The ergonomics of translation tools: Understanding when less is actually more

Martin Kappus and Maureen Ehrensberger-Dow

ZHAW Institute of Translation and Interpreting, Winterthur

Abstract

Professional translation and consequently translation training involve a wide range of general-purpose technological aids as well as specifically-designed translation tools. In response to market demands, a great deal of effort has been devoted to developing and implementing features in such tools that can contribute to higher consistency, productivity and quality. There has been less attention paid to the needs of learners and actual users, however. Previous research with professional translators at the workplace suggests that there is potential to optimize the ergonomics of certain features of computer-aided translation tools currently on the market. An exploratory study of the usability of two such tools with very different interfaces was carried out with students enrolled in a Master of Arts (MA) program in translation. Quantitative measures from process monitoring software and qualitative indicators from post-task interviews highlight how students interacted with the two tools as they translated comparable source texts. The global process measures and the subjective comments suggest that the CAT tool with the leaner interface might be easier for students to use. In-depth analyses of three pairs of comparable segments support this finding, as do the evaluations of the target text segments by a separate cohort of MA students. We consider the implications of our findings in light of recent calls for improved cognitive, physical, and organizational ergonomics at the translation workplace. In conclusion, we make recommendations for incorporating ergonomic considerations into technology modules in translator training.

Keywords: CAT tools; human-computer interaction; interfaces; cognitive ergonomics; translation training

1. Introduction and context

The history of human evolution clearly documents that tool use alters human problem-solving processes as well as the nature of the problems that have to be solved. For example, the invention of the printing press in the mid-15th century and greater accessibility to affordable paper provided the impetus for universal literacy in Europe. Similarly, the speed and modes of written information exchange have changed dramatically since ARPANET[1] first laid the groundwork for the internet in the 1960s (Kleinrock 1961), making life without e-mail hard to imagine today. In the domain of translation, the body of evidence is growing that the introduction of computer-aided translation (CAT) tools into the workplace fundamentally alters the tasks and processes that most commercial, freelance, and institutional translators engage in (cf. Krüger 2016; Mellinger and Shreve 2016).

The current framework for the European Master's in Translation (EMT 2017) recognizes the importance of technology as one of the five main areas of competence. The technology sub-competences concern using and adapting to tools, managing files, handling technologies,

[1] Advanced Research Projects Agency Network

mastering machine translation (MT) basics, assessing the relevance of and implementing appropriate MT systems, and applying other tools in support of translation technology. Interestingly, all of the sub-competences in the EMT framework are formulated in terms of humans coping with the technology rather than developing the meta-competence to choose the tools and settings that meet their needs.

The increasing reliance on technology has serious ergonomic implications for translators, since engaging in complex, bilingual, screen-intensive activity over prolonged periods requires concentration and stamina. In the *International Ergonomics Association*'s terms, ergonomics includes "physical, cognitive, social, organizational, environmental and other relevant factors" of human work and the promotion of conditions that are "compatible with the needs, abilities and limitations of people".[2] Good ergonomic design puts human needs in focus to maximize the usability of technology and minimize unnecessary steps or distractions. Poor ergonomic design can result in inconvenience, cognitive overload, short-term irritation, or even health issues. Although the EMT (2017) refers to ergonomics, it only mentions physical and organizational aspects. The perspective taken in an ergonomic approach to professional translation and training is broader, because it also encompasses cognitive, environmental and social aspects (Lavault-Olléon 2011; O'Brien 2012; Ehrensberger-Dow and Jääskeläinen 2019).

Cognitive aspects of the ergonomics related to translation technology include the design, organization, and operation of user interfaces as well as the number and complexity of functionalities. If CAT tools are easy to use, then more time and cognitive capacity should be available for the decision-making and problem-solving processes integral to translation work. If such tools or certain features are complicated and/or non-intuitive, then human-computer interaction can be compromised, which usually results in less than optimal use and dissatisfaction with tools. This is exactly what research with professional translators has shown. For example, Bundgaard et al. (2016) reported that a translator they observed seemed somewhat resistant to the CAT tool she was using although it was clearly of assistance. Similarly, an international survey that focused on the ergonomics of translation revealed that despite almost all of the CAT tool users reporting that they found them helpful at least some of the time, over half also reported that they were irritated by their tools (Ehrensberger-Dow et al. 2016). This substantiates the finding by Moorkens and O'Brien (2016) that only about half of their survey respondents liked using CAT tools.

Recent research has also suggested that professional translators find the ergonomics of many of their tools less than ideal (Ehrensberger-Dow et al. 2016; O'Brien et al. 2017; Teixeira and O'Brien 2017). If tool settings and features do not align with translators' ways of working, then their flow can be interrupted, their cognitive load increased, and their efficacy compromised. The comments about CAT tools provided in the international ergonomics survey mentioned above were analyzed in more detail in a follow-up study (O'Brien et al. 2017). The most commonly mentioned irritating feature overall was complexity of the user interface, followed by segmentation, formatting issues, visual presentation, and bugs. Other examples from self-report data relate to apparent overload of cognitive resources by the amount of information presented on crowded screens. Many of these irritations could be reduced by adjusting settings, but fewer than half (i.e. 44%) of the translators who used CAT tools reported customizing various aspects, mostly concerning the layout. A between-country

[2] https://www.iea.cc/whats/index.html

analysis suggests that customizing can reduce irritation, due either to empowerment or improved ergonomics or both (Ehrensberger-Dow and Jääskeläinen 2019).

Although published fairly recently, the research discussed above actually reflects the situation of professional translation a few years ago, since the data was collected as early as 2014. In the fast-changing domain of translation, with major advances such as high-quality neural machine translation (NMT) being integrated into translation memory (TM) systems, it would not be particularly surprising if older professionals were having trouble keeping up and might be overwhelmed with crowded screens with a high density of information. Research carried out at the European Commission's Directorate-General for Translation suggests that some of its translators are generally uncomfortable with technology such as machine translation (Cadwell et al. 2016). Drawing conclusions about optimal tool ergonomics on the basis of older professionals' comments and experience might lead to under-estimating the competence and flexibility of younger cohorts with respect to technology. More importantly for our purposes, the relevance of such results for teaching tech-savvy students about CAT tools might be limited. In this article, we therefore focus on the acceptance by MA students of certain new developments in translation tools as well as the possible impact on the translation process and quality of the target texts. In the next section, the new developments are explained in order to provide the necessary background information for the exploratory study presented immediately afterwards.

2. Convergence in translation technology

In recent years, there have been several significant advances in the field of translation technology concerning processes, workflows, and linguistic resources. The availability of new linguistic resources has had an especially notable effect on the way translators work with tools. Two new functionalities stand out, since they have been adopted by virtually all CAT tool vendors. The first is related to further developments in CAT tools, namely the introduction of subsegment matching; the second concerns the improved quality and better integration of machine translation. These features may influence the cognitive ergonomics and the resulting cognitive load that translators experience in two ways. On the one hand, they have affected the quality and quantity of suggestions available to the translators and, on the other, they have changed the way these suggestions are presented. As a result, the task for translators has shifted away from creating their own content (i.e. 'translation from scratch') to selecting the best option out of a number of suggestions from different sources. Furthermore, suggestions are not final but rather are generated and re-generated while translators work, creating an unprecedented amount of information available at any moment.

In the following sections, we explain these two developments and their potential impact with a special focus on two CAT tools with similar functionalities but very different user interfaces (i.e. *SDL Trados Studio 2017*, here referred to as *Trados,* and *Lilt*, an interactive, adaptive translation platform).

2.1. Subsegment matching

Until about ten years ago, CAT tools offered linguistic assistance to translators mainly as segment matches from the TM, where the segments typically correspond to a sentence, and suggestions from the terminology component, typically a word or a short phrase. In both cases, the suggestions are drawn from existing databases. They are computed and presented once the segment is activated by clicking in the target segment field but are not re-computed, even if the translator makes changes to the segment.

Over the course of the last decade, many CAT tools have started incorporating additional linguistic resources based on subsegment matching. The typical use case for subsegment matching is when the match value of the entire segment is below the threshold configured in the CAT tool, but suggestions for parts of the segment might be useful. Based on an analysis of TM content (and possibly other sources), subsegment matches for such fragments are thus shown to the translator. These matches not only differ in size from conventional matches, they also differ in that they can be computed dynamically at the time of typing and presented 'on-the-fly'. The subsegment search is usually triggered by the input typed by the translator (i.e. in auto-complete mode) and dynamically changes depending on the user input. A detailed overview of the concepts behind and the use of subsegment matching in various CAT tools can be found in Flanagan (2015).

In short, subsegment matching has two main consequences for translators: more matches and frequent updating of the matches as they work, substantially increasing the amount of information available that needs to be process. In *Trados*, for example, the suggestions are presented as a list in a 'tooltip' that appears when a translator hovers the cursor over an item in the target text (TT) editor (as shown in Figure 1). The translator can select the best candidate by using the keyboard arrow keys and confirming with the enter key.

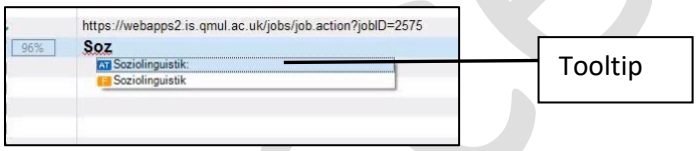

Figure 1. Tooltip display of subsegment matches in Trados

The subsegment suggestions are presented in different areas of the workspace depending on the translation tool. *Trados* has an additional feature called upLift that provides fragment matches in the same area of the screen that the segment-based translation memory matches are presented when no segment matches from translation memory or MT are available. With respect to cognitive ergonomics, it is important to note that subsegment matches and TM matches for the entire segment are displayed in two different areas of the screen in the default window layout of *Trados* (see Figure 2).

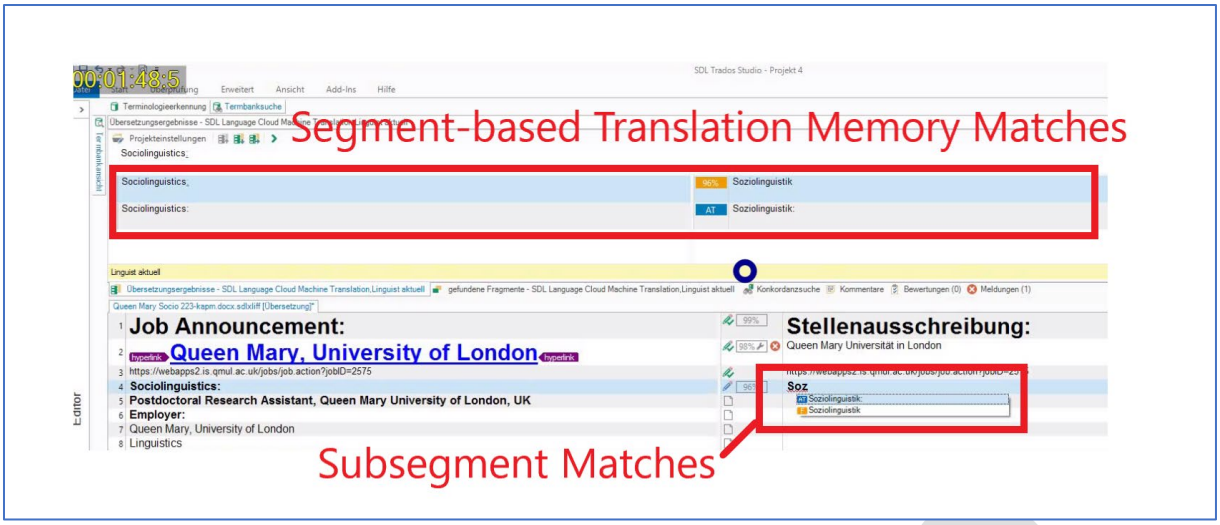

Figure 2. Location of segment-based and subsegment matches in default interface of Trados

In *Lilt,* there is no explicit display of subsegment matching from the TM. However, the MT suggestions are recomputed based on the text entered by the user (see the section below on machine translation in *Lilt*), similar to how classical subsegment matches are computed based on user input.

2.2. Machine translation and CAT tools

As a consequence of the newest approaches (i.e. neural machine translation or NMT), the quality of raw MT output has improved drastically since it was first introduced (Bahdanau et al. 2015), making MT systems more useful for professional translation. The major CAT tool vendors have reacted to this trend by integrating MT systems into their tools and by making the results of MT available to translators in various ways. This integration of MT into translation tools is increasingly replacing the traditional way of carrying out post-editing (Vieira and Alonso 2018). Instead of reviewing and correcting rough versions that are produced after being run through MT and TM, translators are presented with raw MT output in the form of segments in a similar way to traditional TM matches. Due to recent improvements in the fluency of NMT output, these segments may not always be easily distinguishable from TM matches. CAT tools use different mechanisms to include MT results, as described in the next two sub-sections.

2.2.1. Machine translation in Trados

Trados does not have an integrated MT engine, but the tool can connect to a number of MT systems via plug-ins, and MT results are made available in a variety of ways. They can be used in the pre-translation step of project creation, resulting in a document that has either MT output or TM matches for all source sentences. MT output can also be presented in the translation results window where TM matches are typically displayed. MT suggestions are then marked in a different color and labelled *AT* (for 'automated translation'). The suggestions can be transferred to the TT the same way as is done with TM matches. Finally, MT output can be accessed via the tooltip that presents subsegment matches. Suggestions are computed on the fly based on the source text and input from the translator. Various possible translations for the next few words are presented in the form of a list.

2.2.2. Machine translation in Lilt

Lilt also provides access to MT output in the editor window of the user interface, but the modes of presentation and selection are quite different from those in *Trados*. When MT results are presented in *Lilt*, the translator can accept individual words simply by pressing the tab key. Suggestions provided by the MT engine are presented below the line on which the TT appears and are constantly adapted to the emerging target segment produced by the translator (see Figure 3).

Figure 3. Appearance of MT suggestions in Lilt interface

The MT output and the TM matches are displayed in the same area of the screen, but they are marked in different colors so the translator can distinguish them (see Figure 4).

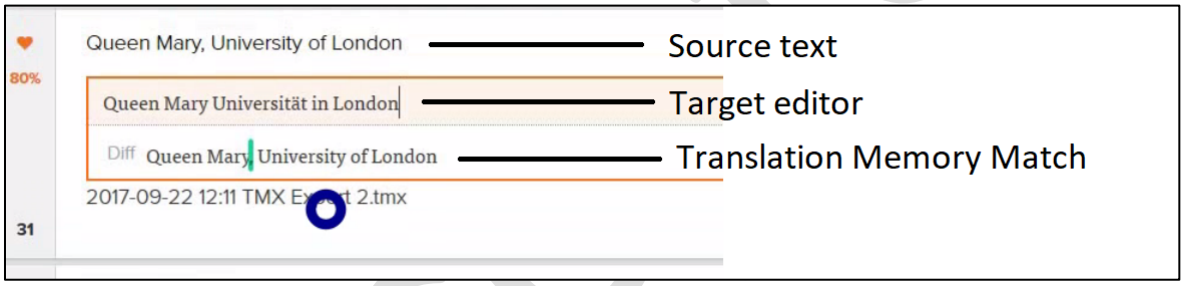

Figure 4.Appearance of TM match in Lilt interface, with percentage match indicated in red

Lilt's implementation of MT differs in two ways from the use of MT in other CAT tools. First of all, its default mode is based on adaptive MT, an approach in which the MT engine is updated using the corrections introduced by the translator. This information is fed back into the MT system after the target segment is confirmed by the translator and serves as training data, changing the output for future computations (for more information about adaptive machine translation, see Bentivogli et al. 2016). Secondly, *Lilt* also uses interactive MT (IMT; see Green 2016 or Peris and Casacuberta 2018), in which the system selects and displays different (initially non-optimal) suggestions updated in real time based on the translator's input, while the translator produces the target segment.

2.3. Presentation of linguistic resources in CAT tools

While both CAT tools under consideration offer a similar range of linguistic resources, the way these displayed in the respective interfaces is quite different (see Table 1) and may present different sorts of challenges to students learning to use them.

Table 1. Overview of linguistic resources displayed in two CAT tool interfaces

Linguistic resource	*Trados*	*Lilt*
TM matches	translation result area	translation result area
MT suggestions (sentence-based)	translation result area	translation result area
Dynamic MT suggestions	tooltip in the target editor	translation result area
Other subsegment matches[3]	tooltip in the target editor	-
Terminology matches	separate window, tooltip in target editor	translation result area

Considering that students' focus of attention during the translation decision-making process is likely to be in the translation result area, the question is whether it is more ergonomic to have as many linguistic resources as possible displayed there (i.e. as *Lilt* does) or in separate areas of the screen (i.e. as *Trados* does). The next section describes how we applied methods familiar from usability testing to address this research question.

3. Empirical comparison of two CAT tools

In the exploratory study reported below, we compare the usability of two CAT tools that have been designed for translation work with MT and TM and are now being taught in some translator education programs, namely *SDL Trados Studio 2017* and the web-based tool *Lilt.*[4] The user interfaces of the two tools differ considerably with respect to the amount of information visible on the screen and the number of functions available. *Trados* has evolved over the course of nine years and has more than 2,000 functions and commands.[5] This range of commands and functionalities is reflected by the large number of menus and menu items as well as by the distribution of relevant linguistic information over several areas on the user interface (see Figure 5). Only a limited amount of information is displayed in the area where the student actively works on the TT segment, which is in the bottom right half or, as the student works further down the text, at the very bottom right of the screen.

[3] The participants in our study were aware of the upLift feature but did not make use of it, probably because there were always TM or MT matches available to them.
[4] Since it is a web-based tool, we cannot provide an exact version number for *Lilt*. The data collection was conducted in October 2017.
[5] D. Brockmann, Product Manager at SDL (p.c.)

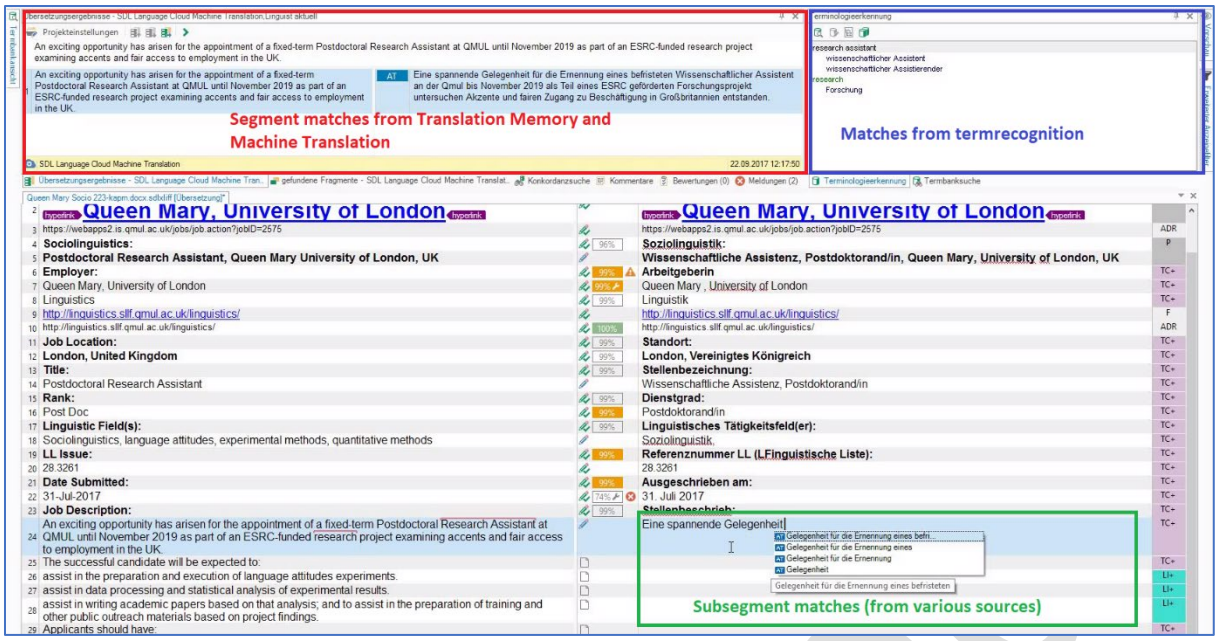

Figure 5. *Trados* interface with areas for various linguistic resources indicated in colored boxes

The other CAT tool under study, *Lilt,* is relatively new to the market and first appeared in 2016. It has a much simpler look and, as described in the previous section, most of the linguistic information is presented in the translation result section where the translator types the TT segment (see Figure 6). Our assumption, based on initial reports from the field (e.g. Zetzsche 2016) and informal pilot tests with professionals as well as with undergraduate students at our institute, was that the leaner interface of *Lilt*, with all the relevant linguistic information available in one area of the interface, could contribute positively to productivity. If less effort is required to actually produce a translation, students who are still becoming familiar with the functionalities of CAT tools can focus their attention on dealing with translation and linguistic challenges.

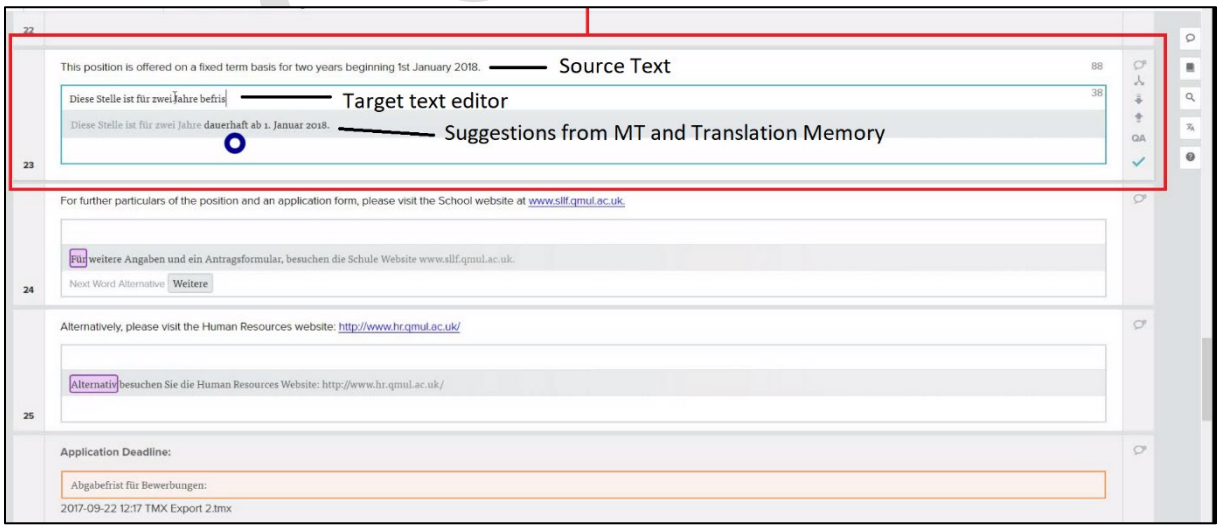

Figure 6. User interface of the web-based CAT tool *Lilt*, with translation result section indicated in red

In order to analyze how the user interfaces affect translation work, we collected various types of process data while MA students completed translations using the respective tools. The testing was done individually in the usability lab in our institute. The quantitative data about the process include the number of deletions, keystrokes, mouse clicks as well as the time spent on the translation and on particular segments. Eye-tracking data were also collected from each participant and used in the form of aggregated heat maps to indicate the primary focus of attention when working with each of the CAT tools. Qualitative data included feedback from post-task interviews with the participants and evaluations of three target segments per text by an independent group of evaluators.

3.1. Participants and their CAT tool experience

Eight second-semester MA translation students who had taken a course dealing with CAT tools participated in the study. The course included an introduction to *Trados* with instructor-led hands-on exercises to explain the typical workflows when translating a document as well as a small-group translation project. *Trados* was also used in some of their translation classes, although without any explicit instruction about the tool. As for *Lilt*, the students had received a short introduction with a hands-on translation task a few days before they participated in the study. Immediately prior to data collection, we asked the participants about their experience and skill level with regard to the two tools (see Table 2). None of the students considered themselves advanced or expert users of either tool, and reported that outside the classroom they used both tools only sporadically.

Table 2. *Self-reported experience and skill levels with each CAT tool (n=8 MA students)*

Skill level / *CAT tool*	*No previous experience*	*Beginner skills*	*Basic skills*	*Advanced skills*	*Expert skills*
Trados	0	2	6	0	0
Lilt	1	6	1	0	0

3.2. Description of stimuli and presentation

The two source texts that the students had to translate were academic job announcements from *linguistlist.org* (see Appendix). The students had a basic familiarity with the genre, since similar texts had been used in classroom exercises in the previous semester. The two texts were always translated using the same CAT tool (i.e. source text 1 with *Trados* and source text 2 with *Lilt*), which is why the texts are referred to this way throughout the article. The texts were comparable with respect to two commonly used readability indices (i.e. Gunning Fog and Flesh-Kincaid; see Table 3). The *Lilt* source text was slightly longer (in terms of words, words per sentence, and number of characters), but the vocabulary density and average word length were both slightly lower than the *Trados* source text.

Table 3. Characteristics of the source texts used in the comparison of two CAT tools

	Trados text	*Lilt text*
Length in characters	1349	1507
Length in words	202 words	233 words
Number of unique words[6]	135 words	145 words
Vocabulary density (unique/total)	0.668	0.622
Average sentence length	16.0 words	19.0 words
Average word length	5.8 characters	5.6 characters
Gunning Fog[7] readability index	14.54	14.53
Flesch-Kincaid grade level	14.7	15.1

To control for potential effects of familiarity and fatigue, the order of presentation was balanced randomly (i.e. half of the participants did the *Trados* text first, and the other half did the *Lilt* text first). In both tasks, a TM with over 6,000 segments and a terminology database containing 55 entries were provided. In the case of the *Trados* setting, its own MT engine *SDL Language Cloud* was activated as a linguistic resource. *Lilt* had its own MT engine, which was activated by default. The participants had unrestricted access to the internet, so they could also use any additional resources they wanted.

Within each text, we defined three segments for more detailed analysis, based on two criteria. The first criterion was the stage of the translation process: one segment appears at the beginning of the source text, one in the middle, and one at the end. The second criterion was the segment length. In each text, we selected two short segments of 8 to 10 words each and a longer segment of 36 words. The short segments were from the beginning and end of the texts, and longer ones were from the middle of each text. The corresponding segments of the two texts were comparable with respect to sentence length, word length, and Flesch-Kincaid grade level (Table 4 and Table 5).

[6] https://planetcalc.com/3205/
[7] http://gunning-fog-index.com/fog.cgi

Table 4. Selected extracts to be analyzed in detail from the *Trados* source text (i.e. Text 1)

Selected extracts from source text translated in Trados		*# words*	*char/ word*	*Flesch-Kincaid*
1	Sociolinguistics: Postdoctoral Research Assistant, Queen Mary University of London, UK	10	7.4	18.9
2	An exciting opportunity has arisen for the appointment of a fixed-term Postdoctoral Research Assistant at QMUL until November 2019 as part of an ESRC-funded research project examining accents and fair access to employment in the UK.	36	5.4	21.7
3	Familiarity with website maintenance and computer programming is highly desirable.	10	7.2	18.9

Table 5. Selected extracts to be analyzed in detail from the *Lilt* source text (i.e. Text 2)

Selected extracts from source text translated in Lilt		*# words*	*char/ word*	*Flesch-Kincaid*
1	Sociolinguistics: Researcher, Queen Mary, University of London, United Kingdom	9	7.3	19.3
2	The School of Languages, Linguistics and Film intends to appoint a Research Assistant to work under the direction of Dr Devyani Sharma on an ESRC-funded research project entitled: 'Dialect Development and Style in a Diasporic Community'.	36	5.5	20.7
3	Alternatively, please visit the Human Resources website: http://www.hr.qmul.ac.uk/.	8	7.6	14.1

3.3. Quantitative analyses

In line with common practice in translation process research (Alves 2003), several quantitative measures were used to determine the effort involved in translating in the two CAT tools. Specifically, effort was calculated in terms of time to produce the TT (i.e. minutes for the entire document and seconds for the individual segments), number of deletions as a measure of regressions, and number of keystrokes and clicks per TT character or word. These measures were expressed in terms of TT at the level of the whole document or at the level of individual segments in order to produce an indication of relative effort, as described below.

3.3.1. Document-level results

To assess the impact of the different user interfaces of the two CAT tools, we first examined measures of effort with respect to the entire TTs. One of the most notable results was the difference in the number of characters or words produced by the students. In the instructions to the study, they had been told that they had about 20 minutes to complete each translation.[8] While we did not stop them after exactly 20 minutes, they all took between 20 and 22 minutes to do each task. The median number of words produced when translating with *Lilt* was significantly higher (i.e. 222, range of 213-227) than the number for the

[8] This estimation was based on the results of pilot testing.

translations produced with *Trados* (i.e. 194, range of 133-206; Z=2.53; p<0.05).[9] Because of this discrepancy, the results below are reported in terms of TT characters or TT words.

We recorded the keystrokes that each student made for each text, using the feature in the *Tobii T60* eye-tracking system. We considered the overall number of keystrokes used to produce the TT as well as the number of keystrokes needed per character in the TT. Six out of eight students used more keystrokes to produce a translation with *Trados* than with *Lilt*. The results in Table 6 show that even though there were more characters in the target texts when translating in *Lilt*, the students needed fewer keystrokes compared to the translations in *Trados*. The reason that the number of keystrokes to produce the TT is smaller than the number of characters in the TT is that matches from the translation memory, terminology, machine translation etc. can consist of units larger than a single character (i.e. words or even an entire sentence). These matches can be selected and inserted by a single keystroke or a combination of keystrokes.

Considering the number of characters in the TT, we find that with *Lilt* the students needed significantly fewer keystrokes for every character produced than when they worked with *Trados* (i.e. a median of 0.81 vs. 0.93; Z=2.36; p<0.05). The pattern and results are very similar with respect to number of keystrokes for each TT word, with significantly fewer needed in *Lilt* than in *Trados* (i.e. median of 7.2 vs. 8.5; Z=2.38; p<0.05).

Table 6. Median of each measure for whole TT and per TT character and word; *p<0.05, significantly fewer in Lilt according to the Wilcoxon signed-ranks test

Measure	*Trados*	*Lilt*
Characters per TT	1777	1918
Words per TT	194	222
Number of keystrokes	1580	1564
Keystrokes per TT character	*0.93*	*0.81**
Keystrokes per TT word	*8.5*	*7.2**
Number of mouse clicks	102	82
Clicks per TT character	*0.06*	*0.04*
Clicks per TT word	*0.52*	*0.37**
Number of deletions	159	105
Deletions per TT character	*0.10*	*0.06**
Deletions per TT word	*0.89*	*0.48**

We also counted the number of times the students clicked the mouse during each task. Similar to the keystroke measures, we found that they needed fewer mouse clicks to translate the *Lilt* source text compared to *Trados* (see Table 6). Taking into account the actual amount of text that was produced in each task, we find that with *Lilt* they made slightly fewer mouse clicks for each character than with *Trados* (i.e. 0.04 vs. 0.06), although the difference was not significant. Calculated in terms of words in the TT, the students needed

[9] Wilcoxon signed-rank test for non-parametric paired-sample data; see https://www.socscistatistics.com/tests/signedranks/default.aspx

significantly fewer mouse clicks per word with *Lilt* than with *Trados* (i.e. median of 0.37 vs. 0.52 clicks per word; Z=1.96, p<0.05).

As an additional measure of effort, we compared the number of deletions that the students made during the two tasks. We counted the keystrokes, using the backspace and delete keys as instances of deletions. When using *Lilt*, the median number of deletions was over 50 lower than when using *Trados* (see Table 6). When these numbers are expressed in relation to the number of characters and words in the target texts, there are significantly fewer deletions made in *Lilt* than in *Trados* for both measures (i.e. Z=2.52, p<0.05 for deletions per TT character and word).

3.3.2. Segment-level results

The overall results at the document level are supported by the analyses of the selected segments, including the supplemental comparison of the precise time spent working on a single unit. Using the eye-tracking data, we determined the number of seconds the participants spent on each segment from the moment they opened it for the first time until they confirmed their translation (see Table 7). In some cases, this included several intervals working on the segment after having moved on to work on other segments in the text, and in other cases, data from fewer than eight participants since there were some data points missing. The results show that for the shorter segments (i.e. segments 1 and 3) the median time spent per segment is considerably lower when working in *Lilt* whereas the difference is not as noticeable for the longer segments (i.e. segment 2).[10]

Table 7. Measures of effort per segment (medians) in each CAT tool

Effort measure	Segment	*Trados*	*Lilt*
Time (seconds)	1	49	33
	2	217	211
	3	63	31
Keystrokes (number)	1	86	28
	2	310	294
	3	94	59
Clicks (number)	1	0	1
	2	16	14
	3	3	4
Deletions	1	3	0
	2	28	19
	3	4	6

The number of keystrokes is lower in *Lilt* across all three segments, with the difference particularly large for the first segment. There is little difference in the number of clicks between the two CAT tools, but there are fewer deletions for segments 1 and 2 of the translations done in *Lilt* and slightly more for the short segment at the end of the text.

Overall, the results for the individual segments are in line with the findings for the whole TTs. However, the advantages of working with *Lilt* seem to be less noticeable in longer segments and segments towards the end of the texts. In future studies, it would be worth investigating

[10] Because of missing data, only descriptive statistics are provided for the segment analyses.

whether this tendency is an effect of the students becoming familiar with the more complex interface of *Trados* during the translation process or whether other factors are at play.

3.4. Qualitative findings

In addition to the quantitative measures of effort presented above, qualitative data were collected and analyzed to derive information about the cognitive ergonomics and relative effectiveness of the two CAT tools for these MA students. The first type of data, drawn from eye-tracking records, provide an indication of the focus of attention compared with accepted good practices in website design. The students themselves also provided information about their preferences in the form of comments in post-task interviews. Finally, the relative success of the translations done in the two CAT tools was assessed by an independent group of the students' peers, as explained further on.

3.4.1. Focus of attention

Although various types of eye-tracking data were collected during the translation processes for both tasks, the decision to maintain relative ecological validity (i.e. using the default settings of the CAT tools and allowing the students to leave the CAT tool interfaces to search for information on the internet) precluded quantitative comparisons of areas of interest on the screen. These would have required inordinate time investments to isolate the scenes with the CAT tools as well as introducing additional sources of measurement error. Instead, the decision was made to compare the focus of attention in the two CAT tools by generating heat maps based on fixation duration aggregated across all eight participants while they were translating the second selected segment of each source text. The heat map for the processes with *Trados* (see Figure 7) shows some degree of attention to the area of the TM segment matches and MT suggestions and a heavy focus in the TT area as well as attention to the source text.

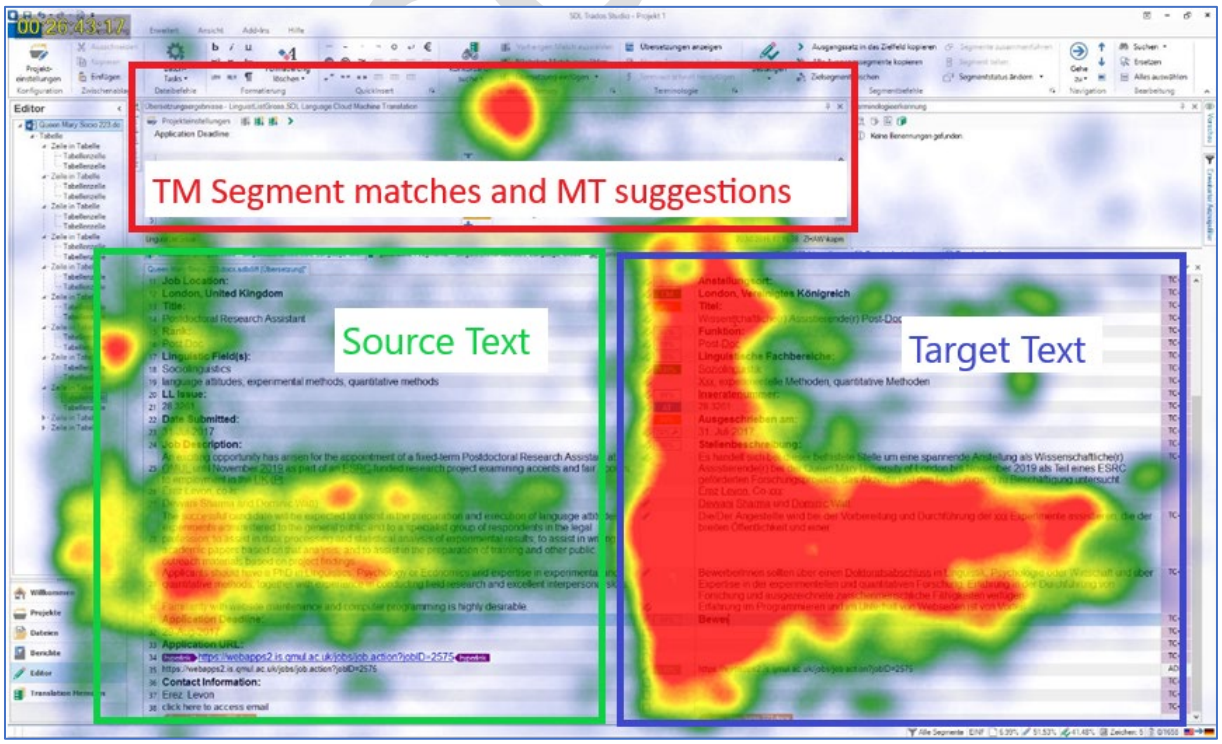

Figure 7. Heat map of visual attention while translating segment 2 of the *Trados* text

The focus of attention in the aggregated heat map for the processes using *Lilt* is much more uniform, with the highest concentration of attention in the upper left third of the screen (see Figure 8). There is relatively little evidence of shifts of attention to other areas of the screen, and the pattern suggests the left-to-right, top-down eye movements typical of reading texts in European languages. This also corresponds roughly to the sweet spot in the upper left area of the screen and the scanning pattern in the 'rule of thirds' recommended for good website design (e.g. Vinh 2011).[11]

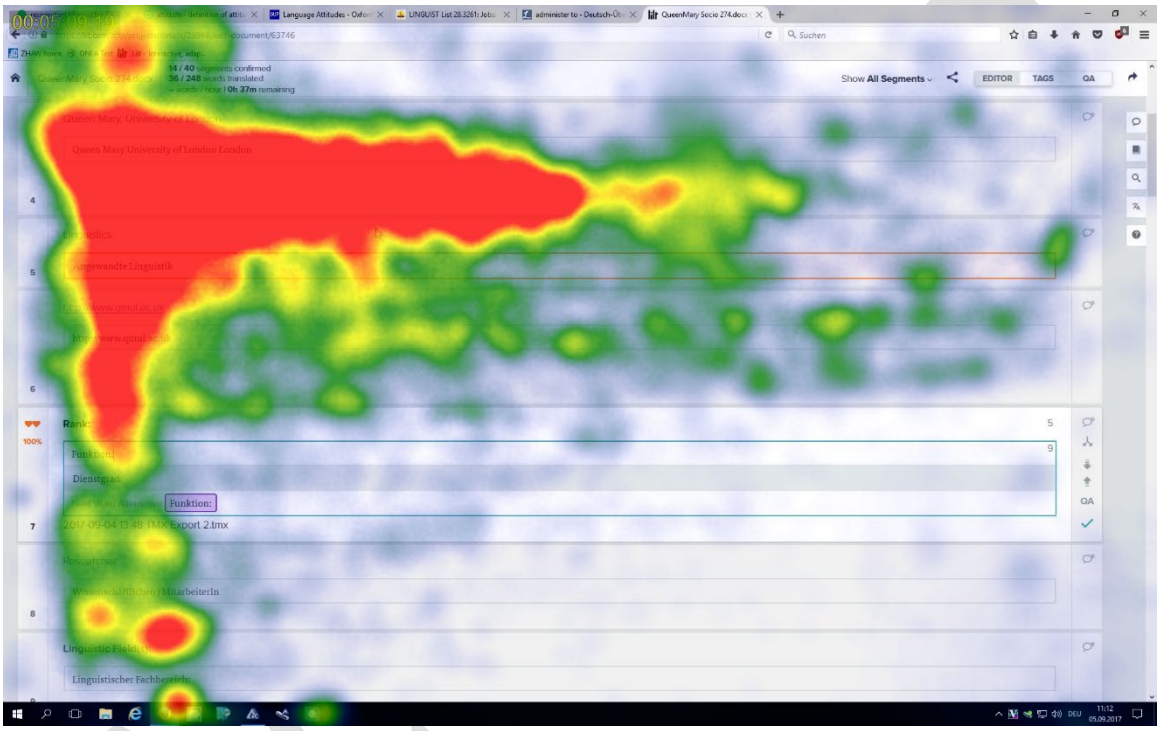

Figure 8. *Heat map of visual attention while translating segment 2 of the Lilt text*

3.4.2. Post-task interviews

Immediately after each student completed both translations, we conducted a short post-task interview in which we asked them about their experience with the two CAT tools.[12] Some of the answers were related to aspects that have little or nothing to do with cognitive ergonomics or the user interfaces (e.g. data security or market share), so we focus here on the replies that were related specifically to the topic of our study. Because of the small sample size and the fact that not all of the students had anything in particular to say in response to some of the questions, the responses are simply summarized.

The first question in the interview was which text the participants found harder to translate. Although the texts were very similar with respect to length, word length, and level as measured by two commonly used readability indices (see Table 3), six of the students judged

[11] See also https://www.interaction-design.org/literature/article/the-rule-of-thirds-know-your-layout-sweet-spots
[12] The interviews were carried out in German; the questions and comments are presented in translation here for the reader's convenience.

the text translated in *Trados* as harder. When asked why, some respondents cited linguistic reasons and two said it had to do with the topic or setting. However, one student explicitly mentioned the other CAT tool as the determining factor (i.e. "*Lilt* shows the complete translation … which can be helpful.").

The participants were also asked which of the two CAT tools is closer to the way they typically work. Five of the seven students who answered this question said that it was *Trados*, probably because they had more experience with that tool (see Table 2), since it contrasts sharply with the responses about productivity in which four out of the five students who answered this question reported that *Lilt* helped them more than *Trados* did. In the next question, we asked which of the two tools supported the participants better in terms of producing the quality they were striving for: two respondents favored *Trados* and three respondents *Lilt*.

When asked to name any advantages of *Trados* compared with *Lilt*, some students mentioned the number of functionalities, the visibility of the entire text, and the table-like arrangement of source and target text in columns. With respect to any disadvantages, half of the respondents mentioned that *Trados* is not "(very) intuitive". We also asked the participants to identify any advantages and disadvantages of *Lilt* compared with *Trados*. Three of the seven respondents who answered this question explicitly mentioned the simple and intuitive user interface as an advantage of *Lilt*, but two participants mentioned that the focus on the current segment in *Lilt* made it harder for them to process the text as a whole.

Summing up, the responses in the post-task interviews suggest that a leaner, simpler interface (e.g. *Lilt*) might be preferable to students both in terms of productivity and self-perceived quality. However, the perceptions of the participants about productivity are less clear than the quantitative results about effort presented in the previous section. While a majority of the respondents judged *Lilt* to be more useful in term of productivity than *Trados*, most said that the latter came closer to the way in which they were used to working. The assessments of the two CAT tools with respect to self-perceived quality are also mixed, which prompted the post-hoc independent evaluation procedure described below.

3.4.3. Evaluations of the target segments

To ensure that any gains in productivity did not happen at the expense of quality, we had the translations of segments 1-3 of each source text evaluated by a different group of students from a later cohort of the same MA program. The student evaluators were asked to assess target segments on a scale from 1-5 according to the fluency and adequacy descriptors often used for the evaluation of MT output (Koehn 2009; see Table 8).

Table 8. Adequacy and fluency scales used in the quality evaluation task (adapted from Koehn 2009)

Rating	*Adequacy*	*Fluency*
5	all meaning	flawless
4	most meaning	good
3	much meaning	non-native
2	little meaning	disfluent
1	none	incomprehensible

The student evaluators were not provided with any information about how the target segments had been translated, but they all had prior experience evaluating translations and were familiar with the scale, mostly in the context of evaluating MT output. The evaluators were presented with all the unique translations for each of the six selected segments (see Table 4 and Table 5). Duplicate target segments were not included; the number of evaluations per segment therefore ranged from 4-7, with a total of 35 evaluations of adequacy and 35 evaluations of fluency made by each evaluator. An analysis of the inter-rater reliability showed higher reliability for the adequacy ratings (Krippendorff's α=0.431) than for the fluency ratings (Krippendorff's α=0.168) but both are unacceptably low to aggregate the evaluations.

A comparison of the average adequacy ratings by each evaluator for each segment shows that in three cases, the segment translated in *Trados* was rated higher than the corresponding segment translated in *Lilt*. In the remaining 21 cases, the translations produced with *Lilt* were rated to be better (see Table 9).

Table 9. Average adequacy rating for each segment by each student evaluator; higher ratings in bold

Translations *Evaluators*	*Trados segment 1 (n=6)*	*Lilt segment 1 (n=4)*	*Trados segment 2 (n=6)*	*Lilt segment 2 (n=7)*	*Trados segment 3 (n=6)*	*Lilt segment 3 (n=6)*
BW1	4.67	**5.00**	3.00	**4.86**	4.83	**5.00**
BW2	**4.83**	4.50	3.17	**4.00**	4.00	**4.50**
BW3	3.67	**4.25**	3.50	**4.00**	4.33	**4.67**
BW4	4.17	**4.75**	3.17	**4.86**	**4.33**	4.17
BW5	4.67	**5.00**	3.33	**4.29**	4.67	**4.83**
BW6	**5.00**	4.75	3.83	**4.86**	4.33	**5.00**
BW7	2.83	**3.00**	3.33	**4.43**	4.17	**5.00**
BW8	3.67	**4.75**	3.33	**4.00**	4.33	**4.50**

A similar pattern emerged for the average fluency ratings: in one case the segment translated in *Trados* was rated higher, in 19 cases the translations produced with *Lilt* were rated to be better, and in four cases there was no difference (see Table 10).

Table 10. Average fluency rating for each segment by each student evaluator; higher ratings in bold

Translations *Evaluators*	*Trados segment 1 (n=6)*	*Lilt segment 1 (n=4)*	*Trados segment 2 (n=6)*	*Lilt segment 2 (n=7)*	*Trados segment 3 (n=6)*	*Lilt segment 3 (n=6)*
BW1	4.17	**4.25**	3.17	**4.14**	3.33	**4.00**
BW2	4.33	**4.75**	3.00	**4.00**	3.50	**4.00**
BW3	3.67	**4.25**	4.00	**4.29**	3.67	**4.50**
BW4	2.83	**3.00**	3.50	**3.71**	3.33	**3.67**
BW5	3.67	**4.25**	3.00	**3.71**	4.00	4.00
BW6	4.50	4.50	3.00	**4.14**	3.50	**3.83**
BW7	**3.33**	2.50	3.50	**4.14**	3.50	**4.17**
BW8	3.50	3.50	3.00	**3.71**	3.50	3.50

The ratings of adequacy and fluency are consistent with the responses in the post-task interview, in which six of the students who had translated texts in both CAT tools reported finding the *Trados* source text harder. If a source text is more challenging, then the students might have difficulty achieving the level of quality that they would be able to deliver with a less challenging text. However, the same could be true of working with a translation tool that is more complex. A simpler tool might ease the task enough for a translation to be perceived as less challenging. Although we tried to control for the comparability of the two texts, only a partial replication of the experiment reversing the assignment of text to tool would adequately address whether the source texts were the cause of the differences in perceived difficulty and quality ratings or whether it was an effect of the tool being used.

4. Implications and conclusions

In this paper, we reported on a comparison between two CAT tools that differ with respect to the amount of information and number of functions available on the screen when the default settings are retained. Usability testing methods were combined with effort measures in order to assess the relative ergonomics of the two tools. According to almost all of the measures, the tool with the less complicated interface (i.e. *Lilt*) seemed to be more ergonomic than the one with the more complicated interface (i.e. *Trados*). The results of the adequacy and fluency evaluations of six target segments suggest that working in a CAT tool in which most of the information is presented in one area of the screen might help students focus on the translation process itself and thereby improve the quality of their translations. The question remains as to whether these two factors are independent of each other or whether the increased productivity when working with *Lilt* leaves students more time to ensure higher quality through review and corrections.

A limitation of the study was the decision to always have source text 1 translated in *Trados* and source text 2 in *Lilt*. It would be advisable to replicate the study with a larger MA cohort with random assignment of source text and CAT tool, but we simply did not have the numbers at the time of data collection. Another factor that could account for at least some of the differences found in this study and that was not controlled for properly was the MT engine. In the domain of translation technology, changes are happening so rapidly that there is little point in collecting data with older versions. The tasks of translating with TM, selecting matches from a variety of sources, and post-editing MT are quickly merging in many professional contexts, so it is imperative to empirically document potential sources of cognitive overload in order to develop ways to best prepare students for working seamlessly between both modalities.

With this study, we hope to have opened the discussion of whether potential ergonomic issues can be reduced by limiting the options available (i.e. 'less is more') at least when students are first learning to use translation technology. A possible consequence for translator training programs would be to introduce students to CAT tools with lean human-computer interfaces so they could concentrate more on the decision-making process than shifting between various areas of the screen. As they become more familiar with the activity of computer-aided translation, they can be taught how to individualize their CAT tool settings to optimize the ergonomics to reduce effort and otherwise suit their own needs.

During the introduction of a new CAT tool (in particular feature-rich tools such as *Trados*), it is important to explicitly discuss the different match types (i.e. TM and MT), their origins, and the way they are presented. Ideally, students should try to systematically activate and de-activate certain match types and re-arrange the setup workspace of the CAT tool they are using in order to explore which matches are the most useful and which settings are the most ergonomic for them. We have started to do screen recordings of students' CAT tool use in class with the goal of identifying inefficient use of the tools and resources that the tools provide. We are planning to include self- and peer-review of the recordings as a component of the course so that the students can identify and monitor their own efficient and inefficient ways of using, adapting, and adapting to CAT tools.

Acknowledgements

We would like to gratefully acknowledge the participation of the MA students and thank colleagues from our institute for contributing to this research and two anonymous reviewers for improving this article.

References

Alves, Fabio, ed. 2003. *Triangulating Translation. Perspectives in Process Oriented Research*. Amsterdam: John Benjamins.

Bahdanau, Dzmitry, KyungHyun Cho, and Yoshua Bengio. 2015. "Neural Machine Translation by Jointly Learning to Align and Translate." *International Conference on Learning Representations, ICLR 2015*, San Diego, CA, arXiv:1409.0473v7.

Bentivogli, Luisa, Nicola Bertoldi, Mauro Cettolo, Marcello Federico, Matteo Negri, and Marco Turchi. 2016. "On the Evaluation of Adaptive Machine Translation for Human Post-editing." *IEEE/ACM Transactions on Audio, Speech and Language Processing* 24 (2): 388-399.

Bundgaard, Kristine, Tina Paulsen Christensen, and Anne Schjoldager. 2016. "Translator-computer Interaction in Action - An Observational Process Study of Computer-aided Translation." *The Journal of Specialised Translation* 25: 106-130.

Cadwell, Patrick, Sheila Castilho, Sharon O'Brien, and Linda Mitchell. 2016. "Human Factors in Machine Translation and Post-editing among Institutional Translators." *Translation Spaces* 5 (2): 222–243.

Ehrensberger-Dow, Maureen, Andrea Hunziker Heeb, Gary Massey, Ursula Meidert, Silke Neumann, and Heidrun Becker. 2016. "An International Survey of the Ergonomics of Professional Translation." *ILCEA Revue de l'Institut des Langues et des Cultures d'Europe et d'Amérique* 27. http://ilcea.revues.org/4004

Ehrensberger-Dow, Maureen, and Riitta Jääskeläinen. 2019. “Ergonomics of Translation: Methodological, Practical, and Educational Implications.” In *Moving Boundaries in Translation Studies*, edited by Helle V. Dam, Matilde Nisbeth Brøgger, and Karen Korning Zethsen, 132-150. London: Routledge.

EMT. 2017. *European Master’s in Translation Competence Framework 2017*. Brussels: European Commission.

Flanagan, Kevin. 2015. “Subsegment Recall in Translation Memory – Perceptions, Expectations and Reality.” *The Journal of Specialised Translation* 23: 64-88.

Green, Spence. 2016. “Interactive Machine Translation: From Research to Practice.” http://www.spencegreen.com/pubs/green.wmt16.pdf

Kleinrock, Leonard. 1961. *Information Flow in Large Communication Nets*. Proposal for a Ph.D. thesis, Massachusetts Institute of Technology.

Koehn, Philipp. 2009. *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Krüger, Ralph. 2016. “Contextualising Computer-assisted Translation Tools and Modelling their Usability.” *Trans-kom* 9 (1): 114-148.

Lavault-Olléon, Elisabeth. 2011. “L’ergonomie, nouveau paradigme pour la traductologie.” *ILCEA* 14. https://journals.openedition.org/ilcea/1078?lang=en.html

Mellinger, Christopher, and Gregory Shreve. 2016. “Match Evaluation and Over-editing in a Translation Memory Environment.” In *Reembedding Translation Process Research*, edited by Ricardo Muñoz Martín, 131-148. Amsterdam: John Benjamins.

Moorkens, Joss, and Sharon O'Brien. 2016. “Assessing User Interface Needs of Post-editors of Machine Translation. In *IATIS Yearbook 2016*, edited by Dorothy Kenny, 109-130. London: Routledge.

O’Brien, Sharon. 2012. “Translation as Human–computer Interaction.” *Translation Spaces* 1: 101-122.

O’Brien, Sharon, Maureen Ehrensberger-Dow, Marcel Hasler, and Megan Connolly. 2017. “Irritating CAT Tool Features that Matter to Translators.” *Hermes Journal of Language and Communication in Business* 56: 145-162.

Peris, Álvaro, and Francisco Cascuberta. 2018. “Online Learning for Effort Reduction in Interactive Neural Machine Translation.” *Computer Speech and Language* 58: 98-126.

Teixeira, Carlos S., and Sharon O'Brien. 2017. “Investigating the Cognitive Ergonomic Aspects of Translation Tools in a Workplace Setting.” *Translation Spaces* 6 (1): 79-103.

Vieira, Lucas Nunes, and Elisa Alonso. 2018. *The Use of Machine Translation in Human Translation Workflows. Practices, Perceptions and Knowledge Exchange*. Bristol: University of Bristol.

Vinh, Khoi. 2011. *Ordering Disorder. Grid Principles for Web Design*. Berkeley, CA: New Riders.

Zetzsche, Jost. 2016. “Lilt: Translation Environment Tool of a Different Kind.” *MultiLingual* 157: 15-17.

Appendix

Source text 1, translated in *Trados*

Job Announcement: Queen Mary, University of London

Sociolinguistics: Postdoctoral Research Assistant, Queen Mary University of London, UK

Employer:	Queen Mary, University of London Linguistics http://linguistics.sllf.qmul.ac.uk/linguistics/
Job Location:	London, United Kingdom
Title:	Postdoctoral Research Assistant
Rank:	Post Doc
Linguistic Field(s):	Sociolinguistics, language attitudes, experimental methods, quantitative methods
LL Issue:	28.3261
Date Submitted:	31-Jul-2017
Job Description:	An exciting opportunity has arisen for the appointment of a fixed-term Postdoctoral Research Assistant at QMUL until November 2019 as part of an ESRC-funded research project examining accents and fair access to employment in the UK. The successful candidate will be expected to: • assist in the preparation and execution of language attitudes experiments. • assist in data processing and statistical analysis of experimental results. • assist in writing academic papers based on that analysis; and to assist in the preparation of training and other public outreach materials based on project findings. Applicants should have: • a PhD in Linguistics, Psychology or Economics • expertise in experimental and quantitative methods, experience in conducting field research • and excellent interpersonal skills. Familiarity with website maintenance and computer programming is highly desirable.
Application Deadline:	29-Aug-2017
Application URL:	https://webapps2.is.qmul.ac.uk/jobs/job.action?jobID=2575
Contact Information:	Erez Levon click here to access email

Source text 2, translated in *Lilt*

Job Announcement: Queen Mary, University of London

Sociolinguistics: Researcher, Queen Mary, University of London, United Kingdom

Employer:	Queen Mary, University of London Linguistics http://www.qmul.ac.uk
Rank:	Researcher
Linguistic Field(s):	Sociolinguistics
LL Issue:	18.3192
Date Submitted:	29-Oct-2017
Job Description:	The School of Languages, Linguistics and Film intends to appoint a Research Assistant to work under the direction of Dr Devyani Sharma on an ESRC-funded research project entitled: 'Dialect Development and Style in a Diasporic Community'. The project examines dialect variation and change within families of Indian origin in London. It will use quantitative methods and discourse analysis to examine changing dialect use in a contact situation. The tasks of the Research Assistant will include: • intensive fieldwork (sociolinguistic interviews, participant observation, coordinating recording) in the Indian community in London;. • processing and archiving of recordings. • quantitative or qualitative analysis of linguistic, discursive and social factors in variation across generations. This position is offered on a fixed term basis for two years beginning 1st January 2018. For further particulars of the position and an application form, please visit the School website at www.sllf.qmul.ac.uk. Alternatively, please visit the Human Resources website: http://www.hr.qmul.ac.uk/
Application Deadline:	16-Nov-2017
Application Address:	Recruitment Administrator School of Languages, Linguistics and Film Queen Mary, University of London Mile End Road London E1 4NS United Kingdom
Application URL:	http://www.sllf.qmul.ac.uk
Contact Information:	Devyani Sharma click here to access email Phone: +44-(0)20-7882-8338