Usability evaluation of a translation memory system

La evaluación de la usabilidad de un sistema de memoria de traducción

Chelo Vargas-Sierra

Universidad de Alicante. chelo.vargas@ua.es
Recibido: 30.04.2019. Aceptado: 02.08.2019

Resumen: El uso de herramientas de traducción asistida por ordenador en las clases de traducción se ha convertido en una práctica común desde hace poco más de una década. Con todo, para los estudiantes actuales, los llamados "nativos digitales", la experiencia de aprendizaje con este tipo de software está lejos de ser sencilla. Esto nos llevó a preguntarnos por las actitudes reales de los estudiantes respecto a la usabilidad de este tipo de software. Este trabajo presenta una evaluación de la usabilidad desde el punto de vista del usuario final de una memoria de traducción líder en el mercado. Más concretamente, el objetivo del estudio era evaluar la percepción de usabilidad de los estudiantes. Para ello, al final de dos cursos académicos, 95 estudiantes de último curso cumplimentaron el cuestionario Software *Usability Measurement Inventory*, que se considera un método de referencia para evaluar la usabilidad de un producto de software. Mide cinco escalas, esto es, Eficiencia, Afecto, Utilidad, Control y Aprendizaje. El análisis de los resultados obtenidos muestra que la opinión de los estudiantes sobre la usabilidad global de la herramienta evaluada está dentro de la media, pero no tanto con respecto a la escala Aprendizaje, que es la peor valorada. La única escala por encima de la media fue Afecto. Estos resultados muestran que se necesita hacer mayor énfasis en el diseño de la herramienta evaluada para que se adapte a las necesidades reales de sus usuarios y mejorar, de este modo, el conocimiento tecnológico de nuestros estudiantes de traducción.

Palabras clave: traducción asistida por ordenador; sistema de memoria de traducción; cuestionario SUMI; usabilidad; SDL Trados Studio.

Abstract: The use of Computer-Assisted Translation tools in translation classes has become a common practice for a little more than a decade. Even so, for students nowadays, the so-called digital natives, the teaching experience with this type of software is still far from being as easy as one would expect. This led us to ask ourselves what the real students' attitudes were regarding the usability of software of this kind. This paper

» Vargas-Sierra, Chelo. 2019. "Usability evaluation of a translation memory system". *Quaderns de Filologia: Estudis Lingüístics* XXIV: 119-146. doi: 10.7203/QF.24.16302

presents a usability evaluation for a leading desktop-based translation environment tool from the end user's point of view. More specifically, the aim of the study was to assess the students' perception of usability. For this purpose, at the end of two academic years, 95 senior students completed the Software Usability Measurement Inventory questionnaire, which is considered as a proven method for evaluating the usability of a software product. It measures five scales, i. e., Efficiency, Affect, Usefulness, Control, and Learnability. The analysis of the results obtained suggests that the students' opinion about the global usability of the tool under evaluation is within the average, but not so much with regard to its learnability, which is the worst-rated scale. The only scale above average was Affect. These results show that greater emphasis is needed on the design of the tool evaluated in order to adapt to the real needs of users and actually improve the technological savvy of our translation students.

Keywords: computer-assisted translation; translation memory systems; SUMI questionnaire; usability; SDL Trados Studio.

1. Introduction

In the last decade, higher education has undergone important changes caused, among other reasons, by the evolution of Information and Communication Technologies (ICTs) and their incorporation into traditional or virtual classrooms. Thanks to the use of these technologies, teachers can undertake different training activities to involve students in active learning processes and carry out translation tasks simulating real-life professional contexts. Furthermore, integrating these technologies into the teaching-learning process increases the motivation of teachers (Álvarez *et alii*, 2011) and learners (Ramírez-Polo & Ferrer-Mora, 2010: 32) and, consequently, using ICT helps to improve this process.

As far as translation and ICT training are concerned, the current scenario does not differ much from what was stated in the previous lines: the incorporation of new technologies in translation classrooms is currently a self-evident reality and technologies and their close relationship with translation do not need to be justified; it is naturally assumed that translators must know how to make use of the multiple tools that intervene in one way or another in their workstation (Vargas-Sierra & Ramirez-Polo, 2011), so in their training, translation students should ideally acquire an expert knowledge of technologies applied to translation. In the field of translation technologies, many tools have already been incorporated into translator-training programmes (Kenny, 1999). Among all these tools, we would like to highlight the most popular one among professionals: "the translation memory –which falls into the broad category named 'computer-aided translation (CAT) systems'" (Vargas-Sierra, 2011: 49). A translation memory system (TM) is a software product designed to speed up the process of translating a text by incorporating, among others, the functions of storing and retrieving previous translations units (or parallel segments) and equivalent terms, ensuring consistency and quality of translations, together with an increase in productivity (Candel-Mora, 2011).

The impact of technology in translation classes is more than evident and, therefore, we need to evaluate how learning to use a particular translation tool affects the development of the so-called instrumental sub-competence that students must achieve. In fact, the acquisition of this competence occupies a relevant position in current models of translation competence, such as those elaborated by the PACTE (2005)

and Transcom (Göpferich, 2009) groups. The European Master's in Translation group (2009) also places emphasis on it, dividing it into two categories: (1) the information mining competence, which includes the development of documentation strategies and the search for terminological information, as well as work with experts and the use of technological tools for this purpose; and (2) technological competence, referring to the agile and efficient use of files and software for translation, documentation, layout, terminology, etc. Thus, once we assume that technological competence should be taught to students, we need to consider the integration of translation software in the learning process.

Even though commercial Translation Memory systems have been available on the market for two decades now, the degree of satisfaction of both professional translators and/or beginners with such tools, as well as their efficiency, usability and helpfulness in practice are little-known aspects. As far as we know, the usability of TMs has not yet been systematically investigated in that we have no statistically documented empirical evidence of this important parameter. At the same time, since the translator is a professional whose workstation is full of computer tools, usability should become an area capable of drawing the attention of TM providers and developers in order to adapt, when necessary, their tools to the real needs of translators.

It is evident that for TMs to be user-friendly, we need to know the different reactions that the tool provokes in new users (students), as well as what factors encourage their motivation to learn to use it and to reach an optimum level in this learning. In this sense, we think it is necessary to assess the students' perception of the usability of a particular software product, since, on the one hand, it has become a key quality factor (Abran *et alii*, 2003) and, on the other, according to our experience, it can have a certain impact on the learning attitudes of students.

An assessment of the process of this acquisition, i. e., of the student's interaction with a translation memory system motivates this study. We seek to assess usability when regarding a leading desktop-based translation environment tool from the end user's point of view. To do so, we gathered a sample of 95 senior students who responded to the SUMI questionnaire, which is considered as a proven and well-known method for evaluating the usability of a software product, as will be argued in section 4.

This paper is then ordered as follows. In Section 2, we briefly examine the concept of usability and methods for assessing it. In Section 3, we refer to related work in the field of translation studies. Section 4 is devoted to the introduction of our evaluation framework and highlight the characteristics of both SUMI questionnaire and the sample. In Section 5, we provide a detailed overview of the data analysis and results obtained. Finally, we offer some concluding remarks.

2. Usability evaluation

Although users are increasingly familiar with ICTs, there is a greater demand for quality in the software that is released to the market so that it fits the real needs of users in real work environments. The concept of quality in use, like so many other concepts, has undergone changes over time, as has the software industry. In the opinion of Bevan (1995), the main change occurred when end users and real contexts of use were placed at the centre in the evaluation of software applications performance. This author also states that usability and quality complement each other and that "usability is quality in use" (*ibid*.: 1).

Usability is "a measure of interface quality that refers to the effectiveness, efficiency and satisfaction with which users can perform tasks with a tool" (Dillon, 2001: 1). Its evaluation is part of the system development process and is a central issue in human-computer interaction, defined as "the processes, dialogues and actions that a user employs to interact with a computer in a given environment" (Preece *et alii*, 1994). From this definition it follows that one of the fundamental purposes of the HCI is to remove obstacles that impede a smooth interaction between the user and the tasks to be performed. And this is where we think usability plays a key role.

For the purposes of this research, we use and understand the term *usability* in the sense given by ISO/IEC 25022, i. e., the "degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use".

According to Dillon (2001: 2), there are multiple methods for evaluating usability, but he highlights three categories among these:

1. User-based: where a sample of intended users try to use the application.
2. Expert-based: where a usability expert performs an evaluation of the application.
3. Model-based: when an HCI expert employs formal methods to predict one or more criteria user performance criteria.

The software under study will meet the requirements of our students if "it is effective (accurate and complete), efficient in use of time and resources, and satisfying, regardless of the specific attributes it possesses" (Bevan, 1997: 2). So, this concept represents an essential point when assessing a software product: software quality from a user's viewpoint, being, then, the user experience placed at the centre of the process. Moreover, as Dillon (2001: 4) points out, "properly executed user-based methods are always going to give the truest estimate".

There is a great deal of academic production in the computing field concerning usability evaluation and quality in use regarding and applied to different software products, since, as we said above, usability is integrated into the development process of software products and is a key quality factor. We found, among many others, evaluations for: web applications (Covella & Olsina, 2006), for web portals (Arh & Blažič, 2008), for semantic web exploration tools (González *et alii*, 2012), for mobile users' interfaces (Alnanih *et alii*, 2013) and, more recently, for Big Data (Merino *et alii*, 2016), or for video conferencing system (Khalid & Hossan, 2017).

3. Related research in translation studies

The study described in the current paper used a proven questionnaire method to assess the quality of use of a leading desktop-based translation environment tool in order to obtain information about its usability in a defined use context. As Krüger said (2016: 128) publications studying "the usability of CAT tools from a user-oriented perspective still seem to be relatively scarce". Empirical research carried out with data collection tools (surveys, questionnaires, interviews) on the use of translation technology is also limited, although the following stand out.

Lagoudaki (2008) carried out one of the first surveys on the usability aspects of TMs, in which she obtained information from 874 transla-

tion professionals who gave their opinion on the positive and negative points of this tool and expressed what they wanted TMs to incorporate in the future. One of the conclusions of that survey was that usability and end-user demands do not appear to have been a priority in the design and development of TM systems and it is recommended that "user engagement be pursued in all stages of software development" (*ibid*.: 205).

LeBlanc (2013) conducted an ethnographic study in three medium-sized translation firms and services in Canada focusing on the advantages and disadvantages of TMs when used at the workplace. One of the author's conclusions coincides with Lagoudaki's, in that part of the dissatisfaction among translators regarding TMs "revolves around the tool's conception or design" (*ibid*.: 10).

Asare (2011) also employed ethnographic methods to examine translators' workflow and their perspectives on the usability of translation tools at a translation agency. This case study investigated how certain features of CAT tools were used and whether some were not being used. The study finally verified that "a number of features in the translation tools were not being used because their purposes were not understood by the tool users" (*ibid*.: 138) and concluded that there is a divergence between the use the tool designer foresees and the way end users understand and use the software.

In the framework of a research project named Smart Computer-Aided Translation Environment (SCATE), Bergh *et alii* (2015) used several data collection tools (a survey, semi-structured interviews and nine contextual consultations) to conduct an empirical research on the translators' working practices and tools. Their study concludes with a set of recommendations that could positively impact translators' workflow and which focus on improving the efficiency, effectiveness usability and control of translation tools.

The large-scale survey titled *ErgoTrans* (O'Brien *et alii*, 2017) included specific items to assess, from the perspective of a professional translator, whether any of the characteristics of CAT tools (in general) were identified as negative –"irritating or missing", in authors' words. The results of this survey revealed that more than half of the users' comments on "irritating features were associated with the User Interface or with CAT Tool Functionalities" (*ibid*.: 160).

4. Evaluation method

The aim of our study, as indicated above, was to test the usability of a leading desktop-based translation memory; more specifically, a single-user version of SDL Trados Studio 2017. We wanted to discover the perceived quality of this kind of software by real users (translation students) in real contexts of use (translation process) by applying a well-established and proven test developed in compliance with technical standards.

Questionnaires are a common method in usability assessment, used with real users and with a simple application. Some of the usability evaluation questionnaires available are, according to Fu & Schmidt (2006: 2021), the following: "User Satisfaction Scale, Computer User Satisfaction Inventory (CUSI), Questionnaire for User Interface Satisfaction (QUIS), Software Usability Measurement Inventory (SUMI), Purdue Usability Testing Questionnaire (PUTQ), After Scenario Questionnaire (ASQ), Post-Study System Usability Questionnaire (PSSUQ), and Computer System Usability Questionnaire (CSUQ)."

4.1 *SUMI questionnaire*

In our case, after an extensive information search and review of some of the questionnaires mentioned above, we decided to employ the SUMI method, developed in the project "Metrics for Usability Standards in Computing" (MUSiC) (Kirakowski, 1996) by usability experts.

As we said before, SUMI can be considered as a proven and well-known method for evaluating the usability of a software product. Proof of this are the many publications in which the questionnaire is mentioned and detailed or in which it has been applied to carry out the evaluation in different applications. According to Kirakowski (1994: 27), Preece *et alii* (1994) suggest that SUMI can be considered "as a standard method for assessing user attitudes" and by Dzida *et alii* (1993) "as a way of achieving measurement of user acceptance in the context of the Council Directive on Minimum Safety and Health Requirements for Work with Display Screen Equipment (EEC, 1990)". Davies and Brailsford (1994) also recommend SUMI in their publication about guidelines for developers of multimedia courseware development. SUMI was also used to assess the usability of an e-learning system for occupational medicine

(Rognoni *et alii*, 2008), a multilingual educational portal (Arh & Blažič, 2008), an eLearning course (Deraniyagala *et alii*, 2015) or a video conferencing system in a university's classroom (Khalid & Hossan, 2017), among many others studies. Good evidence of its representativeness is that SUMI is referred to in ISO 9241-11:1998 about "Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability". Its validity has also been conducted with three different kind of studies (*cf.* Kirakowski, 1996: 174-175).

According to the data stated in the previous paragraph, SUMI can be said to be a valid and reliable method to measure users' perception of a software usability as well as an objective way of assessing user satisfaction. Its selection was also based on the following parameters: it is applicable to any software system, it is available in several languages, including Spanish, it is designed according to recognized technical standards and is easy to implement, as it is ready to use and hosted on the web (http://sumi.uxp.ie/es/index.php). Another important parameter was that a very detailed SUMI report is provided as a zipped encoded file when the evaluation has been completed.

The questionnaire is designed in a simple way. It contains a total of 50 straightforward statements about the users' attitudes to the software they employ divided into batches of 10 that are answered by checking one of the three boxes: "Agree", "Undecided" or "Disagree":

Statements 1 - 10 of 50.	Agree	Undecided	Disagree
This software responds too slowly to inputs.	○	○	○
I would recommend this software to my colleagues.	○	○	○
The instructions and prompts are helpful.	○	○	○
This software has at some time stopped unexpectedly.	○	○	○
Learning to operate this software initially is full of problems.	○	○	○
I sometimes don't know what to do next with this software.	○	○	○
I enjoy the time I spend using this software.	○	○	○

Figure 1. Statements in SUMI questionnaire

The questionnaire assigns each statement to one of the five following scales and in this order: (1) Efficiency; (2) Affect; (3) Helpfulness;

(4) Control; and (5) Learnability. Therefore, to avoid bias towards answering in a certain direction, the first item is assigned to the first scale (Efficiency), the second to the second scale (Affect), and so on. The SUMI scales are defined as follows (Kirakowski, 1996: 10):

- Efficiency refers to the respondent's feeling that the software allows them to perform their tasks quickly, efficiently and economically or, at the opposite extreme, that the software is interfering with performance. The items in this scale are: 1, 6, 11, 16, 21, 26, 31, 36, 41, 46.
- Affect is a psychological term for "the user's general emotional reaction to the software". In this context, it refers to the feeling of the respondent in that he finds the interaction with the tool as stimulating and enjoyable or, on the contrary, as stressful and frustrating. Its items are: 2, 7, 12, 17, 22, 27, 32, 37, 42, 47.
- Helpfulness (items 3, 8, 13, 18, 23, 28, 33, 38, 43, 48) measures whether the software communicates usefully with the user and helps to solve problems that arise when using it. It also refers to "the adequacy of help facilities and documentation".
- Control measures the extent to which the user feels in control of the software or, conversely, that s/he is being controlled by the software in carrying out the task. Items in this scale are: 4, 9, 14, 19, 24, 29, 34, 39, 44, 49.
- Learnability –with items 5, 10, 15, 20, 25, 30, 35, 40, 45, 50– refers to the respondent's feeling concerning to how easy it is for them to become familiar with the software. It also measures whether the user feels that the tool interface, tutorials, manuals, etc. are easy to understand and help them get the job done. Moreover, it refers to the ease with which the user reuses the software after a while without doing so.

In order to represent the "unique construction of the 'perceived quality of use'" (Kirakowski, 1996: 11), SUMI questionnaire has also a global scale, which refers to a general feeling of satisfaction with the users' experience of the software under evaluation. It is calculated by evaluating 25 items that turn out to be the most relevant items in terms of usability. These are: 7, 8, 15, 18, 19, 21, 23, 24, 26, 27, 28, 31, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45, 46, 48 and 49.

At the end of the 50 statements, the questionnaire includes two multiple choice questions, each with four options that only admit one answer and two others that are open-ended. The aim of these four questions is to obtain a general assessment, so the questions ask for (a) the importance of the software for their work; (b) their software skills and knowledge; (c) the best aspect of this software; and (d) which aspects need most improvement:

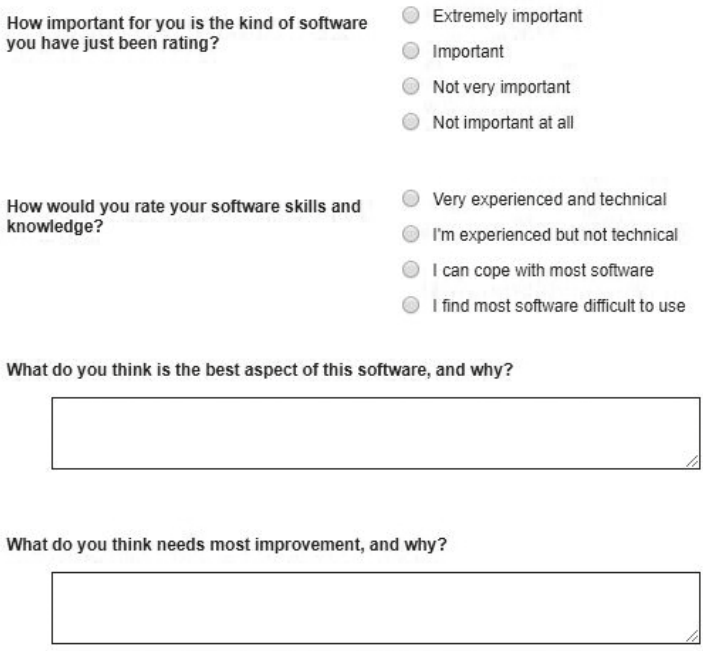

Figure 2. Questions in SUMI questionnaire

4.2 *Sample description*

The minimum user sample size required to perform an analysis with tolerable accuracy using SUMI is in the order of 10-12 users (Kirakowski, 1996: 27). In our case, the total number of responses analysed was 95. For most social science type research, 95 data points is a "good enough" sample. The classic answer is to ask how large the population is of

which this is a sample. Since it is difficult to estimate population sizes realistically in most situations, SUMI developers rely on the fact that after a size of 12 or so, the "t" distribution is close to the normal for the first three decimal places, and that the curve rises steeply so that by 60 it is almost at the level where it will be at 120. In Kirakowski's words, informal benchmark is n=20, since when n=12 the standard deviation of samples using the SUMI scales starts to approach an asymptote, so when n=12, it is OK but when n=20, it is safe.

The participants in this study were students of the General Translation (English-Spanish) course for the BA Translation and Interpreting Degree at the University of Alicante. This is a compulsory 4th year course that is taught during the first four-month period (from September to December). Our method is user-based according to Dillon's classification (2001).

The students used SDL Trados Studio 2017 during the development of the practical translation classes in a computer room. They were also provided with a temporary version so that they could practice at home and carry out a final translation project. When the course started, some students said they had some knowledge of TMs, but most said they had no experience with this commercial tool. For 15 two-hour classes, that are given in the four-month period, translation students were taught progressively how to use the tool, starting from a beginners' level and up to an intermediate one, that is, the stage at which users are able to build a translation project with all the required resources, translate different kind of texts, create the translation text and export the project translation memory in a specific format.

The selection of participants was based on the students who regularly attended classes and corresponds to the groups from two academic years: 2017 and 2018. No student knew s/he was going to evaluate the tool. The group of students in the first academic year answered the questionnaire on 20 December 2017 and the second on 19 December 2018. The idea was to carry out the usability evaluation at the end of the course so that the students would have had time to use the TM system and to practice the main tasks associated with this type of translation software, from beginner to intermediate level.

5. Data analysis and results

Once the questionnaire is completed, an analysis and reporting tool named SUMISCO –included in the SUMI package– scores it and compares the results to an extensive reference database (Kirakowski, 1996). The mean score of the reference database is 50, with a standard deviation of 10 and with a maximum score of 73.

Kirakowski (1996) states that the properties of the standard normal distribution indicate that more than 68 % of software will score on all SUMI scales within the values of a standard deviation (±10) from the mean (50), i. e. between 40 and 60. Software that scores more than 60 or less than 40 is defined as well-above or well-below the average.

5.1 *Summary statistics*

The summary statistics of SUMI results are presented in Table 1 and it contains the mean, the standard deviation and the median values in our study for all scales:

	Mean	**St Dev**	**Median**
Global	44.89	12.50	44.0
Efficiency	42.71	13.52	41.0
Affect	51.08	14.27	52.0
Helpfulness	47.58	12.70	50.0
Control	45.14	12.09	46.0
Learnability	32.97	13.94	31.0

Table 1. Summary statistics of SUMI results

Since the standard deviation values are within a normal distribution, our analysis will focus on the mean values.

The single-user version of SDL Trados Studio achieves a Global usability score of 44.89, indicating that the usability of the system is slightly below average. This global score is lower than you would expect from a market-leading software product. The means are between

32.97 (Learnability) and 51.08 (Affect). Most scales are generally within the average, Learnability being outside the satisfactory range, i. e. well-below average. It is important to point out here that, according to Nielsen (1993), this is "the most fundamental usability attribute", and "technologies should be easy to learn and understand" so that the new user can perform the tasks quickly and efficiently.

Affect is the best-rated scale in this CAT tool and means that students enjoyed working with the software and find it mentally stimulating to use. The low score for Learnability, conversely, demonstrates what we observed in class: that is, that the students found it difficult to master the software, understand its concepts, and often forgot how it worked, how certain tasks were performed.

5.2 *Means with standard deviation*

The following graph (Figure 3) shows the means with standard deviation for all scales, including the global usability one. The mean is in the centre of the ring and is represented by a cross; the staples are extended by one standard deviation on each side of the mean. When the mean is above the 50 line then the staple is shown in green and when it is below this mark, it is shown in red. When it is above the 50 line, it means that

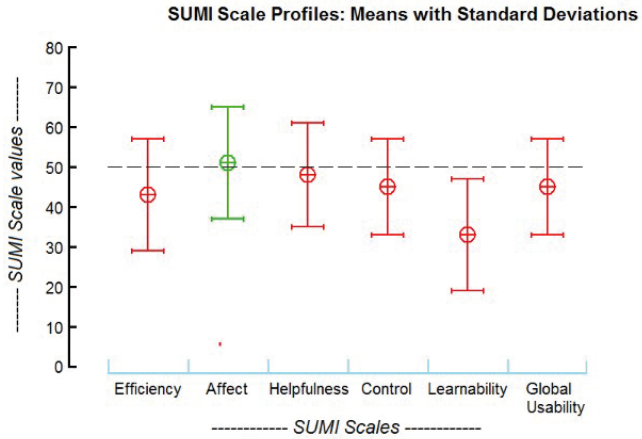

Figure 3. Means with standard deviations for usability evaluation

it is better than the reference database. As we have just seen on the previous table, only the Affect scale is above this line and is, consequently, shown in green. Scores at or below 40 may indicate the need to adjust and improve the tool.

As Figure 3 shows, the higher values were obtained for Affect, Helpfulness and Control, while the lowest ones were given to Efficiency and Learnability. Only the Learnability scale does not reach the 50 line, not even after the standard deviation is added.

5.3 *Item consensual analysis*

The SUMI also generates a strengths and weaknesses analysis named "Item Consensual Analysis", whose focus is to compare these two polar dimensions (strengths and weaknesses) to the statistics in the reference database. Any probabilities of difference equal to or greater than p = 0.95 are treated as "statistically significant". With respect to the statistically significant results, there are three categories of difference:

1. A strong "Agree": when there are many more observed votes in this direction than expected. If the statement has a positive meaning (e. g. "I would recommend this software to my colleagues") and the respondent click on "Agree", then the answer is positive, but if the statement has a negative meaning (e. g. "Using this software is frustrating") and we agree, then it has negative result.
2. A strong "Disagree": when there are many more observed votes in this direction than expected. A "Disagree" to a negative statement is positive but a "Disagree" to a positive statement is negative.
3. Very Undecided: when respondents give a lot more votes to the central "Don't Know" option.

Our report highlighted 37 items worth considering, since they reported statistically significant results. Below are the tables with these results according to above-mentioned categories of difference, which include the item number, the scale it measures, its statement and the result provided.

Item	Scale	Statement	Verdict
1	E	*This software responds too slowly to inputs.*	Very undecided. Chi Square: 27.891 *(p = .999)*
26	E	*Tasks can be performed in a straight forward manner using this software.*	Very undecided. Chi Square: 33.524 *(p = .999)*
27	A	*Using this software is frustrating.*	Very undecided. Chi Square: 6.192 *(p = .954)*
3	H	*The instructions and prompts are helpful.*	Very undecided. Chi Square: 11.247 *(p = .996)*
13	H	*The way that system information is presented is clear and understandable.*	Very undecided. Chi Square: 16.917 *(p = .999)*
23	H	*I can understand and act on the information provided by this software.*	Very undecided. Chi Square: 48.843 *(p = .999)*
43	H	*Either the amount or quality of the help information varies across the system.*	Very undecided. Chi Square: 9.053 *(p = .989)*
4	C	*This software has at some time stopped unexpectedly.*	Very undecided. Chi Square: 7.164 *(p = .972)*
44	C	*It is relatively easy to move from one part of a task to another.*	Very undecided. Chi Square: 16.164 *(p = .999)*

Table 2. Very undecided

The table above (Table 2) shows that the students' indecisiveness was more prominently directed towards issues of the Helpfulness scale (4 questions). If we look at the statements of this scale, we will see that they relate to the concept of "information", a feature that helps the user to carry out different tasks smoothly. This indecisiveness is also shown when students had to answer two statements concerning Control (the ease to move from task to task or whether the software stopped unexpectedly) and Efficiency (a slow response or if tasks are performed in a straightforward manner).

Students were also unsure about the question of whether the use of the software was frustrating. Their indecisiveness may have arisen because the opposite dimensions Affect and Learnability came into play. On the one hand, they feel that the tool is not easy to use, but on the other hand they recognize that it is important to learn it for their professional future, which motivates them.

Item	Scale	Statement	Verdict
2	A	*I would recommend this software to my colleagues.*	Agree. Chi Square: 10.74 *(p = .995)*
12	A	*Working with this software is satisfying.*	Agree. Chi Square: 10.854 *(p = .995)*
29	C	*The speed of this software is fast enough.*	Agree. Chi Square: 11.518 *(p = .996)*

Table 3. Strong Agree in positive statements

A strong agree was obtained in two Affect statements –one referring to software recommendation and the other to satisfaction– and in one of the Control scales (about speed), as shown in Table 3. Thus, we observe that, again, despite the difficulty students have in learning and mastering the software they would recommend it and consider that working with it is satisfactory. This is definitely one of its strengths. As experts in the evaluated tool, we can affirm that it is a very powerful software, with many very practical features oriented to different tasks that are carried out in the whole translation process (preproduction, production and postproduction). It has its defects and virtues, but not in vain is this software the most widely used in the profession. The fact that it is a key software in real working environments may have an impact on these positive assessments regarding Affect, since students may feel that mastering this tool can open doors for them when they go out into the job market, along with other potential benefits, such as reduction of repetitive work, productivity increase and a better quality in translation output (use of consistent terminology, QA features, etc.), sharing of language resources, among others.

Next tables (Table 4 and 5) show the results for Strong Agree in negative statements and Strong Disagree in positive ones, so both reflect negative results and should therefore be considered as weaknesses.

Item	Scale	Statement	Verdict
6	E	*I sometimes don't know what to do next with this software.*	Agree. Chi Square: 36.446 *(p = .999)*
11	E	*I sometimes wonder if I am using the right function.*	Agree. Chi Square: 44.649 *(p = .999)*
16	E	*This software seems to disrupt the way I normally like to arrange my work.*	Agree. Chi Square: 56.702 *(p = .999)*
36	E	*There are too many steps required to get something to work.*	Agree. Chi Square: 40.637 *(p = .999)*
41	E	*The software hasn't always done what I was expecting.*	Agree. Chi Square: 10.443 *(p = .994)*
46	E	*This software occasionally behaves in a way which can't be understood.*	Agree. Chi Square: 8.562 *(p = .986)*
32	A	*There have been times in using this software when I have felt quite tense.*	Agree. Chi Square: 31.861 *(p = .999)*
37	A	*I think this software has sometimes given me a headache.*	Agree. Chi Square: 75.414 *(p = .999)*
47	A	*This software is really very awkward.*	Agree. Chi Square: 54.47 *(p = .999)*
8	H	*I find that the help information given by this software is not very useful.*	Agree. Chi Square: 21.373 *(p = .999)*
18	H	*There is never enough information on the screen when it's needed.*	Agree. Chi Square: 24.565 *(p = .999)*
14	C	*I feel safer if I use only a few familiar functions.*	Agree. Chi Square: 82.622 *(p = .999)*

49	C	*Getting data files in and out of the system is not easy.*	Agree. Chi Square: 24.067 *(p = .999)*
5	L	*Learning to operate this software initially is full of problems.*	Agree. Chi Square: 101.1 *(p = .999)*
10	L	*It takes too long to learn the software functions.*	Agree. Chi Square: 143.863 *(p = .999)*
20	L	*I prefer to stick to the functions that I know best.*	Agree. Chi Square: 55.964 *(p = .999)*
25	L	*There is too much to read before you can use the software.*	Agree. Chi Square: 163.342 *(p = .999)*
30	L	*I keep having to go back to look at the guides.*	Agree. Chi Square: 91.694 *(p = .999)*
35	L	*Learning how to use new functions is difficult.*	Agree. Chi Square: 70.125 *(p = .999)*
40	L	*I will never learn to use all that is offered in this software.*	Agree. Chi Square: 32.758 *(p = .999)*
45	L	*It is easy to forget how to do things with this software.*	Agree. Chi Square: 67.784 *(p = .999)*
50	L	*I have to look for assistance most times when I use this software.*	Agree. Chi Square: 79.423 *(p = .999)*

Table 4. Strong Agree in negative statements

Item	**Scale**	**Statement**	**Verdict**
48	H	*It is easy to see at a glance what the options are at each stage.*	Disagree. Chi Square: 43.327 *(p = .999)*
19	C	*I feel in command of this software when I am using it.*	Disagree. Chi Square: 15.021 *(p = .999)*
34	C	*The software allows the user to be economic of keystrokes.*	Disagree. Chi Square: 20.931 *(p = .999)*

Table 5. Strong Disagree in positive statements

In tables 4 and 5 we can see that the greatest number of negative statements have been received by the Learnability (9) and Effectiveness (6) scales, followed by Control (4), Helpfulness (3) and also Affect (3). We believe that these results are striking in terms of certain weaknesses that the students strongly emphasize. In fact, each of these statements is very illustrative on its own of the issues that can be improved.

The above weaknesses (Tables 4 and 5) identify needs that uncover potential areas of improvement for TM usability:

- Efficiency: The tool could have a task wizard where students say what they want to do next and the software responds in a timely manner. More clarity is needed in order to know well the functions to be carried out and more flexibility to organize the work. Another important issue is the decrease in the number of steps needed to start any task, as well as the simplification of software performance.
- Affect: Ehrensberger-Dow and O'Brien (2015) observed that working with specialized translation technology tools is partially associated with "cognitive friction", which may be caused "by technological and organizational constraints" (*ibid*.: 102). The tool should be easier to use or, better yet, automatically adapt or customize its user interface (UI) according to the users' level of expertise, which would increase their satisfaction and decrease frustration.
- Helpfulness: TM should communicate usefully with the students and help them, in an effective way, to solve the problems that arise. Consequently, it is necessary that all the help facilities of the tool and documentation are adequate for its different users. A beginner does not need to consult or receive the same information as an expert, so help facilities should be adaptable, flexible and scalable to the user level.
- Control: Students do not feel that they have control of the tool, on the contrary, the tool controls them. They don't dare try new functions and find file management complicated. The tool should show the student how to perform certain actions or new, faster ways to do what they already know.
- Learnability: Trados Studio is a complex programme in that it can perform the multiple and varied tasks that a current translator

carries out, from project management to final quality assessment. With this scenario, it shouldn't be easy to design a user interface that is *familiar* to students. Familiarity may come when the design of the UI follows a metaphor from people's real-world experience. Most of our students do not have practical experience in the real world of translation and there are concepts they do not really understand at this point. Identifying the design concepts that beginners learn quickly and contrasting them with those that they don't can be extremely beneficial to the overall ease of use of Trados Studio. For complex software like this, conducting an evaluation of the learnability of the UI design can help uncover usability issues.

5.4 *Answers to multiple choice questions*

The answers to the question *How would you rate your software skills and knowledge?* obtained mean scores according to software skills and knowledge. In Table 6, the number of respondents who selected each of the choices is displayed, and then the SUMI profile of the average of the respondents who selected that button:

	n	**G**	**E**	**A**	**H**	**C**	**L**
Very experienced and technical	13	49.8	47.5	56.8	51.1	49.0	38.5
I'm experienced but not technical	27	43.3	42.7	51.4	43.3	44.0	35.5
I can cope with most software	45	46.8	43.4	52.5	50.8	46.2	31.8
I find most software difficult to use	10	34.2	33.2	36.3	40.0	38.7	24.3

Table 6. Software skills and knowledge

Table 6 shows what we expected: the greater the student's experience and technological skills, the better usability scores the tool obtains. In effect, we see that the usability score of the system is appreciably higher for very experienced and "digital" users (those that can cope with most software) than for respondents who find the software, in general, difficult to use. Most answers (45) are in line with *I can cope with most software* (digital users) and reach the second higher scores (except for

Learnability), before *I'm experienced but not technical*; this is consistent with the sample, as most respondents were students with little or no experience with TMs. The best-rated is Affect, above average for most respondents, and the worst Learnability, which is not within the average for any kind of user, not even for the most experienced or for the digital users. Both these types of users assess Helpfulness above 50, so they feel that software communicates usefully with them and helps them to solve arising problems.

The answers to question *How important for you is the kind of system you have just been rating?* obtained mean scores according to perceived importance (see Table 7). Next table shows the results:

	n	**G**	**E**	**A**	**H**	**C**	**L**
Extremely important	42	49.7	48.9	58.8	50.4	47.9	39.6
Important	36	42.3	39.1	47.6	46.6	44.7	28.2
Not very important	15	38.9	34.7	40.1	42.7	39.7	27.5
Not important at all	2	36.5	37.5	34.0	43.0	36.5	20.0

Table 7. Perceived importance of the system

As Table 7 shows, only two participants selected the option *Not important at all*, and 15 selected *Not very important*. Most of them selected *Extremely important* (42) or important (36), in total 82 %. Obviously, when the tool is evaluated as extremely important all scales get higher scores; here Affect reaches 58.8, Helpfulness 50.5 and the Global usability scale is near 50 (49.7). However, again Learnability is well-below average in all options.

5.5 *Answers to open-ended questions*

The students expressed their views on the TM under evaluation in the last two open-ended questions of SUMI questionnaire (see Figure 2). The first one was: *What do you think is the best aspect of this software, and why?* The students wrote their opinions about one or more characteristics they liked. In order to group the coincidences, we extracted the

information and categorized according to the functions that the students highlighted in their responses.

The results are as follows: in 32 responses, participants emphasized the usefulness of having a term base and a translation memory that stores pairs of segments. In 14, they highlighted the main function of a TM: to retrieve translations when an identical or similar segment appears. 28 students commented that they liked not having to worry about input/output document formats and 21 highlighted questions about increasing productivity. 16 responses emphasized that the system helps ease the burden of translation. 6 respondents commented on quality improvement and consistency of translations. 9 students thought that the system was easy to use and 5 mentioned the leadership of the software in the market. 4 students liked the fact that Trados had many functionalities, 3 liked its interface and 3 highlighted the revision process within the system.

The above characteristics that respondents highlighted could be divided into two: those that are specific to all translation systems (termbase and TM resources, formats, etc.) and those that are specific to the system evaluated (interface, features). As we have not made a comparison of usability between various systems, we must assume that they all refer to Trados; this is an issue that remains pending for future research.

Second open-ended question was *What do you think needs most improvement, and why?* The comments were mostly brief, without providing much detail. 15 responses did not indicate any particular improvement (*No creo que tenga nada que mejorar, Nada en particular, Nada…*). Four opinions referred to the high price of the product. 75 opinions were classified as referring to Efficiency, Helpfulness and Learnability scales. Efficiency obtained a total of 34.67 % of the comments, in which students highlighted aspects such as problems with file formats, start-up low speed, bad functionality, need of a collaborative version, compatible with macOS. Most of the opinions concerned Helpfulness (40%). Here the students referred to the need for clearer error messages, the need to make the tool more user-friendly and improved and clearer instructions. Concerning Learnability (22.67 %), the opinions focused mainly on two aspects: making the software clearer to be able to learn how to use it (*Que sea más claro a la hora de aprender*

a usarlo), and the availability of tutorials in Spanish (*Debería tener más soporte y videos en español*).

6. Concluding remarks

The SUMI questionnaire has proved effective in our study as it identifies critical usability problems that we had been experiencing for several years when teaching how to use this tool in translation classes. The survey has provided a lot of detailed information that back up our experience of using SDL Trados studio.

What I have presented here is still, above all, a case study, and I must stress that the aim is not to generalise at all. This study focuses exclusively on the perceptions of university students of translation when they use a particular, market-leading tool, and does not claim to be representative for all possible types of users, although it does, to a certain extent, for translators with a basic user level.

Based on the usability scores obtained through this study, as shown in Table 1, the result is that on the Global usability scale the tool is situated within an average range but does not reach a scale value of 50, which would be desirable, as was the case with the Affect scale, which is the highest. This is significant given that this tool is the most widely used by translation professionals. Except for the Learnability scale, the results of the rest of the scales were within the average, in the range of 40 to 60. Consequently, we believe there is a lot of room for improvement regarding scales between 40 and 50, and we deem the scale below 40 as critical and propose that it should be acted upon.

The SUMI questionnaire allowed us to detect 25 usability weaknesses in the tool we analysed. The usability mean scores yielded according to both perceived importance and users' knowledge and skills show that usability should be improved, especially regarding factors that have to do with Learnability. What the results show is that this scale is being disregarded by the software developers and designers, in spite of its importance, and the software under evaluation needs to be made simpler and a more comprehensible straightforward interface should be provided.

In view of the results, to improve the scores designers need to make modifications to the tool to improve, above all, its Learnability, but

should go beyond that, as we have seen weaknesses in Efficiency, Control and Helpfulness.

If the software is of good quality in terms of Learnability, the time needed to become proficient in the use of the tool will be reduced, productivity will be increased during this crucial phase and a feeling of satisfaction will be generated in the learner/users. It goes without saying that it will make the work of the teacher in this type of classes much easier and improve the students' level of satisfaction and competence.

Acknowledgements

We would like to express our appreciation to Jurek Kirakowski for his constant assistance in interpreting SUMI data and the additional statistical material provided. We would also like to thank Barry Pennock Speck for his constructive comments and the English revision of this paper.

Bibliography

Abran, Alain; Khelifi, Alain; Suryn, Witold & Seffah, Ahmed. 2003. Usability Meanings and Interpretations in ISO Standards Software. *Quality Journal* 11(4): 325-338. doi: https://doi.org/10.1023/A:1025869312943

Alnanih, Reem; Ormandjieva, Olga & Radhakrishnan, T. 2013. A New Quality-in-Use Model for Mobile User Interfaces. In *Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement. IEEE*.

Álvarez, Susana; Cuéllar, Carmen; López, Belén; Adrada, Cristina; Anguiano, Rocío; Bueno, Antonio & Comas, Isabel. 2011. Actitudes de los profesores ante la información de las TIC en la práctica docente. Estudio de un grupo de la Universidad de Valladolid. *EDUTEC, Revista Electrónica de Tecnología Educativa* 35. http://edutec.rediris.es/revelec2/revelec35/ [Accessed 26/04/2019].

Arh, Tanja & Blažič, Borka Jerman. 2008. A Case Study of Usability Testing - the SUMI Evaluation Approach of the EducaNext Portal. *WSEAS Transactions on Information Science & Applications* 2(5): 175-181.

Asare, Edmund K. 2011. *An Ethnographic Study of the Use of Translation Tools in a Translation Agency: Implications for Translation Tool Design* (Ph. D. Thesis), Kent State University.

Bevan, Nigel. 1995. Usability is Quality of Use. *Advances in Human Factors/Ergonomics* 20: 349-354. doi: https://doi.org/10.1016/S0921-2647(06)80241-8

Candel-Mora, Miguel Ángel. 2011 Computer-Assisted Translation and Terminology Management: Tools and Resources. In Suau-Jiménez, Francisca & Pennock Barry (ed.) *Interdisciplinarity and Languages: Current Issues in Research, Teaching, Professional Applications and ICT*. Bern: Peter Lang Publishing, 145-160.

Covella, Guillermo J., & Olsina, Luis A. 2006. Assessing Quality in Use in a Consistent Way. *Proceedings of the 6th International Conference on Web Engineering - ICWE* '06. http://portal.acm.org/citation.cfm?-doid=1145581.1145583 [Accessed 24/04/2019].

Ehrensberger-Dow, Maureen, and Sharon O'Brien. 2015. Ergonomics of the Translation Workplace: Potential for Cognitive Friction. *Translation Spaces* 4(1): 98-118.

Davies, Patricia M. & Brailsford, Tom J. 1994. *New Frontiers of Learning: Guidelines for Multimedia Courseware Developers in Higher Education*; Volume 1: Delivery, Production and Provision.

Deraniyagala, Rohan, Robert J. Amdur, Arthur L. Boyer & Scott Kaylor. 2015. Usability Study of the EduMod ELearning Program for Contouring Nodal Stations of the Head and Neck. *Practical Radiation Oncology* 5(3): 169-175. Elsevier. doi:10.1016/J.PRRO.2014.10.008.

Dillon, Andrew. 2001. The Evaluation of software usability. In Karwowski, Waldemar (ed.) *Encyclopedia of Human Factors and Ergonomics*. London: Taylor and Francis. https://repository.arizona.edu/handle/10150/105344 [24/04/2019]

Dzida, W., Wiethoff, M. & Arnold, A. 1993. *ERGOguide: the Quality Assurance Guide to Ergonomic Software*. Delft University of Technology.

EEC, EMT EXPERT GROUP. 2009. Competences for professional translators, experts in multilingual and multimedia communication. http://goo.gl/N6Fwya [Accessed 27/04/2019].

González, Jose Luis; García, Roberto; Brunetti, Josep Maria & Gil, Rosa. 2012. SWET-QUM: A Quality in Use Extension Model for Semantic Web Exploration Tools. *Proceedings of the 13th International Conference on Interacción Persona-Ordenador*. doi: https://doi.org/10.1145/2379636.2379651

Göpferich, Sussane. 2009. Towards a Model of Translation Competence and its Acquisition. The Longitudinal Study TransComp. In Göpferich, Susanne; Jakobsen, Arnt Lykke & Mees, Inger M. (ed.) *Behind the Mind: Methods, Models and Results in Translation Process Research*. Copenhagen: Samfundsliteratur, 12-37.

Fu, L, & Schmidt, K. 2006. Usability Evaluation. *International Encyclopedia of Ergonomics and Human Factors* 2019-2022.

ISO. International Standard 9241-1998:11. 2008. Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability.

ISO. International Standard ISO/IEC 25022. 2016. Systems and Software Engineering – Systems and Software Quality Requirements and Evaluation (SQuaRE) – Measurement of Quality in Use.

Kenny, Dorothy. 1999. CAT Tools in an Academic Environment: What Are They Good For? *Target* 11.1: 65-82.

Khalid, Md Saifuddin & Hossan, Md Iqbal. 2016. Usability Evaluation of a Video Conferencing System in a University's Classroom. *19th International Conference on Computer and Information Technology (ICCIT)*. North South University, Dhaka, Bangladesh, 184-190. doi: 10.1109/ICCITECHN.2016.7860192

Kirakowski, Jurek. 1996. The Use of Questionnaire Methods for Usability Assessment. Unpublished manuscript. https://bit.ly/2LkoSUi

Krüger, Ralph. 2016. Contextualising Computer-Assisted Translation Tools and Modelling Their Usability. *Trans-Kom - Journal of Translation and Technical Communication Research* 9(1): 114-148. https://cutt.ly/Sw5a5Dl.

Lagoudaki, Pelagia Maria. 2008. *Expanding the Possibilities of Translation Memory Systems. From the Translator's Wishlist to the Developer's Design* (Ph. D. Thesis), Imperial College London.

LeBlanc, Matthieu. 2013. Translators on Translation Memory (TM). Results of an Ethnographic Study in Three Translation Services and Agencies. *Translation and Interpreting* 5(2): 1-13. doi:10.12807/ti.105202.2013.a01.

Merino, Jorge; Caballero, Ismael; Rivas, Bibiano; Serrano, Manuel & Piattini, Mario. 2016. A Data Quality in Use Model for Big Data. *Future Generation Computer Systems* 63: 123-130. doi: 10.1016/j.future.2015.11.024

Nielsen, Jakob. 1993. *Usability Engineering*. London: AP Professional Academic Press Ltd.

O'Brien, Sharon, Maureen Ehrensberger-Dow, Marcel Hasler & Megan Connolly. 2017. Irritating CAT Tool Features That Matter to Translators. *Hermes* 56: 145-62. doi:10.7146/hjlcb.v0i56.97229.

PACTE. 2005. Investigating Translation Competence: Conceptual and Methodological Issues. *Meta* 50(2): 609-619.

Preece, Jenny; Rogers, Yvonne; Sharp, Helen; Benyon, David; Holland, Simon & Carey, Tom 1994. *Human Computer Interaction*. Essex, England: Addison-Wesley.

Ramírez Polo, Laura & Ferrer Mora, Hang. 2010. Aplicación de las TIC en Traducción e Interpretación en la Universidad de Valencia: experiencias y reflexiones. *Redit: Revista Electrónica de Didáctica de la Traducción y la Interpretación* 4: 23-41.

Vargas-Sierra, Chelo. 2011. Translation-oriented terminology management and ICTs: present and future. In Suau Jiménez, Francisca & Pennock, Barry (ed.) *Interdisciplinarity and languages: Current Issues in Research, Teaching, Professional Applications and ICT*. Bern: Peter Lang Publishing, 45-64. http://goo.gl/dN93hH.

Vargas-Sierra, Chelo & Ramírez-Polo, Laura. 2011. The Translator's Workstations revisited: A new paradigm of translators, technology and translation. *Tralogy: Translation Careers and Technologies: Convergence Points for the Future*, Paris, 3-4 March 2011. http://goo.gl/in0Jmp.