Chapter 6
The Effectiveness of Consulting External Resources During Translation and Post-editing of General Text Types

Joke Daems, Michael Carl, Sonia Vandepitte, Robert Hartsuiker, and Lieve Macken

Abstract Consulting external resources is an important aspect of the translation process. Whereas most previous studies were limited to screen capture software to analyze the usage of external resources, we present a more convenient way to capture this data, by combining the functionalities of CASMACAT with those of Inputlog, two state-of-the-art logging tools. We used this data to compare the types of resources used and the time spent in external resources for 40 from-scratch translation sessions (HT) and 40 post-editing (PE) sessions of 10 master's students of translation (from English into Dutch). We took a closer look at the effect of the usage of external resources on productivity and quality of the final product. The types of resources consulted were comparable for HT and PE, but more time was spent in external resources when translating. Though search strategies seemed to be more successful when translating than when post-editing, the quality of the final product was comparable, and post-editing was faster than regular translation.

Keywords Translation • Post-editing • External resources • Translation process • Translation quality

J. Daems (✉) • S. Vandepitte • L. Macken
Department of Translation, Interpreting and Communication, Ghent University, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
e-mail: joke.daems@ugent.be

M. Carl
Center for Research and Innovation in Translation and Translation Technology, Department of International Business Communication, Copenhagen Business School, Frederiksberg, Denmark

R. Hartsuiker
Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

M. Carl et al. (eds.), *New Directions in Empirical Translation Process Research*, New Frontiers in Translation Studies, DOI 10.1007/978-3-319-20358-4_6

6.1 Introduction

With the increasing need for faster and cheaper translations due to the increasing amount of text to be translated, computer-aided translation has become more and more widespread. While correcting machine translation output by means of post-editing is now a relatively common task for translators, professional translators are still reluctant to do it, and it is still not clear exactly how regular translation differs from post-editing.

A better understanding of the differences between human translation and post-editing can improve the field of translation in numerous ways. On the one hand, the knowledge can be used to improve translation tools to better aid translators with their work, by indicating in which cases a translator should be allowed to work from scratch, or in which cases he can benefit from the presence of machine translation output. On the other hand, insight in these differences can help understand the reluctance of professional translators to post-edit and can help colleges and universities to teach translation students the appropriate skill sets required for the increasingly technological translation work. Recent studies indicate that certain types of college students would make decent post-editors (Yamada 2015).

In this chapter, we focus on the usage of external resources by student translators translating and post-editing newspaper articles from English into Dutch. For both types of activity, we compare the number and type of resources consulted. We also investigate whether consulting different types of resources and spending more or less time consulting external resources leads to a decrease or increase in productivity and/or quality of the final product.

6.2 Related Work

The field of translation process research is rapidly evolving. Where, originally, rather intrusive methods such as think aloud protocols (TAP) had to be used in order to study the translation process, new tools such as keystroke logging tools and eye-trackers have helped researchers gather data in more ecologically valid ways. The Translation Process Research Database (TPR-DB), which contains over 1300 translation and post-editing sessions, is one example of advanced data collection in the field (see Chap. 2 in this volume). Originally containing Translog data (Jakobsen and Schou 1999; Carl 2012), the TPR-DB has since been enriched with data from CASMACAT (Alabau et al. 2013, and Chap. 3), a state-of-the-art workbench for translation and post-editing, with added keystroke logging capacities.

Yet some aspects of the translation process remain elusive even with these advanced tools. The usage of external online resources, for example, which can provide insights into translators' problem-solving strategies (Göpferich 2010) or uncertainty management (Angelone 2010), is not so easily analyzed. For regular

translation, search queries can be related to source text meaning, meaning transfer or target text production. For post-editing, however, the machine translation output comes into play as well. Whereas the presence of this MT output is intended to facilitate and speed up the translation process, professional translators seem to benefit less from post-editing than translation trainees (Garcia 2011). This could be caused by insecurity about the quality of the MT output, which leads to a higher number of consulted resources, which could, in turn, negatively affect productivity. A better understanding of the usage of external resources during translation and post-editing is needed to obtain a more profound insight into successful problem-solving strategies with regard to quality and productivity.

External resources are usually registered by means of screen capture software such as Camtasia Studio (Göpferich 2010). The drawback of this software, however, is the fact that the data still needs to be replayed and manually encoded for automatic analysis, which can be quite time-consuming. TAP can provide some idea of the resources consulted, but participants' utterances are often incomplete and researchers still need to look at the screen recordings in parallel to make sense of their data (Ehrensberger-Dow and Perrin 2009). Some previous research has made use of data gathered with the TransSearch tool to get a better insight in translators' queries (Macklovitch et al. 2008), but they are limited to one type of resource (TransSearch) and don't take other types of resources into account. The present study attempts to solve these issues by introducing a new method for the analysis of external resources by means of Inputlog (Leijten and Van Waes 2013), a keystroke logging tool originally intended for writing research, which logs all Windows-based applications. In a recent study, Inputlog has been used to analyze the external resources used by a professional communication designer when creating a proposal (Leijten et al. 2014). To the best of our knowledge, Inputlog's logging of external resources has not been used for translation research before the present study. We've opted for a combination of CASMACAT and Inputlog to be able to fully grasp the translation process with external resources. As described in Chap. 2, Sect. 2.7.1, an extra table for the TPR-DB can be created, which accommodates the Inputlog data and allows for a more thorough analysis of external resources, adding an extra layer to the translation process research options the TPR-DB currently provides.

6.3 Methodology

6.3.1 Participants

Participants were ten master's students of translation, who had passed their English General Translation exam. Eight participants were female, two participants were male, and ages ranged from 21 years old to 25 years old. Two participants wore contact lenses and one participant wore glasses, yet the calibration with

the eyetracker was successful for all three participants. Students had no previous experience in post-editing. To prevent exhaustion effects, each session was spread over two half days on different days. Participants received a gift voucher of 50 euros for each half-day session, amounting to 100 euros per participant.

6.3.2 Text Selection

We tried to control for text difficulty as much as possible, as we are mainly interested in investigating differences between post-editing and human translation, and wanted to exclude other potential influential factors. A number of newspaper articles were selected from Newsela,[1] a website which offers newspaper articles at various reading levels, originally intended for use in the classroom. What makes this site so useful is the fact that texts are not just ranked according to existing readability metrics, but that context and the difficulty of a topic is taken into account as well. We selected articles from different topics with the highest possible Lexile® levels (between 1160 L and 1190 L[2]), and selected 150–160 words from each article as potential texts. Lexile® measures are a scientifically established standard for text complexity and comprehension levels, giving a more accurate representation of how challenging a text is than existing readability measures. The scores are usually used in classrooms to provide students with texts of their appropriate reading levels. Our study is—to the best of our knowledge—the first one to apply these measures for translation research. As additional control measures, we then manually compared the texts for readability, potential translation problems and machine translation quality. Texts with on average less than fifteen or more than twenty words per sentence were discarded, as well as texts that contained too many or too few complex compounds, idiomatic expressions, infrequent words or polysemous words. The machine translation was taken from Google Translate, and annotated with our two-step Translation Quality Assessment approach (Daems et al. 2013). We discarded the texts that would be too problematic, or not problematic enough, for post-editors, based on the number of structural grammatical problems, lexical issues, logical problems and mistranslated polysemous words. The final corpus consisted of eight newspaper articles of 150–160 words long, each consisting of 7–10 sentences.

[1] newsela.com

[2] The authors would like to thank MetaMetrics® for their permission to publish Lexile scores in the present chapter. https://www.metametricsinc.com/lexile-framework-reading

6.3.3 Experimental Setup

Each participant translated four texts and post-edited four different texts. To counter fatigue effects, the tasks were performed in two sessions, with two translation and two post-editing tasks in each session. We used a Latin square design to eliminate task order effects, as can be seen in Table 6.1. Across all participants, each text was translated five times and post-edited five times.

We used a combination of logging tools to be able to analyze the translation and post-editing process in detail. Whereas think-aloud protocols (TAP) are often used to elicit problem-solving strategies and other steps in the translation process (Angelone 2010; Ehrensberger-Dow and Perrin 2009), they have been shown to influence the translation process itself (Jakobsen 2003; Krings 2001). We therefore opted to use keystroke logging tools, which are capable of logging the process without interfering with it. The first tool is CASMACAT (Alabau et al. 2013, Chap. 3, this volume), a translator's workbench which doubles as a keystroke logging tool. Unlike other keystroke logging tools, it has the functionality and interface of an actual translator's workbench, allowing for a more realistic experimental setup. In this study, we used a simplified version of CASMACAT, without interactive translation. Another reason for selecting CASMACAT was the fact that it is compatible with the EyeLink2000 eye-tracker. We collected the gaze data with the EyeLink2000 to add an extra layer of information to our other data. Though we will not report on gaze data in the present chapter, it must be noted that a chinrest was used to gather the gaze data, which limited participants' movements, and which could have some effect on our results. In addition to CASMACAT, we also used the keystroke logging tool Inputlog (Leijten and Van Waes 2013). Though Inputlog was originally intended for writing research within the Microsoft Word environment, its capability to log all applications and browser tab information enables us to extract information on the usage of external resources. As CASMACAT only logs what happens within the CASMACAT interface, we needed to add Inputlog to our tool set to analyze the entire translation process, including the usage of external resources.

Table 6.1 Latin square design, mixed text order and task order

Participant		P1	P3	P5	P7	P9	P2	P4	P6	P8	P10
Session1	task1	PE_1	PE_8	PE_7	PE_6	PE_5	HT_1	HT_8	HT_7	HT_6	HT_5
	task2	PE_2	PE_1	PE_8	PE_7	PE_6	HT_2	HT_1	HT_8	HT_7	HT_6
	task3	HT_3	HT_2	HT_1	HT_8	HT_7	PE_3	PE_2	PE_1	PE_8	PE_7
	task4	HT_4	HT_3	HT_2	HT_1	HT_8	PE_4	PE_3	PE_2	PE_1	PE_8
Session2	task5	HT_5	HT_4	HT_3	HT_2	HT_1	PE_5	PE_4	PE_3	PE_2	PE_1
	task6	HT_6	HT_5	HT_4	HT_3	HT_2	PE_6	PE_5	PE_4	PE_3	PE_2
	task7	PE_7	PE_6	PE_5	PE_4	PE_3	HT_7	HT_6	HT_5	HT_4	HT_3
	task8	PE_8	PE_7	PE_6	PE_5	PE_4	HT_8	HT_7	HT_6	HT_5	HT_4

Columns are labeled with participant codes (ranging from P1 to P10), cells contain codes for the task type (*PE* post-editing, *HT* human translation) and text (ranging from 1 to 8)

The first session consisted of the following steps: first, participants filled out an introductory survey, asking them about their experience with an attitude towards post-editing; second, they performed the LexTALE test (Lemhöfer and Broersma 2012) to be able to measure their English proficiency; third, they copied a text of 150 words, so that they could get used to the keyboard and the chin rest of the eye-tracker; fourth, they translated a text in the CASMACAT interface, consisting of four segments that were post-edited and four segments that were translated manually, to get them acquainted with the tool and task; and finally, participants translated two texts and post-edited two texts. For both types of task, the students were instructed to make sure the final product was of publishable quality. Each segment in the CASMACAT interface contained one sentence.

The second session started with another warm up task within CASMACAT, consisting of four segments to be post-edited and four segments to be translated manually, followed by the actual tasks: two texts to be translated manually and two texts to be post-edited. After these tasks, participants had to look at the texts again and highlight the most problematic passages for one translation task and one post-editing task. They were asked to add comments to these passages in a Word document. At the end of the session, participants had to fill out another survey, asking them about their experience and their attitude towards post-editing.

6.4 Analysis

The final dataset consisted of CASMACAT and Inputlog data (xml-files) for all 80 sessions. Using the scripts provided with the TPR-DB, the CASMACAT xml-files were prepared for word alignment. A first, automatic, alignment was done with Giza++ (Och and Ney 2003), which we then manually corrected with the YAWAT tool (Germann 2008). Data from the aligned files was extracted and converted to more manageable table formats with another TPR-DB script (see Chap. 2). From the Inputlog data, we extracted the focus events with the provided software (focus events contain information on the opened application or screen, time spent in the application, and keystrokes). We then manually grouped the different events into categories: dictionary, web search, concordancer, forum, news website, encyclopedia, etc. Figure 6.1 shows an overview of the most common categories for human translation and post-editing. As can be seen, most types of external resources are only sporadically used, with the exception of search engines, concordancers, dictionaries, and encyclopedias. We therefore limit ourselves to these four categories for further analysis, and group the other external resources together in a generic category ‘other’.

A next step was to combine the CASMACAT and Inputlog data for subsequent analysis. Since this is the first study where data from both tools are combined, the TPR-DB had to be updated to accommodate for the new data. An InjectIDFX-script was developed to merge Inputlog data with the CASMACAT xml-files. CASMACAT only logs the keystrokes and events within the CASMACAT interface.

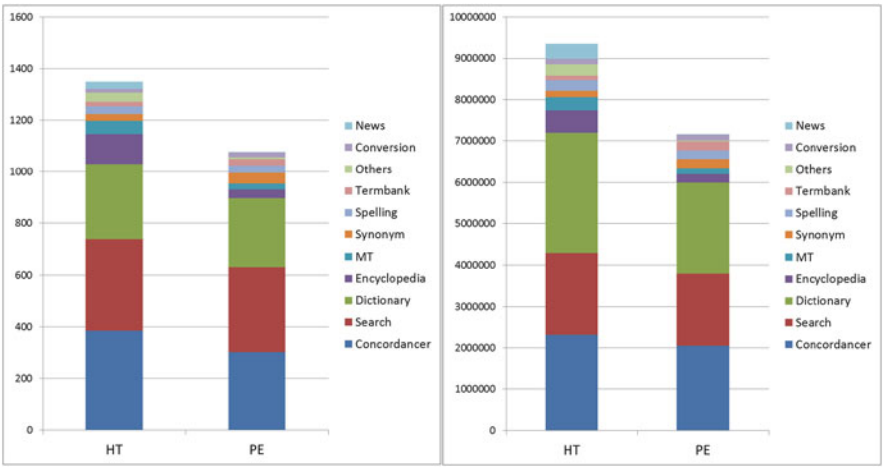

Fig. 6.1 General overview of resource types used in and human translation (HT) and post-editing (PE), expressed in total number of resource hits (*left*) and total duration (*right*) over all 80 sessions

The xml-files themselves contain a 'blur'-event whenever a person leaves the CASMACAT interface and a 'focus'-event whenever they return to the CASMACAT interface, but whatever happens between the blur and the focus-event is unknown. By adding the Inputlog data to the xml-files, we can analyze what happens when a person leaves the CASMACAT interface as well. We added an extra table: the EX-table, containing information on external resources consulted, the time spent in the resource, and keystrokes made within the external resource. We added an extra column to the EX-file where we added the categories we had assigned to the various Inputlog events. An extract from an EX-file can be seen in Table 6.2 below.

Looking at the 'Focus' column and corresponding category label in Table 6.2, we see the participant moving from the main document (CASMACAT, EXid 3) to a new tab in Google Chrome (EXid 4), where he types 'woorden...' (see 'edit'), leading him to the Dutch spelling website 'Woordenlijst' (EXid 5). He then types 'groot-bri' to look up the Dutch spelling of Britain (Groot-Brittannië). After this search, he returns to the CASMACAT interface (EXid 6) for 2 min, after which he again opens a new tab in Google Chrome (EXid 7) for the next search: 'linguee', allowing him to go to the Linguee concordancer (EXid 8), where he looks up the translation of 'in fact' (EXid 9) before returning to the CASMACAT document once more (EXid 10).

It is currently impossible to automatically map external resources to the correct segment. In the data file, there is a column for the last segment that was open before the CASMACAT interface was left, and the first segment to be opened after returning to the CASMACAT interface, but the search itself could be related to either one, or even an entirely different segment. For example, a person can look up a word in a dictionary while translating the first segment of a text. If the person goes back to the CASMACAT interface without closing the screen with the search

Table 6.2 Excerpt from EX-file

EX id	Focus	Time	Dur	ST segN	ST segL	STidN	STidL	KD idN	KD idL	Edit	Category
...											
3	Translate—T1_T5_PE_P9.xlf—204—Google Chrome	−53,975	0	9629	−1	5	−1	0	−1	–	MAIN
4	Nieuw tabblad—Google Chrome	81,778	3360	9630	9629	15 + 16	12 + 13 + 14	122	121	**woorden[..]nlij**	NAVIGATION
5	Woordenlijst Nederlandse Taal—Officiële Spelling—Google Chrome	85,138	3937	9630	9629	15 + 16	12 + 13 + 14	122	121	**groot-bri**	SPELLING
6	Translate—T1_T5_PE_P9.xlf—204—Google Chrome	89,075	123,512	9630	9629	15 + 16	12 + 13 + 14	122	121		MAIN
7	Nieuw tabblad—Google Chrome	212,587	3548	9633	9632	75	70	193	192	**linguee**	NAVIGATION
8	Linguee | Nederlands-Engels woordenboek (en andere talen)—Google Chrome	216,135	2718	9633	9632	75	70	193	192	**n fact**	CONCORDANCER
9	in fact—Nederlandse vertaling—Linguee woordenboek—Google Chrome	218,853	4765	9633	9632	75	70	193	192		CONCORDANCER
10	Translate—T1_T5_PE_P9.xlf—204—Google Chrome	223,618	264,006	9633	9632	75	70	193	192	eed	MAIN
...											

Each time the participant switches to another screen or application, a focus event is recorded, with code EXid and a label found in column 'Focus'. Time is time in ms since the beginning of the session, Dur is the time in ms spent in a particular focus event. STsegL represents the last segment opened in CASMACAT before leaving the tool, STsegN is the next segment opened after returning to the CASMACAT tool. STidL and STidN represent the last source token before leaving CASMACAT and the next token after returning to CASMACAT. KDidL and KDidN contain the ID of the last keystroke before leaving CASMACAT and the next keystroke after returning to CASMACAT. The actual characters typed within a focus event are shown in the column 'edit'. Each focus event is given a corresponding category

query on it, the next time that person opens the search query, this will show up exactly like the search made during the first segment in the data. It would require a lot of extra manual work to label each external resource with the correct segment. In the future, we will try to better map the CASMACAT and Inputlog data by looking at keystrokes or by filtering on the time spent on certain pages. At the moment, however, we grouped the information from the EX-files per session, and not per segment so as to not incorrectly link certain resources to segments. This information was added to the more general SS-file, a table containing an overview of the different sessions. For the different categories (Dictionary, Concordancer, Encyclopedia, Search, and Other) we added a column containing the number of times that resource was consulted in that particular session, and a column containing the time spent in that resource during the session. To be able to better compare the data across all sessions, we normalized the counts and durations by dividing them by the number of source text tokens.

6.4.1 Differences in Usage of External Resources Between HT and PE

Before assessing the impact of the usage of external resources, we wanted to check whether or not there is a difference in the external resources used in regular translation (HT) or post-editing (PE). We used the R statistical software (R Core Team 2014), the lme4 package (Bates et al. 2014) and the lmerTest package (Kuznetsova et al. 2014) to perform a linear mixed effects analysis of the relationship between the total time spent in external resources normalized by dividing by the number of source text tokens, and the type of task (post-editing and human translation). As fixed effect, we entered task. To account for between participant and between text variation, we added intercepts for participants and text as random effects, without random slope. We did test a model with random slope for task, but the slope did not significantly improve the model, so we left it out in the final model. The model with fixed effect was significantly different from the null model without fixed effect ($p = 0.006$), reducing the Akaike's Information Criterion (AIC) value from 1256.8 to 1251.3. AIC (Akaike 1974) is a method designed for model selection, based on a comparison between models. It is shown to have a sound theoretical foundation (Burnham and Anderson 2004). Burnham and Anderson provide the following strategy and rules of thumb when assessing plausible models: the best model is considered the one with the lowest AIC value—in the above case 1251.3—and the plausibility of the model that you compare with it is determined by the difference between both AIC values—in this case the difference between 1256.8 and 1251.3, i.e. 5.5. According to Burnham and Anderson, if the difference is less than 2, there is still substantial support for the model, if the difference is between 4 and 7, there is considerably less support, and models that differ from the best model by more than ten points have basically no support. For the present models, we can

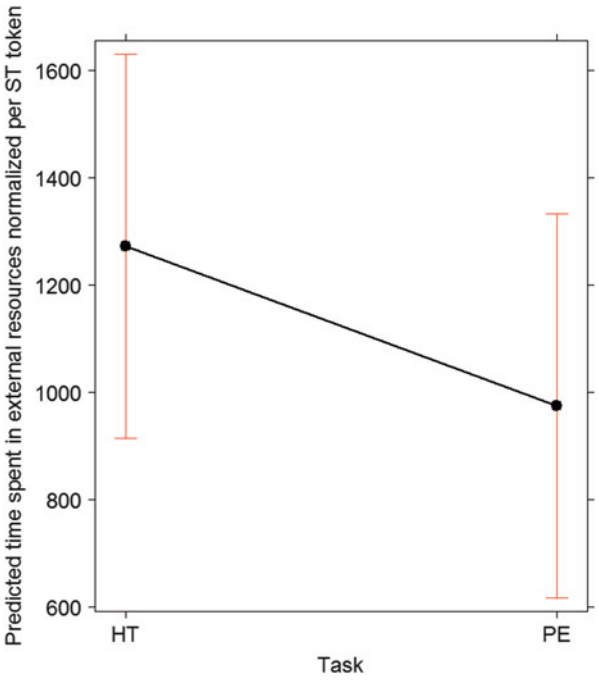

Fig. 6.2 Effect plot of relationship between task (HT = human translation, PE = post-editing) and predicted time (in ms) spent in external resources normalized per ST token. Error bars represent 95 % confidence intervals

conclude that the null model without fixed effects (and AIC value of 1256.8) is not supported enough, so we drop it in favour of the model with fixed effect (and AIC value of 1251.3). The model summary further showed that significantly more time is spent in external resources in human translation, compared to post-editing: about 297 ms $\pm$ 105 (standard errors). The effect plot obtained with the effects package (Fox 2003) is depicted in Fig. 6.2 below. This plot indeed confirms that less time is spent in external resources when post-editing than when translating. Though the confidence intervals in Fig. 6.2 overlap to some extent, this does not affect the statistical significance found (Goldstein and Healey 1995). Visual inspection of normal Q-Q plots indicated right skewed data, which is presumably due to the natural boundary at zero, which is an integral part of the data: It is impossible to spend less than 0 s in external resources, fifty per cent of data points are below 1000 ms, with very few observations above 2000 ms.

In addition to the overall comparison of time spent in external resources, we wanted to check whether the time spent in each type of external resource differed between both methods of translation. We restructured our data of the session summary table (cf. Chap. 2, Sect. 2.3) to be able to perform the appropriate analysis. An excerpt of the new data file can be seen in Table 6.3 below.

Table 6.3 Restructured data for comparative analysis of usage of external resources between human translation and post-editing

Session	Participant	Text	Task	ExternalSource	CountSource	DurSource
P01_P01	P01	T1	P	Dictionary	0.033898305	228,3,785,311
P01_P01	P01	T1	P	Concordancer	0.084745763	369,7,909,605
P01_P01	P01	T1	P	Encyclopedia	0	0
P01_P01	P01	T1	P	Search	0.096045198	417,0,225,989
P01_P01	P01	T1	P	Other	0	0

The column CountSource contains the number of times each resource was consulted during a particular session, normalized per ST token, and the column DurSource contains the time spent in each external resource during a particular session, also normalized per ST token

We again performed a linear mixed effects analysis of the relationship between the time spent in external resources normalized per ST token and the type of task, but this time also in relation to the type of external resource (dictionary, concordancer, encyclopedia, search, other). As fixed effects, we entered task and external resource with interaction term (as we are interested in the combined effect of task and external resource type). Again, we had intercepts for participants and texts, without random slope as random effects (both models were tested, but the model without random slope performed better). The model with fixed effects and interaction was significantly different from the null model without fixed effects ($p < 0.001$), reducing AIC from 5693.7 to 5650.7, but—contrary to our expectations—not significantly different from the model without interaction between task and type of external resource ($p = 0.896$; AIC = 5643.8). The drop1 test showed that none of the predictors (with or without interaction) were significant. We therefore conclude that type of external resource and task are not significantly inter-dependent on each other with regards to the time spent in external resources, even though the overall time spent in external resources was significantly different between human translation and post-editing. The model summary only showed significance for the time spent in encyclopedias and 'other resources'. Both are used significantly less than dictionaries, concordancers and search queries: encyclopedias lowered the duration in the resource per token by about 250 ms (±60 ms), and 'other resources' lowered it by about 150 ms (±60 ms). The effect plot of the model with interaction can be seen in Fig. 6.3. As we can see, there seems to be some trend to spend more time in each resource when translating than when post-editing, but these differences were not found to be significant within the current model.

From these two analyses, we can conclude that overall, the ten translation students spend more time in external resources when translating than when post-editing, though the time spent in each specific resource is not significantly different between the two conditions. In the following sections, we take a closer look at possible effects of the usage of external resources, namely the impact of external resources on overall productivity, and the impact of external resources on the final quality.

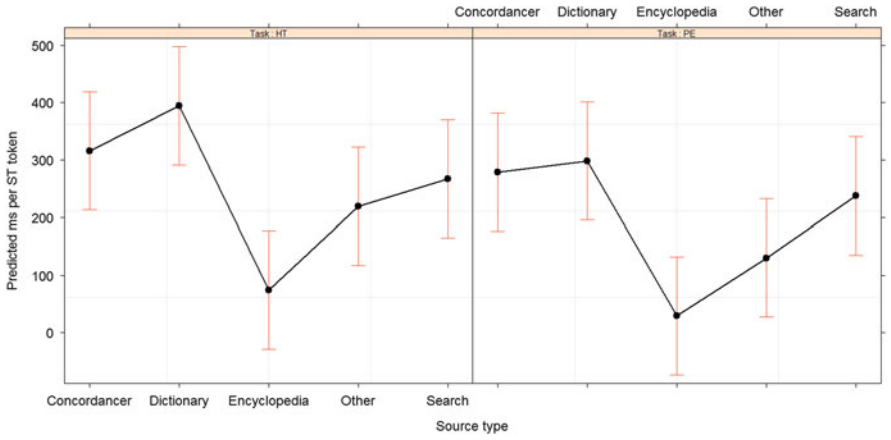

Fig. 6.3 Effect plot of predicted time (in ms) spent in each type of external resource, normalized per ST word, for both task types (*left*: HT = human translation, *right*: PE = post-editing)

6.4.2 Impact of External Resources on Productivity

There are two conceivable ways in which the usage of external resources affects productivity. On the one hand, we can expect total translation time to increase when a person spends more time in external resources, on the other hand, it is possible that the time spent in external resources decreases the overall time needed to translate a text, as a translator looks up external resources to solve problems.

We first take a closer look at the overall difference in time between human translation and post-editing by performing a linear mixed effects analysis. Total time normalized per ST token was taken as the dependent variable, and task as the predictor variable. Intercepts for text and participant were added as random effects. The model with predictor variable performed significantly better than the null model ($p = 0.0116$), reducing the AIC value from 1370.6 to 1366.2. Significantly more time per token was needed for the regular translation task compared to the post-editing task: 523.43 ms (± 202.14; $p = 0.0119$). This effect is visualized in Fig. 6.4 below.

In a next step, we added the time spent in external resources as a predictor, plus the interaction with task, so as to assess the combined effect of task and time spent in external resources on overall time. This model performed significantly better than the model with only task as predictor ($p < 0.001$), reducing the AIC value from 1366.2 even further to 1321.9. However, when we tested the model with interaction against a model without interaction, there was no significant difference, and the model without interaction reduced the AIC value to 1319.9. In addition, the drop1 function showed that only the time spent in external resources was a significant predictor. The AIC value for the final model, which included only the time spent in external resources as predictor, was 1318.6. We can conclude that, even though the

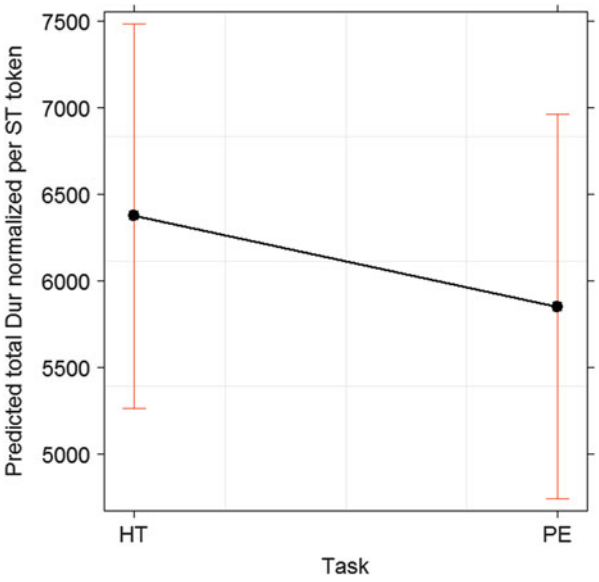

Fig. 6.4 Effect plot of predicted total time (in ms) normalized per ST token for both task types (HT = human translation; PE = post-editing). Error bars represent 95 % confidence intervals

total time, and the time spent in external resources is significantly higher for human translation than for post-editing, the time spent in external resources is a much better predictor of overall time than the task type. The model summary shows that every millisecond spent in external resources per ST token corresponds to a total time per token to increase by 1.348 ms (± 0.145; $p < 0.001$), thus causing us to reject the hypothesis that the time spent in external resources reduces the overall time needed. The effect plot can be seen in Fig. 6.5 below. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality.

6.4.3 Impact of External Resources on Quality

Another crucial aspect to take into account is a text's final quality. Spending more time in external resources (and thus increasing the overall time needed) can be justified if this extra time also brings about an increase in quality. While quality assessment is not always straightforward, we have developed a translation quality assessment approach which allows us to look at the most important problems after translation. It is beyond the scope of this chapter to expand on our methodology, but it has been discussed in more detail in Daems et al. (2013, 2014). The main difference between our approach and other approaches is that we look at acceptability and adequacy as two aspects of quality: quality with regards to the

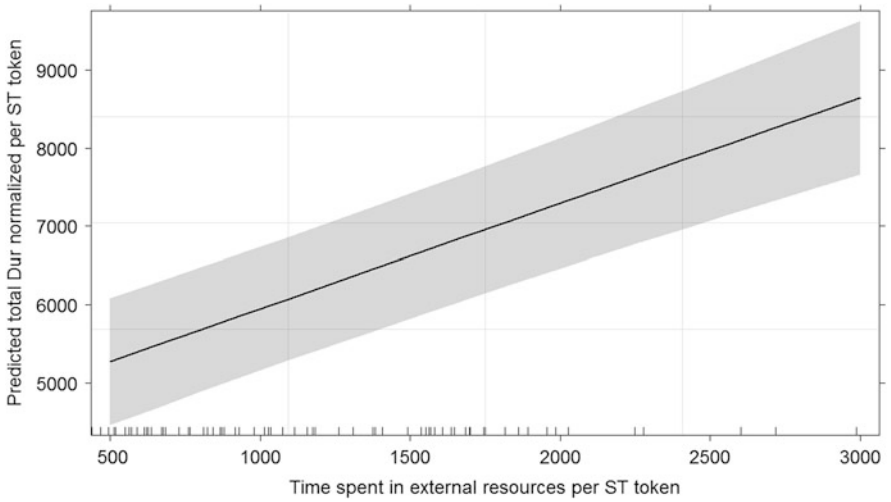

Fig. 6.5 Effect plot of relationship between time spent in external resources normalized per ST token and total time normalized per ST token (both in ms)

final text as a good text in the target language and culture, and quality with regards to the correspondence between source and target text. Acceptability and adequacy each contain various subcategories (such as, for example, grammar, spelling, style and lexicon for acceptability; and word sense, deletions and contradictions for adequacy), allowing for a fine-grained error analysis. Each error category also receives an error weight from zero to four, indicating the severity of the error for the specific text type (for example, a contradiction error receives a weight of four, whereas a capitalization error receives a weight of one). We do also provide an overall quality score. The overall score is calculated by summing up the error scores for acceptability and adequacy and subtracting those acceptability items which were caused by adequacy errors, so as to not penalize the same problem more than once. For example, a word sense error (adequacy) can also lead to a logical problem (acceptability), as is the case in the following situation: The source text contains the verb 'to spend', meaning 'to spend money' (e.g. 'families continue to spend cautiously'), but this is translated as 'doorbrengen' in Dutch, meaning 'to spend time'. The word 'doorbrengen' in this sentence is both a word sense error and a logical problem in the target text. Rather than summing up both error scores in these situations, we only count the error score for the word sense error. Two of the authors highlighted and labeled all errors in the translations, after which we held a consolidation phase where problematic cases were discussed and resolved. Our analyses were conducted on data containing only those errors both annotators agreed on. As with the information on external resources, the error count and score for each category was added to the session file (SS) and normalized by dividing through the number of words in the source text.

6.4.3.1 Overall Quality

Before looking at the effect the usage of external resources has on quality, we looked at the effect of the task on quality. We fit a linear model with normalized total error score as dependent variable and task as predictor variable. In this model, task was not a significant predictor of total error score in itself ($p = 0.669$). We can therefore conclude that there is no significant difference in overall quality between both types of translation (post-editing and human translation).

We then fit a linear mixed effects model to analyze the relationship between overall error score normalized per ST token and the normalized total time spent in external resources. Normalized total error score was the dependent variable, task and time spent in external resources with interaction were added as predictor variables and text and participants were added as random effects, both with random slope for task. This model performed better than the null model without predictors, though only just so ($p = 0.09$), reducing AIC from −306.57 to −306.97, which—according to Burnham and Anderson (2004)—is a negligible reduction. Backward elimination of non-significant effects with the step function showed a significant effect for all variables, with the exception of the slope added to the variable text. In the final model, this slope was left out, leading to a further reduction of the AIC value to −309.59. The main effects of task (post-editing vs. translation) are positive and significant ($p = 0.05$), increasing the average total error score per ST token in the translation condition with 0.035 units (±0.0174). Taking the interaction effect of the total number of external resources into account, however, we see something else entirely. The slope for the time spent in external resources is set at 0.000015 for the post-editing condition (±0.000008587; $p = 0.079$), which is reduced with 0.0000286 points (±0.00001: $p = 0.0118$) in the translation condition. This interaction effect can be seen in Fig. 6.6 below. Inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality.

The differences in slope seem to indicate a difference in the effect of consulting external resources for both types of task. In the case of post-editing, spending a longer time in external resources does not lead to an increase in quality, but rather a decrease, indicating that the resource consulting strategies are not successful. In the case of translation, however, the extra consulted resources do seem to pay off, leading to a decrease in overall error score. This is perhaps not such a surprising result, given that our participants are students with experience in translation, but not in post-editing. It can be assumed that they have developed successful resource consulting strategies when translating throughout their studies, whereas post-editing is a new type of translation, giving rise to different problems, questions, and strategies, which are not always as successful as when translating. We speculate that a possible explanation for these findings can be found in the machine translation (MT) quality. On the one hand, students might be too trusting of MT quality (as evidenced by the fact that less time is spent in external resources when post-editing), on the other hand, they encounter very different problems when post-editing than when translating from scratch, making it hard to find the exact cause of a problem, and—in extension—to decide on the most appropriate external resources

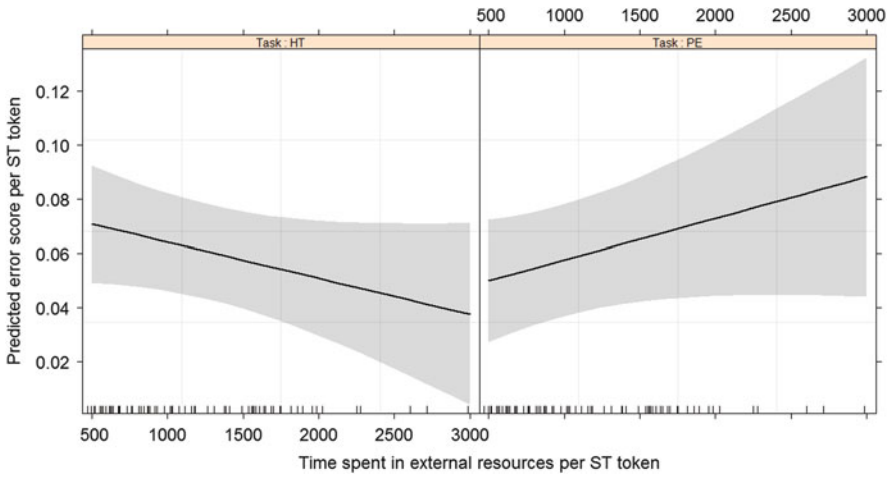

Fig. 6.6 Effect plot of the predicted relationship between time spent in external resources normalized per ST token and overall error score normalized per ST token, for both types of task (*left*: HT = human translation, *right*: PE = post-editing)

to consult. Perhaps the machine translation output primes certain—misguided—search strategies, leading to the students being unable to solve problems even when consulting external resources. Another explanation could be that, when translating from scratch, students look up external resources in sentences that are not so difficult to begin with, which would be reflected in extra time spent in external resources for sentences that already have low error scores.

In addition to this global analysis, we wanted to look at the effect of time spent in the various external resources normalized per ST token on overall quality. We performed a linear mixed effects analysis to assess the relationship between the total error score per ST token and the time spent in the various external resources per ST token. The full model contained the duration of all external resources as possible predictor variables (dictionary, encyclopedia, search, other, concordancer). Text and participant were added as random factors, with added random slope for task. The model with predictor variables did, however, not perform better than the null model ($p = 0.243$), increasing AIC from −309.49 to −306.2. We used the step function from the lmerTest package to assess the necessity of each variable through automatic backward elimination of effects. Only the random effects were significant according to this function. This might indicate that quality is influenced more by differences between texts and differences between participants than the types of external resources consulted. Additional correlation analyses showed no significant correlation between the students' LexTALE proficiency scores and the total error score. We did find a low but significant correlation ($r = 0.296$, $p < 0.01$) between the total error scores and how tiring students perceive post-editing to be. What is remarkable, however, is that the students who perceive post-editing as being less tiring than human translation have higher error scores. This could indicate that

those students are not critical enough: the fact that they perceive human translation as being more tiring could indicate that they struggle with human translation—potentially leading to high error scores—and the fact that they perceive post-editing as less tiring could indicate that they trust the machine translation output too much—again leading to higher error scores. These assumptions warrant further investigation in future research.

6.4.3.2 Acceptability

After looking at quality in general, we took a closer look at our two aspects of quality: acceptability and adequacy, beginning with the first. Inspection of exploratory box plots showed no obvious difference between the acceptability score normalized per ST token for both tasks, which was confirmed by fitting a simple linear model with acceptability error score as dependent, and task as predictor variable. In this model, task was not a significant predictor of the acceptability error score ($p = 0.35$), which is in line with the findings from the overall error score.

We then set out to statistically assess the relationship between time spent in external resources and acceptability error score. We performed a linear mixed effects analysis with normalized acceptability error score as dependent variable and task and normalized time spent in external resources with interaction as predictor variables. Participant was added as a random effect, with added random slope for task. This model, however, did not significantly perform better than the null model ($p = 0.57$). Backward elimination of non-significant effects with the step function showed that none of the predictor variables significantly added to the model. Only participant as random effect with random slope for task was retained, leading us to conclude that neither the overall time spent in external resources nor task type has a significant effect on the acceptability error score, but acceptability error score is most likely influenced by between participant differences. In their 2010 paper, Carl and Buch-Kromann also found no significant relationship between longer translation times and the fluency—which corresponds to our notion of acceptability—of student translators.

The following step was to see whether time spent in specific external resource types had an effect on acceptability error score. We performed a linear mixed effects analysis to assess the relationship between the total acceptability error score per ST token and the time spent in the various external resources per ST token. The full model contained the duration of all external resources as possible predictor variables (dictionary, encyclopedia, search, other, concordancer). Text and participant were added as random factors, with added random slope for task. We used the step function from the lmerTest package to assess the necessity of each variable through automatic backward elimination of effects.

On the basis of this analysis, we again only retained participant as a random effect, with random slope for task, and the duration for dictionary as a predictor variable. This was the only predictor variable found to have an impact on overall acceptability quality. The final model was tested against a null model without

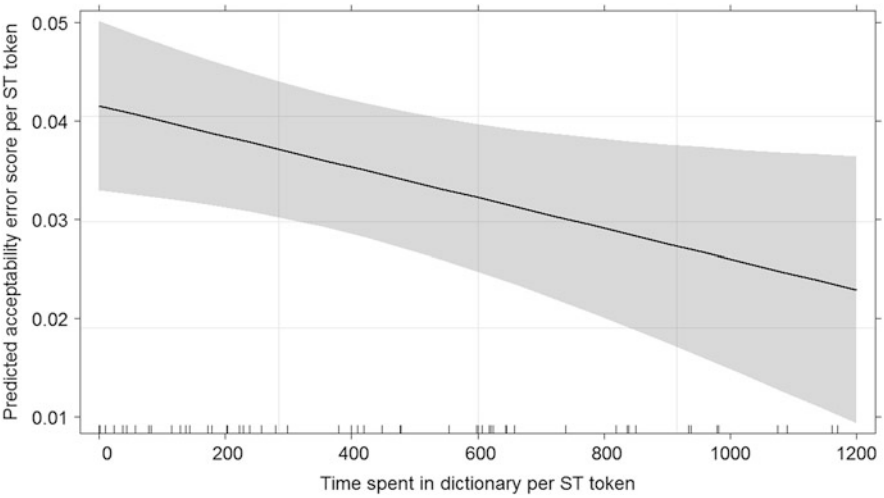

Fig. 6.7 Effect plot of the predicted relationship between time spent in dictionaries normalized per ST token and acceptability error score normalized per ST token

predictor variable, and was found to provide a significantly better fit ($p = 0.01762$), reducing AIC from −384.9 to −388.53.

The effect plot can be seen in Fig. 6.7 below. Residual plots did not reveal any obvious deviations from homoscedasticity or normality. Each millisecond spent in dictionaries affects the acceptability error score per ST token with −0.000016 points (±0.000006). So each second spent to look something up in a dictionary can reduce the acceptability error score for that word with approximately 0.016 units. We can conclude that dictionaries seem to be the only external resource that significantly reduces the acceptability errors made, making it perhaps the most useful resource with regards to acceptability issues.

6.4.3.3 Adequacy

A second aspect of quality is adequacy. We again fit a linear model, this time with normalized adequacy error score as dependent variable and task as predictor variable. As was the case for acceptability, no significant effect was found ($p = 0.527$).

We then performed a linear mixed effects analysis with normalized adequacy error score as dependent variable and normalized time spent in external resources as predictor variable to assess the relationship between time spent in external resources and adequacy quality. Participant and text were added as a random effects, with added slope for task. This model, however, did not perform better than a model without fixed effects ($p = 0.7$), increasing the AIC value from −346.67 to −344.82. Backward elimination of non-significant effects with the step function from the lmerTest package showed only text to be a significant random effect, without

slope. We can conclude that the overall time spent in external resources does not significantly influence the obtained adequacy error score. This finding is in line with the findings by Carl and Buch-Kromann (2010) that there is no notable correlation between accuracy—which corresponds to our notion of adequacy—and translation time.

The next step was to look at the influence of the different types of resources. We applied the same methodology to assess the relationship between the total adequacy error score normalized per ST token and the time spent in the various external resources normalized per ST token. Again, the full model contained the duration of all external resources as possible predictor variables (dictionary, encyclopedia, search, concordancer, other), as well as the task predictor variable. Text and participant were added as random factors, with added random slope for task. We used the step function from the lmerTest package to assess the necessity of each variable.

On the basis of this analysis, we only retained task as a random effect, without random slope. This time, the only predictor that came out of the analysis as having a significant effect on overall adequacy error score, was the time spent in encyclopedias. The final model was tested against a null model without predictor variable, and was found to provide a significantly better fit ($p = 0.04182$), reducing AIC from -352.39 to -354.53.

The effect plot can be seen in Fig. 6.8 below. Residual plots did not reveal any obvious deviations from homoscedasticity or normality. Each millisecond spent in encyclopedia affects the adequacy error score per ST token with 0.000056 points (± 0.000027). So each second spent to look something up in an encyclopedia can increase the adequacy error score for that word with approximately 0.056 units.

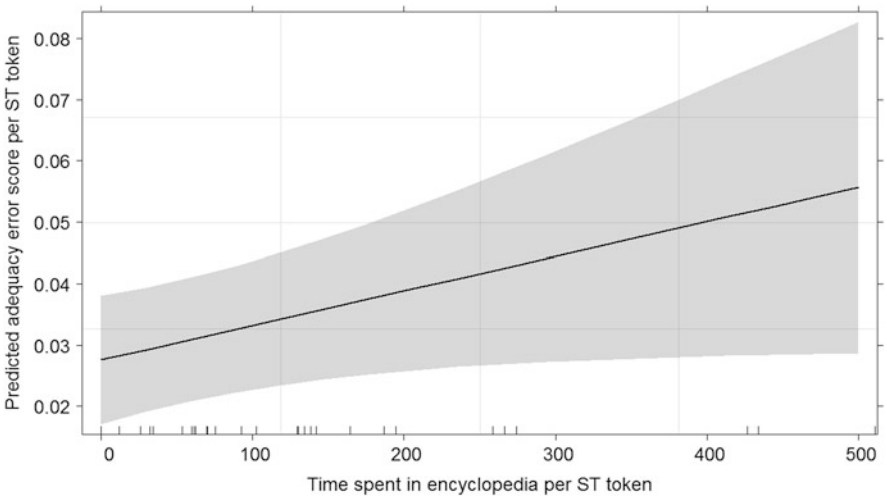

Fig. 6.8 Effect plot of the predicted relationship between time spent in encyclopedias normalized per ST token and adequacy error score normalized per ST token

Of course we do not claim this relationship to be causative. It is presumably not the consulting of the encyclopedia which increases the error score, but the need to consult more encyclopedias can be an indication of the difficulty of the translation. The fact that the effect on adequacy error score is positive might mean that consulting encyclopedias is not always a successful strategy. A possible explanation could lie in the nature of encyclopedias: they provide additional information on a topic, but they do not always provide clues on how to translate terms. Closer inspection of the data shows that sometimes, participants try to look up concepts that are not typical encyclopedia entries, such as 'officially enforced anger'. Additionally, an encyclopedia such as Wikipedia sometimes provides corresponding pages in other languages, but these pages do not always exist or are not always informative. One participant, for example, looked up 'Federal Bureau of Investigation' in Wikipedia, of which the corresponding Dutch page also uses the English term. While the participant spent almost half a minute looking at the Wikipedia pages for 'Federal Bureau of Investigation', this did not help him find an adequate translation. Another participant looked up 'law enforcement agency' and unsuccessfully opened the German page because there was no corresponding Dutch page. The above findings need to be considered with caution, as the overall time spent in encyclopedias is negligible compared to the time spent in other types of external resources (see Fig. 6.1).

6.5 Conclusion

We have conducted a balanced experiment comparing the usage of external resources in human translation and post-editing for general text types, and the effects on time and quality of a text, using a unique combination of state-of-the-art keystroke logging tools. We discussed the addition of Inputlog data to the TPR-DB by means of EX-files (see Chap. 2), containing information on the usage of external resources in a format that is easy to use with the existing TPR-DB tools. This study moves beyond the limitations of previous studies, that either had to make do with manual observation of external resources (Göpferich 2010) or looked at data from within one type of external resource only (Macklovitch et al. 2008).

We found a significant difference in time spent in external resources for both task types (with translation requiring more time). In contrast with our expectations, we found no statistical evidence for the hypothesis that translators use different types of resources, and in different quantities when translating or post-editing, though there seems to be a trend to spend more time in each resource when translating than when post-editing. Significantly less time is spent in encyclopedias and other types of resources compared to dictionaries, concordancers and search engines, for both types of translation.

The overall time needed to translate a text was significantly higher for translation than for post-editing, which is in line with previous findings (Plitt and Masselot 2010). We further found that the time spent in external resources significantly

increases the total time needed to translate a word, indicating that even though the resources might help translators solve translation problems, this goes at the cost of overall productivity. While participants needed significantly more time to translate than to post-edit a word, the effect of time spent in external resources was greater than the effect of the task type.

In a final analysis, we looked at the effect of external resources on the quality of a text. The overall quality of a translation did not seem to be significantly influenced by one specific type of resource, but rather by the overall time spent in external resources, as well as by the task type. When looking at post-editing, longer consultation of external resources was accompanied by higher overall error scores, whereas the opposite was true for human translation, where longer consultation of external resources was accompanied by lower overall error scores. This leads us to believe that participants are more successful in problem solving by consulting different resources when translating than when post-editing. This finding is in line with the suggestion by Yamada that post-editing requires different skills from human translation (2015). With regards to the acceptability aspect of quality, we found no significant difference between human translation and post-editing. When looking at the effect of each type of external resource on acceptability quality, we found that extra time spent consulting dictionaries does bring about an increase in acceptability quality, perhaps making it worth the loss in productivity. With regards to the adequacy aspect of quality, we again found no significant difference between human translation and post-editing. When looking at the effect of each type of external resource on adequacy quality, we found that spending more time in encyclopedias does not bring about a decrease in error score, but rather an increase. This indicates that longer searches do not necessarily lead to better translations with regards to adequacy.

In sum, we can conclude that, whereas search strategies during the translation process are more effective than those used when post-editing, post-editing is still faster than human translation without negatively affecting the final quality of the product.

6.6 Future Work

While the analyses in this chapter have given us a general idea of the effects of external resources and the differences between human translation and post-editing, it might be interesting to look at the texts more closely as well. Due to practical constraints, we performed our analyses on the text level, whereas a more fine-grained approach might give us more practical insights. In the future, we want to better map the resource events to the relevant segments, so that we can perform analyses on the segment level rather than the text level. Taking a closer look at search queries might also provide useful insights in the type of things translators look up in both conditions. Perhaps the external resources used are comparable, but the types of queries are not, or the time spent on each type of query is not.

In addition, we want to take a closer look at the problematic passages as highlighted by the participants and the machine translation quality for the post-editing task. As between participant differences seemed to have a great effect on the results, it can be interesting to perform more in-depth analyses of individual problem solving strategies.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., Martínez, M., et al. (2013). CASMACAT: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics, 100*, 101–112. doi:10.2478/pralin-2013-0016.

Angelone, E. (2010). Uncertainty, uncertainty management and metacognitive problem solving in the translation task. In G. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 17–40). Amsterdam; Philadelphia: Benjamins.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.*http://CRAN.R-project.org/package=lme4

Burnham, K., & Anderson, D. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*, 261–304.

Carl, M. (2012). The CRITT TPR-DB 1.0: A database for empirical human translation process research. In S. O'Brien, M. Simard, & L. Specia (Eds.), *Proceedings of the AMTA 2012 workshop on post-editing technology and practice (WPTP 2012)* (pp. 9–18). Stroudsburg, PA: Association for Machine Translation in the Americas (AMTA).

Carl, M., & Buch-Kromann, M. (2010). Correlating translation product and translation process data of professional and student translators. In *Proceedings of EAMT*, Saint-Raphaël, France.

Daems, J., Macken, L., & Vandepitte, S. (2013). Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE. In *Proceedings of the MT summit XIV workshop on post-editing technology and practice* (pp. 63–71).

Daems, J., Macken, L., & Vandepitte, S. (2014). On the origin of errors: A fine-grained analysis of MT and PE errors and their relationship. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 62–66). Reykjavik, Iceland: European Language Resources Association (ELRA).

Ehrensberger-Dow, M., & Perrin, D. (2009). Capturing translation processes to access metalinguistic awareness. *Across Languages and Cultures, 20*(2), 275–288.

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*(15), 1–27. http://www.jstatsoft.org/v08/i15/

Garcia, I. (2011). Translating by post-editing: Is it the way forward? *Machine Translation, 25*, 217–237.

Germann, U. (2008). Yawat: Yet another word alignment tool. In *46th annual meeting of the association for computational linguistics: Human language technologies; demo session*, 20–23. Columbus, OH.

Goldstein, H., & Healey, M. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 158*, 175–177.

Göpferich, S. (2010). The translation of instructive texts from a cognitive perspective. In F. Alves, S. Göpferich, & I. Mees (Eds.), *New approaches in translation process research* (pp. 5–65). Frederiksberg: Samfundslitteratur.

Jakobsen, A. (2003). Effects of think aloud on translation speed, revision and segmentation. In F. Alves (Ed.), *Triangulating translation: Perspectives in process oriented research* (pp. 69–95). Amsterdam: Benjamins.

Jakobsen, A., & Schou, L. (1999). Translog documentation. In G. Hansen (Ed.), *Probing the process in translation: Methods and results* (pp. 1–36). Frederiksberg: Samfundslitteratur.

Krings, H. (2001). *Repairing texts. Empirical investigations of machine translation post-editing processes*. Kent, OH: Kent State University Press.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2014). *lmerTest: Tests in linear mixed effects models. R package version 2.0-20*. http://CRAN.R-project.org/package=lmerTest

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392. doi:10.1177/0741088313491692.

Leijten, M., Van Waes, L., Schriver, K., & Hayes, J. (2014). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research, 5*(3), 285–337.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44*, 325–343.

Macklovitch, E., Lapalme, G., & Gotti, F. (2008). TransSearch: What are translators looking for? In *AMTA-2008: MT at work: Proceedings of the eighth conference of the association for machine translation in the Americas* (pp. 412–419), Waikiki, Hawai'i, St. Honolulu.

Och, F., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19–51.

Plitt, M., & Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localization context. *Prague Bulletin of Mathematical Linguistics, 93*, 7–16.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Yamada, M. (2015). Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings. *Machine Translation, 29*, 49–67.