https://doi.org/10.12797/MOaP.24.2018.41.03

Olga Witczak
Adam Mickiewicz University, Poznań, Poland
owitczak@wa.amu.edu.pl
Rafał Jaworski
Adam Mickiewicz University, Poznań, Poland
rjawor@amu.edu.pl

CAT Tools Usability Test with Eye-Tracking and Key-Logging: Where Translation Studies Meets Natural Language Processing

Introduction[1]

Contemporary translators have at their disposal a variety of tools and resources to aid their translation workflow, ranging from general-purpose information searching tools to specialised Computer-Assisted Translation (CAT) environments. Over a decade ago Lagoudaki [2006: 1] spoke of technology-driven software for translators, as opposed to user-driven applications, which feature "an abundance of useless features and a complex, impractical and difficult to learn interface". Giammarresi [2008: 428] wrote that translators were asked for their opinions at the concluding phase of software development, as their experience was thought not to facilitate the engineering process. This has changed over

[1] We would like to thank the reviewers for their insightful comments and suggestions on our manuscript.

the years and CAT developers have begun recognising the essential part that user experience plays in designing such specialised software.

For instance, SDL carried out a Translation Technology Insights survey[2] that collected feedback from translators in 115 countries on work styles, perceptions of quality and productivity, as well as how crucial user experience is. They reported that 66% of respondents described contemporary translation tools as easy to use, compared to 44% in the survey carried out five years before. This could either mean that the tools are becoming more user-friendly (e.g. developers are simplifying interfaces and functions) or people are more used to (this type of) technology. However, regardless of the reason it means that translation technology has become an integral part of the translation process. Ehrensberger-Dow and Massey [2014: 59] described two cases of translator-software interaction, i.e. translators adapting the tool to their needs and translators adapting to its features as they keep utilising it. Moreover, Krüger [2016: 115] suggested that young translators who attended university courses on translation technology are likely to think of using CAT tools as the default in the translation process rather than an exception. Therefore, the present study takes into account the needs of potential users at the early stage of the translation aid development process. To that end, we designed a usability experiment in a laboratory setting, which involved translation students as feedback providers. As theoretical framework, we used Situated Translation [Risku, 2004] and ISO 9241-11 Usability Model [2011] definition of usability [cf. Krüger, 2016], relying on effectiveness, efficiency, and satisfaction in a process-oriented experiment. The tool that underwent testing was Concordia [Jaworski, 2015].

1.1. Concordia

Regular translation memory searching scheme involves using the whole sentence as a pattern and retrieving the most similar sentences from the memory. This may result in omitting valuable fragments. For instance, let us suppose that the search pattern is a sentence *S*, and the memory contains a sentence *C1SC2*, where *C1* and *C2* are contexts of considerable length. This match is either retrieved with a low score or not returned at all.

[2] <http://blog.sdltrados.com/user-experience-crucial-translation-technology>, visited on October 27, 2017.

In order to overcome this shortcoming, this article introduces a new translation memory search algorithm, Concordia, inspired by another CAT technique - concordance searching. The main difference between Concordia and a standard concordancer is the fact that Concordia search uses the whole sentence as a pattern and returns all fragments that cover it. The search algorithm is based on a suffix array index in order to ensure fast lookup times.

Let us consider an example illustrating the Concordia search procedure. Let the index contain the following sentences:

- Alice has a cat *(id=56)*
- Alice has a dog *(id=23)*
- New test product has a mistake (*id=321)*
- This is just testing and it has nothing to do with the above *(id=14)*

The results of Concordia searching for pattern: "Our new test product has nothing to do with computers" are presented in the below table:

Table 1. Example Concordia lookup results

Sentence id	Matched fragment
14	has nothing to do with
321	new test product has
14	nothing to do with
321	test product has
14	to do with
56	has a
23	has a

Concordia automatically identifies the longest non-overlapping fragments of the search pattern found in the index ("new test product has" and "nothing to do with"). Concordia search thus makes up for standard translation memory lookup shortcomings. If the above search was performed using standard translation memory search techniques, it would probably return the results: "New test product has a mistake" and "This is just testing and it has nothing to do with the above" with low resemblance scores. These low-scored matches would probably be discarded by the CAT system (as falling below a given threshold) or ignored by

the translator because of insufficient similarity to the pattern. Concordia search results, on the other hand, draw the translator's attention to the coverage of specific fragments of the pattern.

1.2. Situated Translation in a lab: usability testing and CAT tools

From the viewpoint of Situated Translation [Risku, 2004], a human translator and their environment co-create the cognitive translational ecosystem [Strohner, 1995: 56], so that they are influenced by the tools they regularly use for translation, i.e. the situational factors involved in the process [Krüger, 2016: 117]. Thus, the translation process can also be referred to as translator-information interaction, as any tool utilised for translation is likely to have high cognitive relevance [Zapata, 2016: 136].

According to Hutchins [1995, as quoted in Risku and Windhager, 2013: 36], cognition is, similarly to an aircraft cockpit, "an interplay of multiple dynamic systems, i.e. pilots, instruments, aircraft, ground control and the surrounding airspace", which can be understood analogically in the context of translation in the sense that it is a sum of a variety of factors. Translation is therefore "a highly complex problem-solving process embedded in social and physical environments", which is scaffolded by artefacts that include electronic aids [Risku, 2002]. These artefacts are encompassed by translation technology in the narrow sense (TM tools) and in the wider sense, i.e. "the many tools that are part of modern translation work". These include, for instance, text processing software or online resources.

A key concept for the present study is also the notion of process-oriented usability which denotes "[t]he extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specific context of use" [ISO 9241-11, 2011; Krüger, 2016: 130]. In the context of usability research, CAT tools with high usability should be able to facilitate cognitive processing during such a complex task as translation, while those with low usability will likely cause cognitive friction [Krüger, 2016: 116, cf. Ehrensberger-Dow and O'Brien, 2015:102, Cooper, 2004: 19].

Thus, in line with the cognitive view on the translation process, the way translation aids are designed influences how this process proceeds. Translation memory is one of the core concepts behind modern CAT tools, as translation workstations, such as SDL Trados Studio, are even

commonly referred to as TM systems. Therefore, bearing in mind the interconnectedness of humans and artefacts in a translation environment, what is crucial to both designing a translation aid and testing its usability is that: "[t]he tool was developed according to a particular view of the cognitive process of translation, and no matter what the real cognitive processes, it has the potential to discreetly but firmly guide the cognitive processes in that direction" [Risku and Windhagen, 2013: 38]. In the present study, we intended to examine how translators interact with a new translation aid, thus enabling further development of the aid, informed by the needs of users.

Furthermore, despite the emphasis on ethnographic and naturalistic studies that Risku and Windhager introduce in the context of situated and distributed cognition research, laboratory settings can be beneficial in examining the interaction with translation artefacts such as translation aids. In fact, as O'Brien put it:

> More experimental studies of translator-tool interaction could be carried out using formal usability research methods such as screen recording, eye tracking, and observation, the results of which could then be used by translation technology developers to improve the specifications of tools for the benefit of translators and, ultimately, the end users of those translations [O'Brien, 2012: 116–117].

To date, user-oriented usability research on CAT tools includes Höge [2002], Dragsted [2004], Lagoudaki [2006 and 2009], Guillardeau [2009], Dillon and Fraser [2006], Dragsted [2006], Colominas [2008], Christensen and Schjoldager [2011], Campbell, Weyland *et al.* [2013]. Most of them deal with CAT tools in the narrow sense, and therefore there clearly exists a gap in the process-oriented research on tools that are not integrated workstations. Laboratory setting is beneficial for gauging translation aid usability. Most importantly, it is possible to avoid confounds that are inherently a part of a naturalistic setting, as all participants are subjected to the same conditions and procedure. In our study Translation Process Research converges with Natural Language Processing, so that the development of new tools can be informed by empirical research.

To gauge usability, we used effectiveness as the qualitative aspect of the process, capturing how well the user managed to perform the experimental task [Krüger, 2016: 131]. Effectiveness was measured with an

accuracy score based on their terminological choices during the experiment. Another element of usability evaluation is efficiency which Krüger [2016: 131] defined as the amount of effort put into the task. In our experiment, we triangulated the effort measurement into three types, i.e. technical, temporal, and cognitive [cf. Krings, 2001, for detailed description of methodology, see Section 2.6.1.]. Both effectiveness and efficiency are objective measures of usability. However, there is also a subjective dimension involved, i.e. satisfaction. User satisfaction is simply "the way users feel about working with the software" [Krüger, 2016: 131, cf. Rudlof, 2006: 15] and we gauged it through a standardised Software Usability Scale [Brooke, 1996], which was adapted to evaluate Concordia. As per Rudlof's [2006: 15] and Krüger's [2016: 131] general model of CAT tool usability, usability is an *in vivo* quality and all three dimensions are investigated while the user pursues a specific experimental task.

2. Experiment

2.1. Aims and hypotheses

We aimed to investigate the usability of Concordia as a translation aid, i.e. whether Concordia facilitated the efficiency and effectiveness of translation as well as elicited satisfaction from its users. We hypothesised the following:

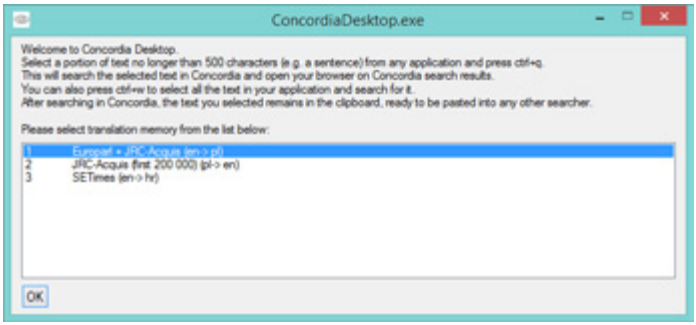
H1. A smaller amount of effort will be generated in the conditions involving Concordia when compared to the condition without Concordia
H2. Accuracy scores will be better for the conditions involving Concordia when compared to the condition without Concordia
H3. Users will be satisfied with Concordia.

2.2. Pilot study

A pilot study was conducted with one participant in order to test the design and carry out a pre-evaluation of Concordia. A PhD student from the Faculty of English (specialisation in linguistics) and freelance translator was recruited to participate. The procedure of the experiment was the same as later in the study proper (see the subsequent sections detailing the materials and methodology). Based on the questionnaire answers, the participant expressed a clear dissatisfaction with the tool, which is why

more texts were added to the database and a special desktop widget (Figure 1) was created that featured an explanation of what can be done with Concordia and what the shortcuts allowed.

Figure 1. Concordia Desktop widget



2.3. Participants

The participants of this study were six students of written Polish <> English translation (5 from the same group and specialisation, second year of a 2-year Master's programme at the Faculty of English, Adam Mickiewicz University in Poznań, while one was an exchange student from a translation programme at a different Polish university). Three took the course on translation of EU texts, but all of them knew about available EU online resources for terminology.

All six participants are highly proficient in English, as their LexTALE scores were all indicative of C1/C2 proficiency level [Lemhöfer and Broersma, 2012: 341], all falling within the accuracy range of 80–100% (*M*=90.42%; *SD*=6%). Their typing skills were also tested in a typing task, in which they were asked to copy a text from the same domain. Their text production per minute was on average 171 characters (*SD*=55) during the typing task, while during translation tasks their speed ranged from 40 to 99 characters per minute (*M*=56, *SD*=13 for all conditions). This shows that the participants were not equally fast when it comes to typing during both tasks and there is some degree of individual variation among them.

2.4. Materials

The source texts used in the present study were balanced in terms of readability (FRES [Flesch, 1948], Gunning Fog index[3] [Bond, 2016]) and word/character count, as per Table 2.

Table 2. Readability scores for source texts

Readability score:	**TEXT A**	**TEXT B**	**TEXT C**	**MEAN**	**SD**
Gunning Fog index	17.01	16.01	18.36	17.13	1.18
Sentences	4	4	4	4	0
No. of words	79	83	87	83	4
Characters	445	427	459	443.67	16.04
Characters per word	5.5	5.0	5.1	5.2	0.26
Flesch Reading Ease	31.5	33.9	32.0	32.47	1.27
No. of key phrases	10	10	9	9.67	0.58

Key phrases selected for analysis later were unique for each text so that participants were not advantaged or biased by their previous searches (except for *[European] Commission* found both in text A and C, which was not counted in the analysis). Three texts were selected for the following three conditions, featuring translation with access to:

1. Internet resources only
2. Concordia only
3. Both Concordia and Internet resources

Out of 10 or 9 key phrases 6 could be found in Concordia when searched as a whole phrase, not each word separately (i.e. the search yielded meaningful results, albeit not necessarily accurate).

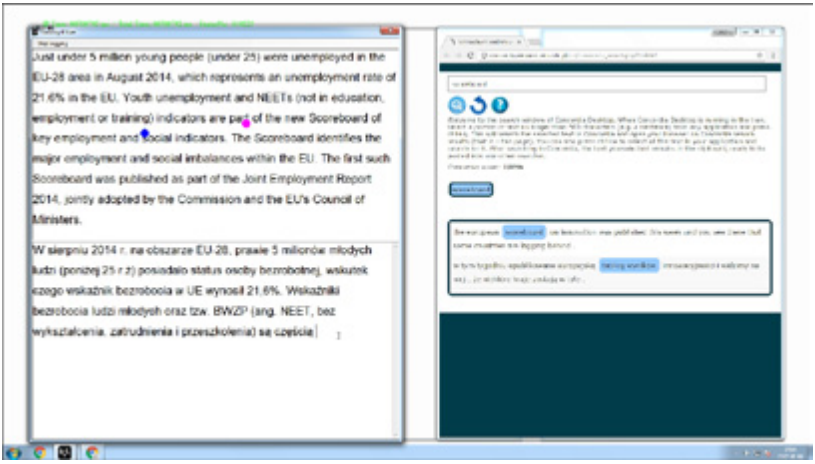
2.5. Methodology and procedure

The experiment was conducted in a laboratory setting at the Faculty of English, Adam Mickiewicz University. Data collection was triangulated via eye-tracking, screen-recording, and keystroke logging methods as well as supplemented with a usability questionnaire. The equipment and software used for this study was EyeLink® 1000 Plus, Morae Recorder, Translog-II, Concordia Desktop, and Google Chrome.

[3] The Gunning Fog index was calculated via http://gunning-fog-index.com, while all other readability statistics were calculated in MS Word (spell check function).

The experiment started with a typing task so that the participants could acquaint themselves with the keyboard. They retyped a text from the same domain, which did not contain the same terminology that the experimental texts did. Then participants were calibrated for eye-tracking and commenced the translation task for text A in Translog-II in Condition 1 (Internet only). After another calibration, they proceeded to translate either text B or C (they were counterbalanced) using only Concordia as a translation aid, first reading the instructions. The third translation task (text B or C) involved both Concordia and other Internet resources. Finally, participants filled in the questionnaire about their translation background and favourite resources as well as impressions of Concordia. They also took the LexTALE test. Figure 2 depicts the experimental setup during a recording session.

Figure 2. Experimental setup: Translog-II and Concordia in Google Chrome.



Texts B and C were purposefully chosen for translation involving Concordia, as around 60% of the key phrases in both of those texts could be found in Concordia. Therefore, the attempts at exploring Concordia functions would require a more complex interaction than if everything had been found in the database. We were interested what actions would be taken to elicit that information from Concordia. Both texts were similar in terms of readability and the number of targeted phrases, which is why

it was the texts that were counterbalanced rather than conditions. Counterbalancing conditions instead of texts would have allowed participants to use Concordia with other Internet resources before exclusively using Concordia, which would have diluted the interaction with this translation aid and highly distorted the results.

2.6. Data analysis and results

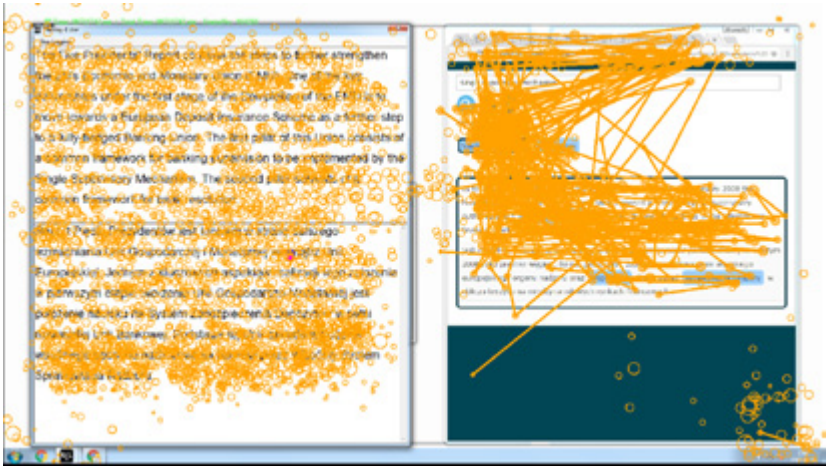
2.6.1. Efficiency

As mentioned in Section 1.2., we defined the efficiency dimension of usability as consisting of three effort types [Krings, 2001], i.e. technical, temporal, and cognitive. Technical effort was operationalised as total user events which were captured by Translog-II during the recording session. The recording events included such actions as insertions, deletion, and navigation. As per Table 3, there was no effect of Concordia use on technical effort for three conditions. But a supplementary analysis revealed an effect of Concordia use on technical effort at the $p<.05$ level for two conditions, i.e. Concordia (Internet + Concordia and Concordia only) and No Concordia [$F(1, 16) = 4.56$, $p=0.049$]. Temporal effort was measured as total task time and time spent in browser, however, there was no effect of Concordia use on this type of effort for any conditions at the above-mentioned *p*-value.

When it comes to cognitive effort, measured via total fixation time (dwell time), in line with the eye-mind assumption we anticipated an approximation of cognitive processing reflected in the visual focus [cf. Hvelplund, 2014: 209]. In the context of complex tasks such as translation, most eye movements can be assumed to be synchronous with cognitive processing as "there is arguably little room for much mind wandering" [Hvelplund, 2014: 210]. Also, simple or automatic tasks are more likely to trigger mind wandering than tasks that demand a participant's attention [Smallwood and Schooler, 2006: 947, 956]. We used fixations as measures of eye movements in order to quantify cognitive effort. A fixation is a period of time when the eye remains relatively stable and longer ones are thought to indicate more effort involved in the processing [Duchowski, 2007: 46; Hvelplund, 2014: 212]. There was no effect of Concordia use on cognitive effort. However, a supplementary analysis was carried out for effect between Internet and No Internet conditions. This showed that there was an effect of Internet use on cognitive effort at the

p<.05 level for two conditions, i.e. Internet (M=127 s; SD=61 s) and No Internet (M=194 s; SD=38 s) [F(1, 13) = 4.84, p=0.046]. This could mean that the A text used exclusively for the Internet only condition was easier to process than the other two texts.

Figure 3. Fixations in the browser Interest Area during a recording session (P05, Concordia only condition)



Eye-tracking data, apart from allowing to quantify cognitive effort, enables to depict the spatial distribution of eye movements. Figure 3 presents a spatial rendering of fixations during a recording session of Participant 5 (henceforth P05, for all other participants as well). In the browser area (right-hand side), the fixations are joined to show fixation distribution in the browser.

This indicates that the minimal design of the interface focused all attention of the participant, rarely making them stray from its essential parts. The only possible distractor might have been the fact that each use of Ctrl+Q shortcut resulted in an automatic opening of a new tab, where the gaze paths are also plotted in the upper part of the browser.

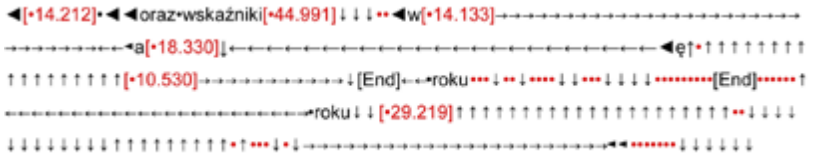
Table 3. Efficiency: Technical, temporal, and cognitive effort per condition

		Internet	Concordia	Internet + Concordia
Technical effort	Mean total events (*N*=6)	843	1103	1097
	SD	242	246	450
Temporal effort	Mean total task time [s] (*N*=6)	693	899	893
	SD	238	268	562
	Mean time spent in browser [s] (*N*=6)	172	174	246
	SD	83	75	106
Cognitive effort	Total dwell time (*n*=5)*	417	572	623
	SD	227	151	286

* One participant was removed from the analysis due to low quality data. All other participants' datasets contained good quality eye-tracking data, with mean dwell time over 200 ms for all interest areas. 200 ms was the quality threshold that Hvelplund [2011: 105] as well as Pavlović and Jensen [2009: 99] set as the minimum amount to filter out unacceptable data.

A supplementary analysis was carried out on text effect independently of the conditions. There was an effect of text on the technical effort, i.e. user events, at the $p<.05$ level for three conditions, i.e. Internet, Concordia, and Internet+Concordia [$F(2, 5) = 17.13$, $p=0.0001$]. It means that independently of the condition, participants generated a significantly different total number of user events for each text (A, B, C). There is a possible explanation for this and the Concordia/No Concordia user events effect. Technical effort is heavily dependent on participant's style of work, especially during the revision phase, where sometimes translators navigate and read the target text at the same time (e.g. with the use of Ctrl+→/← or just the arrows). This generates a lot of user events and, as can be seen from Figure 4, the participants did resort to this technique while revising (P01, P02 and P03).

Figure 4. Technical effort: Target text navigation during the revision phase (P01).



What is more, the use of Ctrl+Q shortcut might further explain this particular effect on technical effort. The shortcut Ctrl+Q pasted a highlighted portion of text into Concordia and searched the database for results. The other shortcut, i.e. Ctrl+W (allowing to paste into Concordia all the text from an active window), was not used by any of the participants. As per Figure 5, high individual variation for both conditions could be explained by how the shortcut was used. P03 used the shortcut while selecting larger portions of text, thus economising on technical effort and getting the same results others got when using the shortcut multiple times for short phrases or individual words. One-way ANOVA on the mean instances of shortcut use showed no effect of condition on the number of used shortcuts at the $p<.05$ level for both Concordia conditions. P03 selected as much text as the function allowed, using the tool in line with the instructions which specified that whole sentences could be processed by the system. Therefore, it is interesting that only one participant out of six utilised Concordia like this, thus saving additional effort. It may have been caused by the transfer of skill from how online searching usually proceeds, i.e. with precise and short phrases to narrow down the search results as much as possible. Nevertheless, this is a result that may indicate the use of such shortcuts to be caused by individual working style.

Figure 5. Efficiency: Instances of hot key use (Ctrl+Q) for searching in Concordia (N=6)

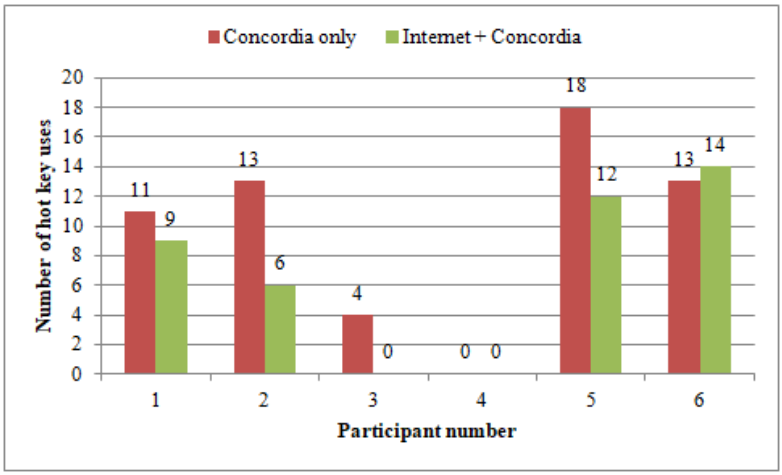


Table 4 shows efficiency quantified by instances of accessing online resources. We conducted a one-way ANOVA on the mean instances of use and there was no effect of condition on the number of resources consulted at the $p<.05$ level for three conditions. This is most likely because they were all aware that it was Concordia they were testing. It is then not surprising that in the condition which allowed the use of both Concordia and other Internet resources participants in total used Concordia almost as many times as they did in the Concordia only condition (43 vs. 49 respectively). Typing from memory meant that a participant either typed a translation solution without consulting a resource or formulated a different translation despite having consulted a resource. Such a high number of those could be found in the Concordia condition, which was likely to have been caused by the fact that Concordia failed to display a satisfactory result and participants then had to rely on their own guesses or literal translations.

Table 4. Efficiency: Instances of accessing resources, typing from memory, and shortcut use per condition

	Condition		
Instances of use	**Internet**	**Concordia**	**Internet + Concordia**
Google	14	-	12
Wikipedia PL	10	-	2
Wikipedia EN	9	-	3
IATE	10	-	21
Linguee2	4	-	0
Glosbe3	0	-	1
Diki4	2	-	0
bab.la5	3	-	1
Reverso Context6	0	-	1
ec.europa.eu	0	-	2
Concordia	-	49	43
Mean number of resource consultations per person (N=6)	9	8	14
Instances of typing from memory	37	39	25
Instances of shortcut use [Ctrl+Q]	-	59	41

One of the most often consulted websites was IATE (Inter-Active Terminology for Europe),[4] which is the largest database with multilingual EU terminology. All participants except P05 were familiar with the resource and its reliability for EU texts. The resource that acted as an intermediary to access other websites was Google as the engine was used to find such websites as IATE or Diki.[5] As for the complexity of search phrasing, only P05 used quotation marks once to find an exact phrase, but otherwise the searches did not use any operators to narrow down the search. This is probably due to the fact that the phrases or terms were

[4] <http://iate.europa.eu>, visited November 1, 2017.

[5] <https://www.diki.pl/>, visited November 12, 2017.

almost immediately accessible via IATE or other resources connected to EU documents. Another common search strategy was switching between Wikipedia language versions from English to Polish. Furthermore, search phrases in Concordia were also simple, participants searched for single words, such as *adopt* or whole phrases like *Council of Ministers* with one exception when P03 pasted whole sentences into Concordia.

Participants (*N*=6) also reported their preferences pertaining to translation resources on the scale of 1–5 (1 – never, 2 – rarely, 3 – sometimes, 4 – often, 5 – always). They rarely consult printed resources when translating (*M*=4.83, *SD*=0.4), more often utilising electronic resources (*M*=3.5, *SD*=0.22) and most of the time online ones (for paper resources *M*=2.5, *SD*=1.05). They also specified their online resource preferences. All reported Wikipedia among their favourite choices and that also could be seen in the distribution of resources accessed during the translation tasks. 83% listed Advanced Google Search, PROZ,[6] and IATE, while 67% mentioned Corpus of Contemporary American English.[7] Half of the participants also picked Linguee, Translatica,[8] and Google Translate. 33% reported Glosbe, Pons, Diki, Lin.pl and Polish National Corpus. A minority of respondents (17%) checked EUR-Lex,[9] Getionary,[10] and Mediteka.[11] In the process data Wikipedia was the most often consulted source after Google and IATE.

As can be seen in the data on the efficiency of Concordia, H1 can be only partially rejected, as it was the technical effort that increased in the Concordia only condition. There was no effect between other conditions and variables. However, in order to get a more complete image of a given tool's usability, in addition to efficiency, it is crucial to examine its other dimensions. Thus, we also tested H2 by analysing the product's quality, i.e. effectiveness.

6 <https://www.proz.com>, visited November 12, 2017.
7 <https://corpus.byu.edu/coca/>, visited November 12, 2017.
8 <https://translatica.pl/>, visited November 12, 2017.
9 <http://eur-lex.europa.eu/homepage.html>, visited November 12, 2017.
10 <http://getionary.pl/>, visited November 12, 2017.
11 <http://www.mediteka.pl/>, visited November 12, 2017.

2.6.2. Effectiveness

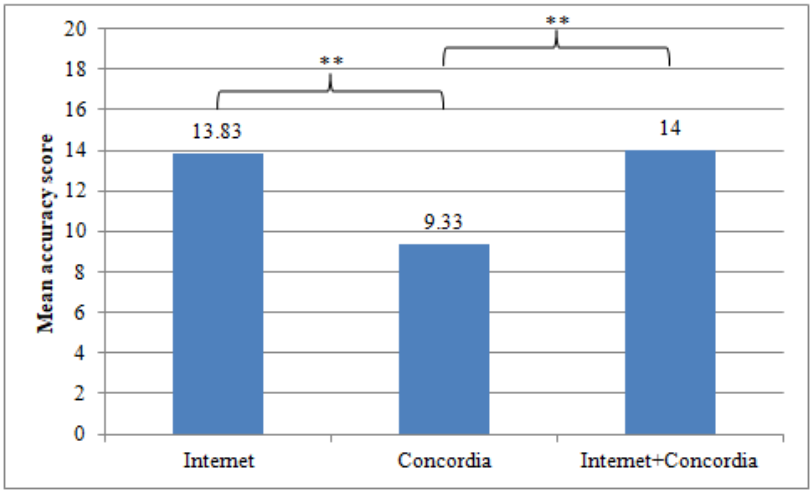
The effectiveness aspect of usability was calculated by means of accuracy scores. Each key phrase was worth 2 points and the solution was awarded 2 points when the translation was both terminologically and linguistically correct. We performed the rating of target texts on our own, using reliable resources to cross-check the suggestions. One point was awarded when there were spelling mistakes but the solution was not a mistranslation. Figure 6 shows that the lowest mean score was in Concordia condition, which was less than 50% accuracy per 10 key phrases in that condition. A one-way ANOVA was conducted to compare the effect of the inclusion of Concordia on the accuracy scores. There was an effect at the $p<.01$ level for all three conditions [$F(2, 15) = 8.74$, $p=0.003$]. A post hoc Tukey HSD test showed that mean accuracy score for the Concordia only condition ($M=9.33$, $SD=2.34$) differed significantly from the Internet condition ($M=13.83$, $SD=2.32$) and from Internet + Concordia condition ($M=14$, $SD=1.9$) at the $p<.01$ level. However, there was no significant difference between the Internet only condition and Internet + Concordia. This means that higher accuracy scores in the present study could be attributed to using other online resources and the condition with access exclusively to Concordia introduced a significant drop in accuracy. This can also be explained by the participants' own subjective impressions provided in the questionnaire (see Section 2.6.3.), in which they reported dissatisfaction with the quality and quantity of suggestions provided by Concordia.

Furthermore, Internet + Concordia condition was both the condition with the highest mean score as well as highest mean number of resource consultations and lowest number of instances of typing from memory. It might be the case that despite a balanced selection of texts and their key phrases, text A might have been easier for the participants, which also showed in the dwell time on source text differences (see Section 2.6.2). They typed a lot of terms from memory and there was also a high accuracy score for that condition. What is more, on average, participants generated the least amount of effort during the control condition with Internet only.

As can be seen from the data in Figure 6, Internet consultations were effective in the Internet + Concordia as well as Internet only conditions. Therefore, from the viewpoint of effectiveness, Concordia did not facilitate translation, which means that the H2 can be rejected. This was an

issue also reflected in the questionnaires which measured the participants' satisfaction with it.

Figure 6. Accuracy: Mean scores per condition (N=6)



2.6.3. Satisfaction

The third and final dimension of usability that was examined in the present study is satisfaction. We measured it through ratings of 10 Likert type statements based on Brooke's [1996] Software Usability Scale also utilised by Krüger [2016]. The scale was as follows: 1 – strongly disagree, 2 – disagree, 3 – neither disagree nor agree, 4 – agree, 5 – strongly agree. Figure 7 shows that participants thought Concordia was both easy to use and quick to master. They were relatively confident with how they used the tool, but were indifferent to the idea of using Concordia regularly.In the open question about what they thought Concordia should be able to do, the recurrent answer was that more than one suggestion should be displayed and that they could not find what they needed in the database (P01, P02, P04, P06). The chunking that Concordia did to the sentences and phrases engendered mixed reactions. It was reported that it chunked proper names and collocates (P02, P03).

Figure 7. Satisfaction: Usability questionnaire

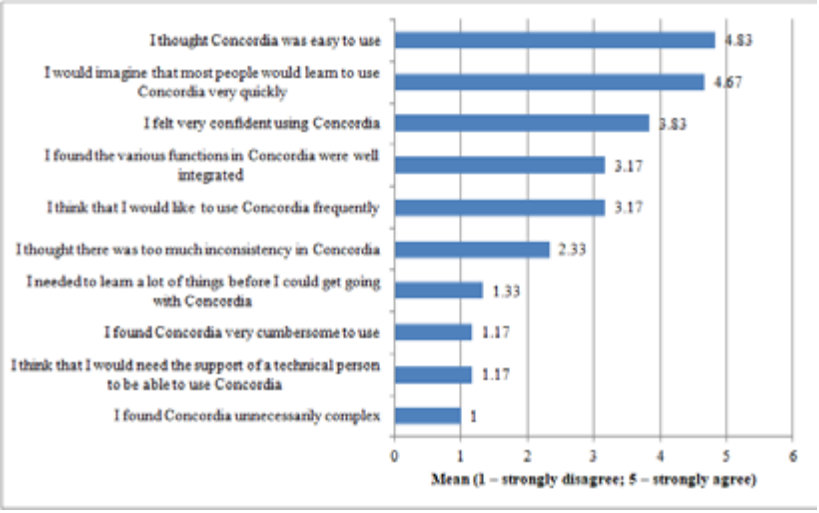
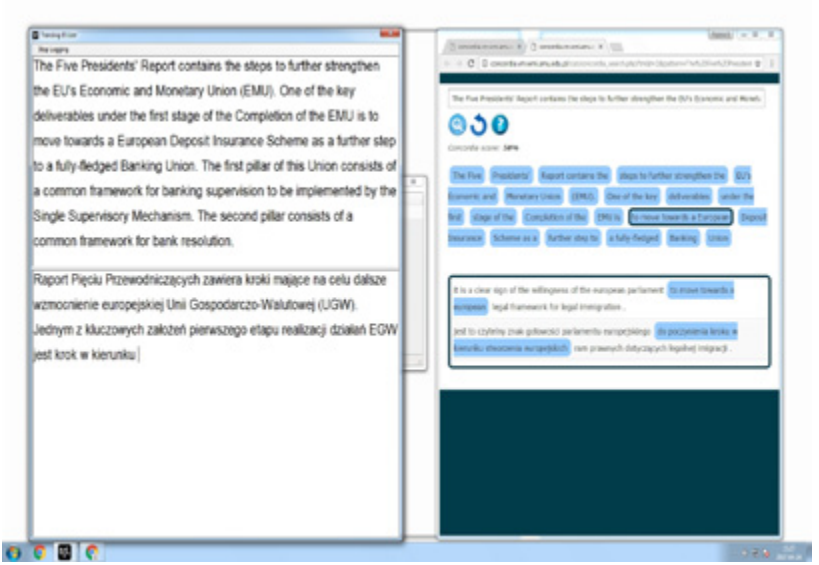


Figure 8. Close-up of the highlight feature in Concordia (P03)



Interestingly, P06 described Concordia in the 'additional comments' section of the questionnaire as "a great foundation for further research on the Internet", which indicates limited trust towards suggestions offered by Concordia and that further validation was needed through other sources. This might be a trait typical of student translators, who tend to overuse resources [Whyatt, 2012]. Additionally, all participants were also asked by the researcher what they thought about Concordia in comparison with Linguee which is slightly similar in terms of user interface. P03 described Concordia to be a "combination of EUR-Lex and IATE and good for [looking-up] sentences", also mentioning that there should be an indication of result confidence, similarly to the exclamation mark in Linguee which marked low-confidence results. Most of the participants (P02, P03, P04, P05, P06) reported the highlight feature of the searched phrase as useful and accurate, as opposed to Linguee's often misplaced and slightly misleading highlighting.

Thus, the participants were generally satisfied with Concordia. They were happy with its interface and features, while being relatively dissatisfied with the quality of its suggestions, contrary to what was anticipated in H3. From a quantitative perspective (see Figure 7), users were satisfied with Concordia. However, based on their comments in the open questions, there was some degree of dissatisfaction, thus making H3 only partially confirmed.

3. Discussion and further directions

O'Brien [2012: 115] once said about CAT tools that "there is little evidence to suggest that (...) [they] have been designed from the point of view of the humans who have to use them." With this study we intended to address the gap in Translation Studies that exists for process-oriented usability research. Our experiment generated both objectively measured data (eye-tracking, keylogging, screen recording) and subjective impressions. These types of data complement each other in a translation tool evaluation. We found that Concordia did not facilitate the translation process of the experimental texts from the viewpoint of two dimensions of usability, i.e. efficiency and effectiveness. However, the general impression of Concordia was quite positive albeit with criticism of the accuracy of displayed suggestions. It is worth emphasising how most of the comments from the questionnaire relate to the issue of trust towards

translation suggestions, as there was no information provided about the document's source, so it is apparent that the translators were not ready to take the single suggestion displayed by the system at face value. Despite this critical attitude towards Concordia's performance, the participants were clearly satisfied with their experience with the interface and functions, which means that the design of the tool itself was not the source of cognitive friction [Ehrensberger-Dow, O'Brien, 2015:102]. This may suggest that the simplicity which was appreciated by the participants is the key to the design of translation tools so that they can become familiar with them in a short amount of time.

There were also some limitations to the study. The most important shortcoming was the insufficient number of texts in the database. It is quite probable that with simpler terminology, Concordia's performance could have been very satisfactory. But had the terms been simpler, participants might type them from memory right away, not even thinking about checking them. Furthermore, the texts were balanced in terms of readability, but the difficulty of the terminology might have been a confound. To remedy this at the design level, probably a norming study of key phrases might have provided a clearer indication of their difficulty levels. Also, converting the open questions pertaining to satisfaction into Likert-type statements might have enabled better quantification of specific features of Concordia and their potential shortcomings.

Concordia is still at an early stage as a tool that has the potential to be incorporated into a TM environment or be a standalone aid, so the relatively poor performance is not a disaster, but a starting point to work towards a better final version. With its searching mechanism allowing to retrieve more information than traditional translation memory algorithms, Concordia, as an artefact in a translator's ecosystem, has the potential for high positive cognitive relevance in the process of translator-information interaction [Zapata, 2016: 136].

References

Bond, S. (2016), "Gunning Fog Index", [online:] http://gunning-fog-index.com/ – 13.11.2016.

Brooke, J. (1996), "SUS-A quick and dirty usability scale", *Usability Evaluation in Industry*, 189(194), pp. 4-7.

Campbell, S.G., Wayland, S.C., Goldman, A., Blok, S., Powell, A.L. (2013), "Speaking the user's language: Evaluating translation memory software for a linguistically diverse workplace", in: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting: Volume 57*, SAGE Publications Sage CA, Los Angeles, CA, pp. 2042-2046, [online:] http://dx.doi.org/10.1177/1541931213571456.

Christensen, T.P., and Schjoldager, A. (2011), "The Impact of Translation-Memory (TM) Technology", in: Sharp, B., Zock, M. (eds.), *Human-Machine Interaction in Translation: Proceedings of the 8th International NLPCS Workshop*, Samfundslitteratur, pp. 119-130.

Colominas, C. (2008), "Towards chunk-based translation memories", *Babel*, 54(4), pp. 343-354, [online:] http://dx.doi.org/10.1075/babel.54.4.03col.

Cooper, A. (2004), *The Inmates Are Running the Asylum: Why High-tech Products Drive Us Crazy and How to Restore the Sanity*, Sams Indianapolis, IN, USA.

Dillon, S., Fraser, J. (2006), "Translators and TM: An investigation of translators' perceptions of translation memory adoption", *Machine Translation,* 20(2), pp. 67-79, [online:] http://dx.doi.org/0.1007/s10590-006-9004-8.

Dragsted, B. (2004), *Segmentation in translation and translation memory systems: An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process*, Ph.D. series, 5, Samfundslitteratur, København.

Dragsted, B. (2006), "Computer-aided translation as a distributed cognitive task", *Pragmatics & Cognition,* 14(2), pp. 443-464, [online:] http://dx.doi.org/10.1075/pc.14.2.17dra.

Duchowski, A. (2007), *Eye Tracking Methodology: Theory and Practice: Volume 373*, Springer Science & Business Media, London.

Ehrensberger-Dow, M., Massey, G. (2014), "Translators and machines: working together", *Man vs. Machine*, 1, pp. 199-207.

Ehrensberger-Dow, M., O'Brien, S. (2015), "Ergonomics of the Translation Workplace: Potential for Cognitive Friction", *Translation Spaces*, 4(1), pp. 98-118.

Flesch, R. (1948), "A new readability yardstick", *Journal of Applied Psychology*, 32(3), pp. 221-233, [online:] http://dx.doi.org/10.1037/h0057532.

Giammarresi, S. (2008), "Second Generation Translation Memory Systems and Formulaic Sequences", *Łódź Studies in Language*, 17, pp. 419-430.

Guillardeau, S. (2009), *Freie Translation Memory Systeme für die Übersetzungspraxis*, University of Vienna.

Hutchins, E. (1995), "How a cockpit remembers its speeds", *Cognitive Science*, 19(3), pp. 265-288.

Hvelplund, K.T. (2011), *Allocation of Cognitive Resources in Translation: An Eye-tracking and Key-logging Study*, PhD Series, 10, Samfundslitteratur, Frederiksberg.

Hvelplund, K.T. (2014), "Eye tracking and the translation process: reflections on the analysis and interpretation of eye-tracking data", *MonTI Special Issue – Minding Translation*, 1(1), pp. 201-223.

ISO 9241 (2011), "Ergonomics of Human-System Interaction", Geneva: International Organization for Standardization.

Jaworski, R. (2015), "Approximate sentence matching and its application in corpus-based research", *The Future of Information Sciences: E-Institutions, Openness, Accessibility and Preservation*, 5, pp. 21-30.

Krings, H.P. (2001), *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*, Koby, G.S. (ed.), The Kent State University Press, Kent, Ohio.

Krüger, R. (2016), "Contextualising Computer-Assisted Translation Tools and Modelling Their Usability", *Trans-Kom-Journal of Translation and Technical Communication Research*, 9(1), pp. 114-148.

Lagoudaki, E. (2006), "Translation memories survey 2006: Users' perceptions around TM use", in: *Proceedings of the ASLIB International Conference Translating & the Computer: Volume 28*, pp. 1-29.

Lagoudaki, E. (2009), *Expanding the Possibilities of Translation Memory Systems: From the Translator's Wishlist to the Developer's Design*, Imperial College, London.

Lemhöfer, K., Broersma, M. (2012), "Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English", *Behavior Research Methods*, 44(2), pp. 325-343, [online:] http://dx.doi.org/10.3758/s13428-011-0146-0.

O'Brien, S. (2012), "Translation as human–computer interaction", *Translation Spaces*, 1(1), pp. 101-122, [online:] http://dx.doi.org/10.1075/ts.1.05obr.

Pavlović, N., Jensen, K. (2009), "Eye tracking translation directionality", in: Pym, A., Perekrestenko, A., (eds.) *Translation research projects 2*, Intercultural Studies Group, Universitat Rovira i Virgili, Tarragona, pp. 93-109.

Risku, H. (2002), "Situatedness in translation studies", *Cognitive Systems Research*, 3(3), pp. 523-533, [online:] http://dx.doi.org/10.1016/S1389-0417(02)00055-4.

Risku, H. (2004), *Translationsmanagement: Interkulturelle Fachkommunikation im Informationszeitalter: Volume 1*, Gunter Narr Verlag.

Risku, H., Windhager, F. (2013), "Extended translation: A sociocognitive research agenda", *Target. International Journal of Translation Studies*, 25(1), pp. 33-45, [online:] http://dx.doi.org/10.1075/target.25.1.04ris.

Rudlof, C. (2006), "Handbuch Software-Ergonomie", *Usability Engineering*.

Smallwood, J., Schooler, J.W. (2006), "The restless mind", *Psychological Bulletin*, 132(6), pp. 946-958, [online:] http://dx.doi.org/10.1037/0033-2909.132.6.946.

Strohner, H. (1995), *Kognitive Systeme: eine Einführung in die Kognitionswissenschaft*, Springer-Verlag.

Whyatt, B. (2012), *Translation as a Human Skill: from Predisposition to Expertise = Przekład jako umiejętność człowieka: od predyspozycji do poziomu eksperta*, Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza, Poznań.

Zapata, J. (2016), "Investigating Translator-Information Interaction: A Case Study on the Use of the Prototype Biconcordancer Tool Integrated in CASMACAT", in: Carl, M., Bangalore, S., Schaeffer, S. (eds.), *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB,* Springer, pp. 135-152, [online:] http://dx.doi.org/10.1007/978-3-319-20358-4_7.

APPENDIX A: Source texts and key phrases

xxxx – phrase found in Concordia
xxxx – phrase not found in Concordia

Typing task	The extension of the European project to countries wishing to take part only in its economic aspect, the European Economic Area (EEA) covers Iceland, Norway, Liechtenstein and Switzerland. Our European Economic Area Consultative Committee (EEA-CC) includes representatives from each of these countries and is responsible for making recommendations to policy-makers.
Text A*	The Lisbon Treaty strengthens the European Economic and Social Committee's (EESC) consultative role in relation to the Parliament. The Treaty gives the latter institution the same prerogatives as the Commission and the Council in terms of consulting the EESC. This will enhance the EESC's „bridging" role between civil society and EU institutions in the policy-shaping and decision-making processes. It also opens up new prospects for an increased involvement of the Committee at all stages of the EU legislative procedure.
Text B**	The Five Presidents' Report contains the steps to further strengthen the EU's Economic and Monetary Union (EMU). One of the key deliverables under the first stage of the Completion of the EMU is to move towards a European Deposit Insurance Scheme as a further step to a fully-fledged Banking Union. The first pillar of this Union consists of a common framework for banking supervision to be implemented by the Single Supervisory Mechanism. The second pillar consists of a common framework for bank resolution.
Text C***	Just under 5 million young people (under 25) were unemployed in the EU-28 area in August 2014, which represents an unemployment rate of 21.6% in the EU. Youth unemployment and NEETs (not in education, employment or training) indicators are part of the new Scoreboard of key employment and social indicators. The Scoreboard identifies the major employment and social imbalances within the EU. The first such Scoreboard was published as part of the Joint Employment Report 2014, jointly adopted by the Commission and the EU's Council of Ministers.

* www.eesc.europa.eu/resources/docs/faq-eesc-lisbon-treaty-en.doc, visited on October 10, 2017. All texts were abridged and adapted.

** http://europa.eu/rapid/press-release_MEMO-15-6153_en.htm, visited October 10, 2017.

*** http://europa.eu/rapid/press-release_MEMO-14-571_en.htm, visited October 10, 2017.

Abstract

Computer assisted tools used to seem as though not made from the point of view of their targeted users [O'Brien, 2012:15]. However, their usability has been improving. In Translation Studies there exists a gap in research on process-oriented usability involving data triangulation. In our study based on the assumption that translation is a situated and complex activity [Risku, 2002, 2004], we aimed to address this gap with our experiment testing a new tool for translators, Concordia. This usability experiment with eye-tracking, keylogging, and screen recording directly involved translators (six translation trainees) in the development process through objectively collected data on effectiveness and efficiency of Concordia. Their satisfaction with Concordia was also a part of the usability test. We hypothesised that participants would be more efficient and effective when translating European Union texts with Concordia and that they will be satisfied with the tool. The results indicate that Concordia at its current state of development does not facilitate the process, but the participants were generally satisfied with the tool's features.

Key words: translation studies, usability research, CAT tools, natural language processing, eye-tracking