Interactive Assistance to Human Translators using Statistical Machine Translation Methods

Philipp Koehn
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk

Barry Haddow
School of Informatics
University of Edinburgh
bhaddow@inf.ed.ac.uk

Abstract

We investigate novel types of assistance for human translators, based on statistical machine translation methods. We developed Caitra, a tool that makes suggestions for sentence completion, shows word and phrase translation options, and allows post-editing of machine translation output. A user study validates the types of assistance and provides insight into the human translation process.

1 Introduction

While machine translation has made tremendous progress over the last years, this progress has made little inroads into tools for human translators. Although it has become frequent practice in the industry to provide human translators with machine translation output for post-editing, typically no deeper integration of machine translation and human translation is found in translation agencies.

An interesting new approach was pioneered by the TransType project (Langlais et al., 2000). The machine translation system makes sentence completion predictions in an interactive machine translation setting. The users may accept them or override them by typing in their own translations, which triggers new suggestions by the tool (Barrachina et al., 2009). But also other information of the machine translation system may be useful for the human translator, such as alternative translations for the input words and phrases.

We developed the web-based translation tool Caitra that offers various types of assistance and carried out a study involving ten human translators, whose interaction with the tool was logged in great detail. Our study showed that most translators were able to produce translations faster and better with such assistance. The detailed log also allowed us to explore how translators spend their time, and how this changes with assistance.

Recently, with the availability of key loggers there has been increasing interest in the process studies of translation (Fraser, 1996). Such studies of user activity data focused on key strokes and considered statistics such as revision ratios (Buchweitz and Alves, 2006). Carl et al. (2008) presents a study that also uses eye trackers. Studies may also make use of *think aloud protocols* (Jääskeläinen, 2001) in which the translator narrates the thought process behind her actions. The interactive machine translation assistance (that we present in Section 2.1) was evaluated by Macklovitch (2006) with an emphasis on the user experience.

Our user study is an extension of this prior work. It extends to different and novel types of assistance. We use a relatively large corpus and a large number of test subjects.

2 Types of Assistance

Caitra is implemented as a web-based client-server architecture, using Ajax Web 2.0 technologies. The machine translation back-end is powered by the Moses decoder. The tool is delivered over the web to allow for easier user studies, but also to expose it to a wider community to gather additional feedback. You can find the tool online at http://www.caitra.org/

2.1 Prediction of Sentence Completion

In the sentence-completion paradigm, the human translator is still in charge of creating the translation word by word, but she is aided by a system that interactively makes suggestions for completing the sentence, and updates these suggestions based on her input. The scenario is similar to the auto-completion function for words, search terms, email addresses, etc. in modern office applications or predictive text entry in mobile phones.

See Figure 1 for a screenshot of the incarnation of this method in Caitra. The user is given an input sentence and a standard web text box to type in her translation. The system makes suggestions about the next word (or phrase) to be added to the translation. The user may accept this (by pressing the TAB key), or type in her own translation. The tool updates the prediction based on the user input.

The predictions are based on a statistical machine translation system. Given the input and the partial translation of the user, the machine translation system computes the optimal translation of

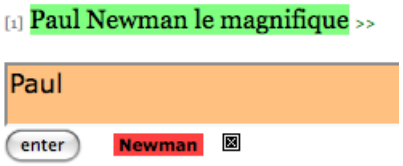


Figure 1: **Interactive Machine Translation.** Caitra uses the search graph of the machine translation decoder to suggest words and phrases to continue the translation.

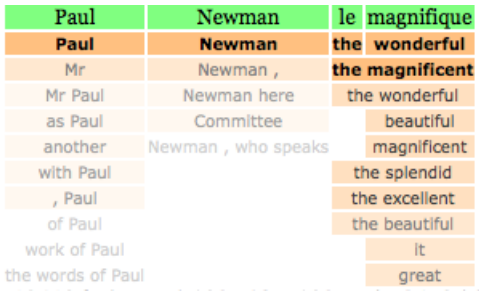


Figure 2: **Translation Options.** The most likely word and phrase translation are displayed alongside the input words, ranked and color-coded by their probability.

the input sentence, constrained by matching the user input. This translation is provided to the user in form of short phrases (mirroring the underlying phrase-based statistical translation model).

2.2 Options from the Translation Table

Phrase-based statistical machine translation methods acquire their translation knowledge in form of phrase translation tables automatically from large amounts of translated texts. For each input word or input word sequence, this translation table is consulted for the most likely translation options.

These translation options may also be of interest to a human translator, so we display them in Caitra. See Figure 2 for an example. The options are color-coded and ranked based on their score. Note that since these options are extracted from a translated corpus using various automatic methods, often inappropriate translations are included, such as the translation of *Newman* into *Committee*.

2.3 Post-Editing Machine Translation

The addition of full sentence translation of the machine translation system is trivial compared to the other types of assistance. When a user starts a new sentence using this aid, the text box already contains the machine translation output and the user only makes changes to correct errors.

See Figure 3 for an example. The tool also compares the user's translation in form of string edit distance against the machine translation. This is illustrated above the text box, to possibly alert the user to mistakenly dropped or added content.

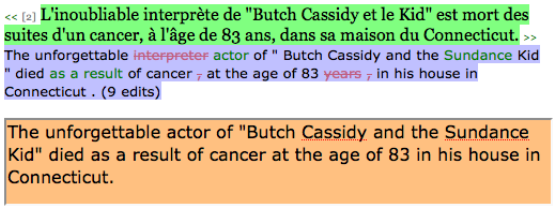


Figure 3: **Post-Editing Machine Translation.** Starting with the sentence translation of the machine translation system, the user post-edits and the tool indicates changes.

3 User Study

Caitra tracks every key stroke and mouse click of the user, which then allows for a detailed analysis of the user's interaction with the tool. See Figure 4 for a graphical representation of the user activity during the translation of a sentence. The graph plots sentence length (in characters) against the progression of time.

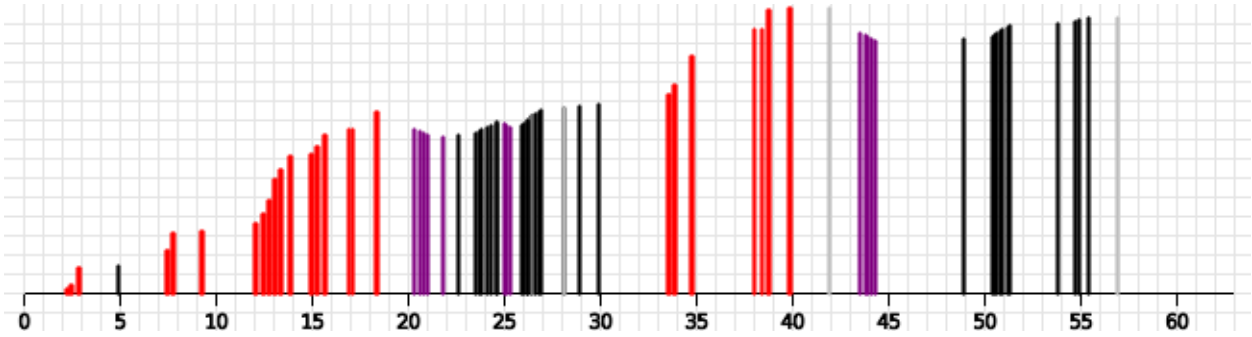
3.1 Experimental Design

We recruited 10 human translators for our study. Half of the translators are native speakers of French (L1) studying in an English speaking country, the other half native speakers of English (L2) with university-level French skills. In the following, the translators are referred to as L1a, L2a, L1b, L2b, and so on. All translators are associated with the University of Anonymous, being either students or staff. The delivery of Caitra over the web allowed the translators to work at their own convenience within a two week period. They were rewarded for their efforts with a fixed amount of money instead of an hourly wage to give them an incentive to be productive.

Each translator translated the same set of documents with the total size of 194 sentences from French to English. The document set was taken from the 2009 EACL WMT workshop and consists of news paper articles from *Le Devoir*, *Le Figaro*, *Les Echos* and *Liberation*.

The text is broken up into five blocks of about 40 sentences, so that each block consists of one to three complete documents each. Table 1 gives details about the distribution of the blocks to the translators under the five different types of assistance: (1) unassisted, (2) postediting machine translation output, (3) options from the translation table, (4) prediction (sentence completion), (5) options and predictions.

While it is not possible to give the same block to the same translator with different types of assistance, we distributed them in a way that each block



Input: "Un échange de coups de feu s'est produit, et la moitié des ravisseurs ont été tués, les autres s'enfuyant", a dit ce responsable qui a requis l'anonymat.

MT: "A exchange of fire occurred, and half of the kidnappers were killed, the other is enfuyant," said this official who has requested anonymity.

User: "An exchange of fire occurred, and half of the kidnappers were killed, the others running away", said the source who has requested anonymity.

Figure 4: **User Activity.** The graph plots the time spent on translation (in seconds, x-axis) against the length of the sentence (y-axis) with color-coded activities (bars). Bars indicate the sentence length at each point in time when a user action takes place. Acceptance of predictions are red, DEL key strokes purple, key strokes for cursor movement grey, and key strokes that add characters are black. The user first slowly accepted the interactive machine translation predictions (second 0-12), then more rapidly (second 12-20), followed by a period of deletions and typing that did not make the translation longer (second 20-30). After a short pause, predictions were accepted again (second 33-40), followed by deletions and typing (second 40-57).

Block	Time	a	b	c	d	f
A	3.9s	U	O+P	P	O	PE
B	3.4s	PE	U	O+P	P	O
C	3.7s	O	PE	U	O+P	P
D	3.1s	P	O	PE	U	O+P
E	3.2s	O+P	P	O	PE	U

U = Unassisted, PE = Postedit, P = Prediction, O = Options

Table 1: **Permutation of Assignments.** Translation blocks A–E (one to three documents with 40 sentences) are assigned to the human translators a–e to translate under varying types of assistance.

is translated by each type of translator (L1/L2) under each condition. One concern is that different blocks pose different degrees of difficulty. This is true to some extent in our data set, where the average translation time for the five blocks varies from 3.1 to 3.9 seconds per word.

3.2 Evaluation

Since Caitra logs the time spent on each sentence, it is straight-forward to compute the average time per input word which we use as our evaluation of translation speed. Speed is not the only criterion of success, the translations have to be correct as well. Evaluation of translation quality is a difficult problem, since ten different translators will almost always produce ten different translations, and it hard to assess which ones are correct.

We relied on human judges to check each translation. Given the French source sentence in context (two preceeding and two following sentences), they were asked to classify translations as correct with the following instructions:

> *Indicate whether each user's input represents a fully fluent and meaning-equivalent translation of the source. The source is shown with context, the actual sentence is bold.*

A web-based tool was deployed to solicit these judgements. All ten translations for each sentence were displayed on the same screen. The judges were fluent in both French and English.

4 Results and Analysis

The detailed logs of the translators actions offer a wealth of data. We are not only interested in translation speed and quality, but we would also like to gain some insight into the translation process and the behavior of the translators.

4.1 Speed and Quality

The most important questions from the view of the tool developer are: do human translators produce better translations and are they faster? The short answer is: mostly, yes.

Table 2 gives a slightly longer answer. On average, the human translators are faster and also achieve better translation quality using any type of assistance offered. Only in very few instances, they are both slower and worse. Individual results vary, see the table for details. Translators are fastest with postediting and obtain highest translation performance when post-editing and using prediction and options.

When post-editing, 8 translators are faster and better, when using the options 4 translators are faster and better, when using the predictions 6 translators are faster and better, and when using both predictions and options 6 translators are

User	Unassisted	Postedit		Options		Prediction		Prediction+Options	
L1a	3.3sec/word	**1.2s**	**(-2.2s)**	**2.3s**	**(-1.0s)**	**1.1s**	**(-2.2s)**	**2.4s**	**(-0.9s)**
	23% correct	**39%**	**(+16%)**	**45%**	**(+22%)**	**30%**	**(+7%)**	**44%**	**(+21%)**
L1b	7.7sec/word	**4.5s**	**(-3.2s)**	**4.5s**	**(-3.3s)**	**2.7s**	**(-5.1s)**	**4.8s**	**(-3.0s)**
	35% correct	**48%**	**(+13%)**	**55%**	**(+20%)**	**61%**	**(+26%)**	**41%**	**(+6%)**
L1c	3.9sec/word	**1.9s**	**(-2.0s)**	**3.8s**	**(-0.1s)**	**3.1s**	**(-0.8s)**	**2.5s**	**(-1.4s)**
	50% correct	**61%**	**(+11%)**	**54%**	**(+4%)**	**64%**	**(+14%)**	**61%**	**(+11%)**
L1d	2.8sec/word	**2.0s**	**(-0.7s)**	2.9s	(+0.1s)	2.4s	(-0.4s)	**1.8s**	**(-1.0s)**
	38% correct	**46%**	**(+8%)**	59%	(+21%)	37%	(-1%)	**45%**	**(+7%)**
L1e	5.2sec/word	**3.9s**	**(-1.3s)**	4.9s	(-0.2s)	**3.5s**	**(-1.7s)**	4.6s	(-0.5s)
	58% correct	**64%**	**(+6%)**	56%	(-2%)	**62%**	**(+4%)**	56%	(-2%)
L2a	5.7sec/word	**1.8s**	**(-3.9s)**	**2.5s**	**(-3.2s)**	**2.7s**	**(-3.0s)**	**2.8s**	**(-2.9s)**
	16% correct	**50%**	**(+34%)**	**34%**	**(+18%)**	**40%**	**(+24%)**	**50%**	**(+34%)**
L2b	3.2sec/word	2.8s	(-0.4s)	*3.5s*	*(+0.3s)*	*6.0s*	*(+2.8s)*	*4.6s*	*(+1.4s)*
	64% correct	56%	(-8%)	*60%*	*(-4%)*	*61%*	*(-3%)*	*57%*	*(-7%)*
L2c	5.8sec/word	**2.9s**	**(-3.0s)**	4.6s	(-1.2s)	**4.1s**	**(-1.7s)**	**2.7s**	**(-3.1s)**
	52% correct	**53%**	**(+1%)**	37%	(-15%)	**59%**	**(+7%)**	**53%**	**(+1%)**
L2d	3.4sec/word	3.1s	(-0.3s)	4.3s	(+0.9s)	3.8s	(+0.4s)	3.7s	(+0.3s)
	49% correct	49%	(+0%)	51%	(+2%)	53%	(+4%)	58%	(+9%)
L2e	2.8sec/word	**2.6s**	**(-0.2s)**	*3.5s*	*(+0.7s)*	2.8s	(-0.0s)	*3.0s*	*(+0.2s)*
	68% correct	**79%**	**(+11%)**	*59%*	*(-9%)*	64%	(-4%)	*66%*	*(-2%)*
avg.	4.4sec/word	**2.7s**	**(-1.7s)**	**3.7s**	**(-0.7s)**	**3.2s**	**(-1.2s)**	**3.3s**	**(-1.1s)**
	47% correct	**55%**	**(+8%)**	**51%**	**(+4%)**	**54%**	**(+7%)**	**53%**	**(+6%)**

Table 2: **Speed and Quality.** On average, translators are faster and also achieve better translation quality using any of the assistances offered. Individual results vary.

faster and better. 4 Translators are faster and better with all of the assistances offered, and only two translators achieved no gains in both dimensions with any assistance.

A note on the quality judgments: We were surprised by the low correctness numbers we obtained from the human judges (the overall average is 50%). When using this metric in machine translation evaluation, human reference translations were judged 85-90% correct using the same metric. After querying some of the human judges, we were left with the impression that they were overly critical (*"this translation sounds funny to me"*), and may also be tempted, when given 10 translations at a time, to label half of them as correct and the other half as wrong — an implicit ranking of the translations.

4.2 Utilizing Assistance

Let us now take a closer look at how translators used the assistance offered to them.

The log of each sentence translation is a sequence of events (key strokes, clicks) at specific time points. We would like to characterize broader activities, such as *typing* or *pauses*, and break up the very detailed sequence of actions into larger intervals of such activities.

We define an activity as a time interval, in which we observe specific events. Say, the activity *typing* is an interval of time that only consists of keystrokes without any significant pauses and no other event. By *significant pause*, we imply that the window of one second before and after a keystroke is part of the typing activity.

Definition: Activity Intervals. Each event e has a timepoint $t(e)$ and a type $y(e) \in Y = \{\text{key, click, tab}\}$. Let L be the set of all events for the translation of a sentence, and w the window size (one second). We define an activity is an interval $I = [t_1, t_2]$ of the type $A \subset Y$ as

$$\begin{aligned} &I[t_1,t_2] \text{ has type } A \Leftrightarrow \\ &\quad \forall e \in L : t_1 - w \leq t(e) \leq t_2 + w \\ &\qquad\qquad \rightarrow y(e) \in A \\ &\quad \text{and} \\ &\quad \forall y \in A, t \in I : \exists e \in L : y(e) = y, \\ &\qquad\qquad t - w \leq t(e) \leq t + w \end{aligned} \tag{1}$$

Under this definition, the period of translating a sentence segments into a unique sequence of maximal intervals of activities (meaning, no neighboring intervals have the same activity).

The set of different activity types is a power set of the types of events, but we collapse all activities with multiple types of events into one type: the *mixed* activity. We further break up pauses into

- initial pauses: the pause at the beginning of the translation, if it exists
- end pause: the pause at the end of the translation, if it exists
- short pause of length 2–6 seconds
- medium pauses of length 6–60 seconds
- big pauses longer than 60 seconds

Note that there are no pauses shorter than 2 seconds, since these are necessarily part of non-pause activities.

We are less interested in the number of intervals, but rather how much time is spent on each type of activity. Does the translator spend most of her time in big pauses, or on typing keys? Table 3 gives a breakdown for each translator for each type of assistance. The timing information is given as seconds per input word (meaning that the total time spend on each activity is divided by the total number of words in the input documents).

Let us take a closer look at two translators: L1b and L2e. L1b is the slowest and a worse than average translator when unassisted. She makes good use of both types of assistance, spending 0.38 seconds on clicking, 0.41 seconds on tabbing (accepting predictions), and using both (0.47 seconds, 0.24 seconds, respectively), when both are offered. This cuts down the time spend on regular typing by 0.9–1.4 seconds. Also, much less time is spend on pauses of various types.

L2e is one of the best translators, but gets hardly any gains from the assistance. The table reveals why: She hardly uses clicks and tabs when offered, and not at all when both are offered. The time spend on typing changes hardly. Nevertheless, she is faster in post-editing, mostly due to spending a second less on typing, although some of those gains are eaten up by more pausing, mostly medium pauses.

4.3 Origin of Characters

Time spent on activities is one way to measure the utilization of assistance. Another is to trace back the origin of the characters in the final translation to their generating activity. We follow the construction of the translation and record how each character is generated.

Table 4 gives a breakdown for each translator for each type of assistance. The break-down into different origins mirrors the time spend on the activities. For instance, translator L1b spent 0.89s, 0.47s, and 0.24s on typing, clicking and tabbing (0.04s on mixed activities — no translator spends significant time on this). The resulting translations contain characters that originate 21%, 44%, and 33%, respectively, from these activities. These numbers do suggest that clicking and tabbing is more efficient in generating characters in the translation.

User: L1a	key	click	tab	mt
Postedit	9%	-	-	90%
Options	13%	86%	-	-
Prediction	10%	-	88%	-
Prediction+Options	21%	31%	46%	-
User: L1b	key	click	tab	mt
Postedit	18%	-	-	81%
Options	59%	40%	-	-
Prediction	14%	-	85%	-
Prediction+Options	21%	44%	33%	-
User: L1c	key	click	tab	mt
Postedit	18%	-	-	81%
Options	43%	56%	-	-
Prediction	45%	-	54%	-
Prediction+Options	30%	68%	1%	-
User: L1d	key	click	tab	mt
Postedit	14%	-	-	85%
Options	99%	0%	-	-
Prediction	22%	-	77%	-
Prediction+Options	15%	0%	84%	-
User: L1e	key	click	tab	mt
Postedit	17%	-	-	82%
Options	70%	29%	-	-
Prediction	32%	-	67%	-
Prediction+Options	73%	4%	22%	-
User: L2a	key	click	tab	mt
Postedit	11%	-	-	88%
Options	8%	91%	-	-
Prediction	17%	-	82%	-
Prediction+Options	15%	10%	74%	-
User: L2b	key	click	tab	mt
Postedit	17%	-	-	82%
Options	36%	63%	-	-
Prediction	100%	-	-	-
Prediction+Options	10%	89%	-	-
User: L2c	key	click	tab	mt
Postedit	13%	-	-	86%
Options	14%	85%	-	-
Prediction	17%	-	82%	-
Prediction+Options	14%	71%	13%	-
User: L2d	key	click	tab	mt
Postedit	26%	-	-	73%
Options	93%	5%	-	-
Prediction	100%	-	-	-
Prediction+Options	59%	40%	-	-
User: L2e	key	click	tab	mt
Postedit	20%	-	-	79%
Options	77%	22%	-	-
Prediction	61%	-	38%	-
Prediction+Options	100%	-	-	-

Table 4: **Origin of Characters.** For each character in the final translation, we trace back its origin, which is either a keystroke, a click on an option, a TAB key stroke to accept an prediction, or the MT output as starting point for edits.

User: L1a	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab	mixed
Unassisted	3.31s	0.07s	0.11s	0.18s	1.04s	0.07s	1.84s	-	-	-
Postedit	1.16s	0.48s	0.08s	0.05s	0.27s	-	0.27s	-	-	-
Options	2.28s	0.19s	0.09s	0.32s	0.62s	-	0.34s	0.68s	-	0.04s
Prediction	1.11s	0.04s	0.02s	0.07s	0.22s	-	0.27s	-	0.42s	0.06s
Prediction+Options	2.38s	0.13s	0.12s	0.22s	0.73s	-	0.60s	0.27s	0.25s	0.07s
User: L1b	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab	mixed
Unassisted	7.74s	1.29s	0.11s	0.25s	1.83s	1.94s	2.32s	-	-	-
Postedit	4.50s	1.47s	0.43s	0.14s	0.95s	0.41s	1.09s	-	-	-
Options	4.46s	0.59s	0.11s	0.36s	0.85s	0.70s	1.46s	0.38s	-	0.01s
Prediction	2.67s	0.29s	0.27s	0.19s	0.74s	0.09s	0.63s	-	0.41s	0.05s
Prediction+Options	4.79s	0.58s	0.35s	0.41s	1.31s	0.48s	0.89s	0.47s	0.24s	0.04s
User: L1c	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab	mixed
Unassisted	3.88s	0.23s	0.16s	0.33s	0.71s	-	2.45s	-	-	-
Postedit	1.92s	0.59s	0.16s	0.10s	0.49s	-	0.57s	-	-	-
Options	3.77s	0.36s	0.19s	0.55s	0.88s	-	1.15s	0.58s	-	0.07s
Prediction	3.11s	0.20s	0.27s	0.38s	0.46s	-	1.28s	-	0.44s	0.07s
Prediction+Options	2.53s	0.27s	0.18s	0.41s	0.29s	-	0.71s	0.56s	0.02s	0.08s
User: L1d	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab	mixed
Unassisted	2.79s	0.23s	0.04s	0.20s	0.39s	0.14s	1.78s	-	-	-
Postedit	2.05s	0.53s	0.15s	0.10s	0.50s	0.23s	0.56s	-	-	-
Options	2.89s	0.13s	0.12s	0.30s	0.50s	-	1.83s	0.01s	-	0.00s
Prediction	2.38s	0.18s	0.11s	0.29s	0.36s	0.08s	0.73s	-	0.60s	0.03s
Prediction+Options	1.78s	0.13s	0.11s	0.23s	0.18s	-	0.50s	0.00s	0.60s	0.04s
User: L1e	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab	mixed
Unassisted	5.17s	0.28s	0.04s	0.33s	1.86s	0.48s	2.18s	-	-	-
Postedit	3.87s	0.76s	0.08s	0.22s	0.94s	0.73s	1.15s	-	-	-
Options	4.94s	0.28s	0.10s	0.56s	1.36s	0.38s	1.99s	0.26s	-	0.02s
Prediction	3.46s	0.19s	0.04s	0.40s	0.89s	0.14s	1.19s	-	0.53s	0.08s
Prediction+Options	4.64s	0.18s	0.10s	0.55s	1.02s	0.46s	2.03s	0.06s	0.23s	0.02s
User: L2a	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab	mixed
Unassisted	5.68s	0.54s	0.12s	0.31s	1.78s	0.71s	2.21s	-	-	-
Postedit	1.82s	0.66s	0.10s	0.09s	0.46s	0.20s	0.31s	-	-	-
Options	2.46s	0.36s	0.13s	0.25s	0.60s	0.12s	0.24s	0.73s	-	0.03s
Prediction	2.70s	0.32s	0.20s	0.14s	0.80s	0.43s	0.48s	-	0.26s	0.06s
Prediction+Options	2.82s	0.21s	0.42s	0.17s	1.20s	0.07s	0.44s	0.13s	0.16s	0.02s
User: L2b	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab	mixed
Unassisted	3.19s	0.14s	0.07s	0.23s	0.43s	0.08s	2.24s	-	-	-
Postedit	2.84s	0.76s	0.20s	0.16s	0.81s	0.13s	0.78s	-	-	-
Options	3.50s	0.21s	0.13s	0.39s	1.03s	0.07s	0.98s	0.62s	-	0.07s
Prediction	5.97s	0.60s	0.21s	0.55s	1.30s	0.49s	2.82s	-	-	-
Prediction+Options	4.64s	0.38s	0.31s	0.61s	1.74s	0.07s	0.46s	1.00s	-	0.07s
User: L2c	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab	mixed
Unassisted	5.82s	0.27s	0.15s	0.52s	1.51s	0.26s	3.11s	-	-	-
Postedit	2.86s	0.61s	0.32s	0.16s	1.02s	0.11s	0.64s	-	-	-
Options	4.60s	0.34s	0.32s	0.49s	1.69s	0.27s	0.50s	0.91s	-	0.08s
Prediction	4.11s	0.24s	0.24s	0.42s	1.46s	0.10s	0.97s	-	0.61s	0.08s
Prediction+Options	2.72s	0.17s	0.16s	0.44s	0.69s	-	0.48s	0.63s	0.08s	0.07s
User: L2d	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab	mixed
Unassisted	3.42s	0.71s	0.09s	0.27s	0.56s	-	1.79s	-	-	-
Postedit	3.10s	0.81s	0.23s	0.14s	1.09s	-	0.83s	-	-	-
Options	4.35s	0.77s	0.15s	0.30s	1.00s	0.33s	1.76s	0.04s	-	0.00s
Prediction	3.83s	0.57s	0.13s	0.37s	0.72s	-	2.03s	-	-	-
Prediction+Options	3.71s	0.55s	0.15s	0.40s	1.10s	-	1.18s	0.32s	-	0.03s
User: L2e	total	init-p	end-p	short-p	mid-p	big-p	key	click	tab	mixed
Unassisted	2.84s	0.28s	0.17s	0.16s	0.32s	0.06s	1.86s	-	-	-
Postedit	2.62s	0.39s	0.25s	0.16s	0.97s	0.12s	0.72s	-	-	-
Options	3.49s	0.14s	0.26s	0.36s	0.56s	0.21s	1.72s	0.21s	-	0.03s
Prediction	2.79s	0.10s	0.32s	0.31s	0.31s	-	1.38s	-	0.30s	0.06s
Prediction+Options	3.01s	0.13s	0.30s	0.18s	0.47s	-	1.94s	-	-	-

Table 3: **Time Spent on Activities.** We break down user actions into a sequence of intervals of specific activities: pause (initial, end, short, medium, big), key strokes, clicking on options, TAB key strokes to accept predictions, and mixed activities (key/tab/click within the same interval). The table shows how much time (measured as seconds per input word) is spent on each activity.

4.4 Analysis of Pauses

One important question that we are trying to answer is: What do translators spend their time on? This has consequences for the design of a translation aid, since we want to alleviate the most time-consuming aspects of the translation process to increase its productivity.

We already included pauses in the analysis above. But strictly speaking, when examining the log of a translator's actions, all we see are pauses interrupted by actions — key strokes and mouse clicks — that take no measurable amount of time. The length of these pauses reveals valuable information about the state of mind of the translator.

We categorized pauses into four categories: Pauses of less than 2 seconds are considered part of an sequence of actions, e.g., the time between key strokes when typing a word. Short pauses of 2–6 seconds indicate some hesitation. Medium size pauses of 6–60 seconds indicate that the translator is thinking and planning her next actions, maybe reading source words or reconsidering some of the already produced output. Longer pauses indicate that the translator is stuck and is trying to solve a difficult translation problem.

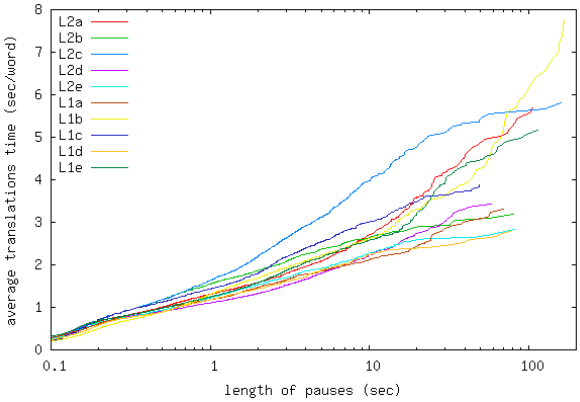
However, the thresholds of 2, 6, and 60 are arbitrary and have no more basis than an intuitive understanding of the translation process. Pauses may be of any length. Instead of classifying pauses into arbitrary categories, we may want to look at the whole range of pauses.

See Figure 5 for an analysis of the pauses of our translators when translating without assistance. Recall that user actions according to our log take no time at all (they happen at specific points in time), and all the time is consumed by pauses between actions. The figure plots on the y-axis the sum of all time periods that last between 0 seconds and the length on the x-axis.

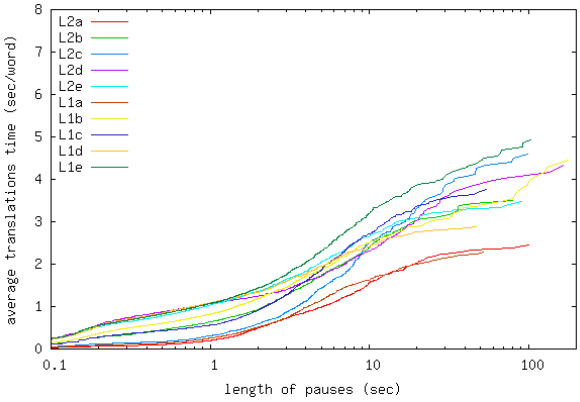
Definition: Sum of Pauses. If P is the set of all pauses p in the translation log and $l : p \rightarrow t$ is the function that maps each pause p to its length in seconds t, then the figure shows the graphs of the function

$$sum(t) = \frac{1}{Z} \sum_{p \in P, l(p) \leq t} l(p) \qquad (2)$$

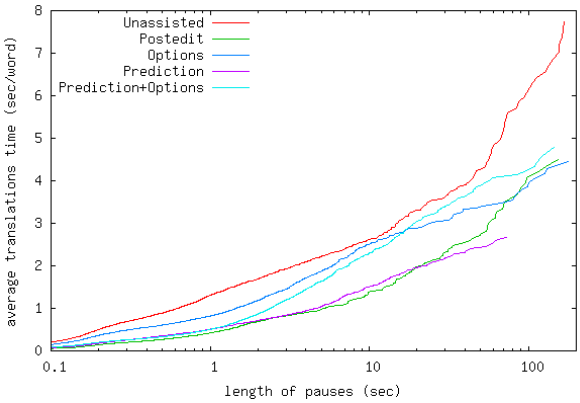
Z is the normalization so that $sum(\infty)$ corresponds to the total translation time per input word that we use in all our other tables. Formally the pauses P are generated when translating a set of input sentences S, and each $s \in S$ has a length of $w(s)$. So, $Z = \sum_s w(s)$.



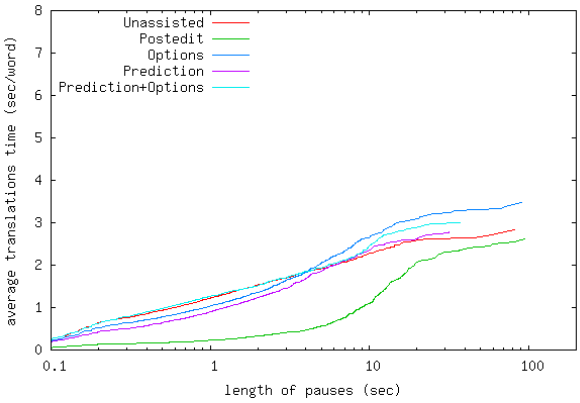
(Figure 5a) Unassisted: All Translators



(Figure 5b) Options: All Translators



(Figure 5c) Translator L1b



(Figure 5d) Translator L2e

Figure 5: **Analysis of Pauses**. Translation time when including pauses of increasing length.

Consider Figure 5a. According to the graph, all translators spend a similar short amount of time in pauses of less than 1 second. Then, the translators diverge. The slowest translator L1b spends about half of their time in pauses of more than 30s. Contrast this to the second slowest translator L2e who spends roughly three quarters of her time in pauses between 3–20s. The fastest translator L2e, who spends hardly any time pausing more than 20s.

This difference in pauses reflects the strikingly different behavior of the translators. Clearly, as mentioned above, the different lengths of pauses indicate different problems the translators are dealing with. We do not yet feel equipped to further qualify the behavior of translators. We are more concerned with the effect the assistance of the tool has on the translation process.

Figure 5c shows the graphs for translator L1b under all five different types of assistance. The graph clearly shows that long pauses during unassisted translation are greatly reduced with assistance. The maximum length of pauses is shortest with the prediction. Figure 5d shows the graph for translator L2e, whose curves, except for post-editing, are almost identical — another indicator that the assistance is not used. When post-editing, most time is taken up by pauses of about 7–20s.

4.5 User Feedback

We requested the translators to fill out a questionary after they completed their translation tasks, and seven did so in time. We ask two multiple choice questions: *Which of the five conditions did you enjoy the most?* Allowing for multiple answers, unassisted was chosen once, post-editing once, options twice, prediction twice, and prediction+options three times.

In which of the five conditions did you think you were most accurate? Post-editing was chosen once, predictions was chosen once, options was chosen twice, and predictions+options was chosen five times. This self-assessment of quality mostly did not match the human judgement, but it was not completely off the mark either.

We also asked the translators to rank the different types of assistance on a scale from 1 to 5, where 1 indicates *not at at all* and 5 indicates *very helpful*. Post-editing received an average rating of 2.9, options a rating of 3.7, prediction a rating of 3.9, and prediction+options a rating of 4.6. It is striking that post-editing was ranked so low, not only in terms of enjoyment, but also in subjective usefulness, while it proved to be as productive as the other types of assistance.

When asked for suggestions for improving the tool, the translators focused on interface issues such as a too small font, being able to finish the translation without clicking the submit button, be able to insert translation options at the cursor position and not just appending them at the end, as well as including a spell checker and grammar checker. Some noted that the options (especially for prepositions) are often wrong and confusing. Some noted the same mistakes over and over again, and should be able to learn from the corrections — as also observed by Macklovitch (2006).

5 Conclusions and Outlook

We described novel types of assistance for human translators and compared them. The study of the human translation process has shown that this assistance improves both speed and accuracy.

Further study of the cognitive processes of translation are needed both to gain insight into what the most time-consuming translation processes are and how they can be alleviated. We are also interested in the varying degree in which Caitra aids novice and more experienced translators. We would like to expand this scale to professional translators on one end and monolingual speakers of the target language at the other end.[1]

References

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Buchweitz, A. and Alves, F. (2006). Cognitive adaptation in translation. *Letras de Hoje*, 41(2):241–272.

Carl, M., Jakobsen, A. L., and Jensen, K. T. H. (2008). Studying human translation behavior with user-activity data. In *NLPCS*, pages 114–123. INSTICC Press.

Fraser, J. (1996). The translator investigated: Learning from translation process analysis. *The Translator*, 2(1):65–79.

Jääskeläinen, R. (2001). Think-aloud protocols. In *Routeledge Encyclopedia of Translation Studies*, pages 269–273. Routeledge.

Langlais, P., Foster, G., and Lapalme, G. (2000). Transtype: a computer-aided translation typing system. In *Proceedings of the ANLP-NAACL 2000 Workshop on Embedded Machine Translation Systems*.

Macklovitch, E. (2006). TransType2: the last word. In *LREC*.

[1]This work was supported by the EuroMatrixPlus project funded by the European Commission (7th Framework Programme).