patients. 32 year old Norris from Glasgow killed the four women in 2002 by giving them large amounts of sleeping medicine. Yesterday, he was found guilty of four counts of murder following a long trial. He was given four life sentences, one for each of the killings. He will have to serve at least 30 years. Police officer Chris Gregg said that Norris had been acting strangely around the hospital. Only the awareness of other hospital staff put a stop to him and to the killings. The police have learned that the motive for the killings was that Norris disliked working with old people. All of his victims were old weak women with heart problems. All of them could be considered a burden to hospital staff.

### Source Text B

*Families hit with increase in cost of living*
British families have to cough up an extra £31,300 a year as food and fuel prices soar at their fastest rate in 17 years. Prices in supermarkets have climbed at an alarming rate over the past year. Analysts have warned that prices will increase further still, making it hard for the Bank of England to cut interest rates as it struggles to keep inflation and the economy under control. To make matters worse, escalating prices are racing ahead of salary increases, especially those of nurses and other healthcare professionals, who have suffered from the government s insistence that those in the public sector have to receive below-inflation salary increases. In addition to fuel and food, electricity bills are also soaring. Five out of the six largest suppliers have increased their customers' bills.

## Evaluation categories

### Accuracy

5 All meaning expressed in the source fragment appears in the translation fragment
4 Most of the source fragment meaning is expressed in the translation fragment
3 Much of the source fragment meaning is expressed in the translation fragment
2 Little of the source fragment meaning is expressed in the translation fragment
1 None of the meaning expressed in the source fragment is expressed in the translation fragment
0 Untranslated fragment

### Fluency

5 The translation is perfect both stylistically and grammatically
4 Slightly unnatural stylistics, lexicalzation, or word order, or minor spelling mistakes
3 Few, rather minor grammatical errors
2 Many, possibly major grammatical errors
1 Completely unintelligible
0 Untranslated fragment