Project funded by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317471.

Project reference: 317471

Project full title: EXPloiting Empirical appRoaches to Translation

D6.5: Final User Evaluation

Authors: Carla Parra Escartin (Hermes), Lianet Sepúlveda Torres (Pangeanic)

Contributors: Constantin Orasan (UoW), Anna Zaretskaya (UMA), Santanu Pal (USAAR), Hernani Costa (UMA), Rohit Gupta (UoW), Liling Tan (USAAR), Varvara Logacheva (USFD), Carolina Scarton (USFD), Liangyou Li (DCU), Chris Hokamp (DCU), Joachim Daiber (UvA), Hoang Cuong (UvA), Hanna Bechara (UoW)

Document Number: EXPERT_D6.5_20160830

Distribution Level: Public

Contractual Date of Delivery: 30.06.16

Actual Date of Delivery: 30.08.16

Contributing to the Deliverable: WP6

WP Task Responsible: UoW

EC Project Officer: Matias Pandolfo

D6.5 Final user evaluation

Carla Parra Escartín[1], Lianet Sepúlveda Torres[2]

[1]Hermes Traducciones, Spain; [1]Pangeanic, Spain
carla.parra@hermestrans.com, lisepul@gmail.com

August 30, 2016

Contents

1 Introduction **2**

2 User Perspective **2**

3 Data collection **7**

4 Language technology, domain ontologies and terminologies **8**

5 Learning from and informing translators **12**

5.1 Learning from translators 12

5.2 Informing translators 13

5.2.1 Quality Estimation 13

5.2.2 Machine Translation Evaluation 17

6 Hybrid corpus-based approaches **18**

7 Conclusion **19**

1 Introduction

The end users of Machine Translation (MT) and translation technologies are manifold and thus their views on what constitutes good or bad quality may differ greatly. While the general public may use MT as a way of understanding a text written in a foreign language and may be more tolerant to errors done by an MT system as long as the message is conveyed, professional translators may find the same errors intolerable. At the same time, some translation tools such as Computer Assisted Translation (CAT) or terminology management tools are mainly used by professional translators or translation project managers. One key aspect when developing these technologies is thus carrying out user evaluations with the aim of assessing to which extent the tool satisfies user needs and requirements.

In this Deliverable, we report on all user evaluations carried out within the EXPERT project. Most of the advances produced in the project have already been presented in previous deliverables[1]. Although some overlapping is unavoidable, here, we focus on user evaluation exclusively and report not only on the user evaluations carried out within the project, but also on the methodologies used for carrying out such user evaluations[2].

In order to facilitate the reading of this Deliverable, it has been organized using the different work packages of the EXPERT project. Section 2 summarizes research in Work Package 2, *User Perspective* and focuses on translators' requirements from translation technologies. Section 3 is devoted to the advances in Work Package 3: *Data Collection*. Section 4 covers Work Package 4, *Language technology, domain ontologies and terminologies*, Section 5 covers Work Package 5, *Learning from and informing translators*, and Section 6, reports on the user evaluations carried out within Work Package 6, *Hybrid corpus-based approaches*. Finally, Section 7 summarizes the main advances in terms of user evaluation and suggests new avenues for exploration in this matter.

2 User Perspective

As described in the Description of Work of the EXPERT project many MT systems force the users to change their working style by imposing the use of sentential segments and not allowing reuse of translations. The proposed EXPERT solution was to "consider the real needs of translators, involving

[1] See http://expert-itn.eu/?q=content/deliverables for a complete list of the Project Deliverables.

[2] For an overview of the work done with regard to corpus-based approaches, please see Deliverable *D6.3. Improved corpus-based Approaches* [Parra Escartín and Sepúlveda Torres, 2016a], and for work on hybrid translation tools please see *D6.4. Improved Hybrid Translation Tool* [Parra Escartín and Sepúlveda Torres, 2016b].

them in the development of technologies, and providing training to prepare them with new skills."

Two Early Stage Researchers (ESR) in the project (ESR1 and ESR2) have worked on this matter. While ESR1's main focus has been on investigating translators' requirements from translation technologies, ESR2 has focused on investigating an ideal translation workflow for hybrid translation approaches.

As part of her work, ESR1 conducted an extensive survey on translators' attitudes and requirements regarding different types of translation technologies. The questionnaire was distributed among professional translators. 736 responses from 88 countries were gathered and most of the respondents were freelance translators. The results of the survey are analyzed in Zaretskaya et al. [2015] and Zaretskaya [2015b]. Deliverable D2.1 [Zaretskaya, 2015a] offers a further analysis of the results along with other work on user requirements.

As far as Translation software evaluation, Zaretskaya [2016] presents an interesting methodology for evaluating translation software using user preferences as features for such evaluation. She applies this evaluation methodology to four popular CAT tools: *SDL Trados Studio 2014*[3], *memoQ 2013*[4], *Memsource Web Editor*[5] (February 2015) and *MateCAT*[6] (June 2015, Federico et al. [2014]). Zaretskaya [2016] suggests an evaluation scheme for CAT tools taking into consideration three software quality characteristics: functionality, adaptability and interoperability. Selected evaluation features related to these three major categories are then used. The features were selected from the list of features used in the original questionnaire on user requirements [Zaretskaya et al., 2015], and on the features indicated by the respondents to the questionnaire as their favorites in CAT tools. She performs weighted and non-weighted evaluation, suggesting that features which are considered more important by translators should also be given a greater weight in the evaluation. The weights, ranging from 1 to 3, were assigned using the average usefulness score obtained by each feature in the questionnaire. The list of features and their assigned weights are summarized in Table 1.

The evaluation results show that *Studio 2014* seems the best CAT tool of the four under study, closely followed by *memoQ*. Although a comparison between a weighted and non-weighted evaluation did not render different results as to which CAT tool seems best, it did show differences in terms of individual categories. In fact, in terms of adaptability, when applying the non-weighted evaluation both *Memsource* and *Studio 2014* are tied, while once the weighted evaluation is applied *Studio 2014* outstands. In terms of

[3] http://www.sdl.com

[4] https://www.memoq.com/

[5] https://www.memsource.com/en

[6] https://www.matecat.com/

Feature		Weight
Functionality	Concordance	3
	Auto propagation	2
	Aligner	1
	Storing TM in the cloud	1
	Real-time QA	2
	Access to online TM	1
	Access to online terminological resources	2
	Sub-segment suggestions	1
	Real-time target preview	2
	Good grammar checker	1
	Merge TMs	1
	Easily add terms	1
	Segment assembly	1
	Dictation	1
	Simple handling of tags	1
	Terminology management	3
	Machine Translation	1
	Work with >1 TM	1
Adaptability	Different OS	1
	Adjustable keyboard shortcuts	2
	Adjustable segmentation	2
	Adaptable/modular interface	1
	Web-based version	1
Interoperability	Share TM	1
	Number of TM formats	3
	Number of document formats	3

Table 1: Features used to evaluate CAT tools and weight attributed to each feature in Zaretskaya [2016].

functionality *memoQ* seems to be the best CAT tool, while *Studio 2014* is better in terms of adaptability and interoperability.

Terminology Extraction Tools (TET) are also of great importance for professional translators. Zaretskaya et al. [2015] reported that only 25% of the respondents to her questionnaire said that they used TET tools. However, these tools may be extremely useful not only for terminology look-up, but also for improving translation coherence and consistency at document level and also across documents at a more general client/product level.

Costa et al. [2016] compare nine different TET tools and the different

	memoQ		Memsource		MateCAT		Studio 2014	
	w.	non-w.	w.	non-w.	w.	non-w.	w.	non-w.
Functionality	**42**	27	28	18	27	18	40	26
Adaptability	8	4	7	5	4	3	**9**	5
Interoperability	8	4	11	5	11	5	**14**	6
Total	58	35	46	28	42	26	**63**	37

Table 2: CAT tools evaluated and scores obtained for each major category [Zaretskaya, 2016]. *w.* stands for "weighted" and *non-w.* stands for "non-weighted".

features they offer. More concretely, they compared *SDL Multiterm*[7]; *SimpleExtractor*[8]; *TermSuit*[9]; *Sketch Engine*[10]; *Translated Labs - Terminology Extraction*[11]; *Terminus*[12]; *KEA (Keyphrase Extraction Algorithm)*[13]; *Okapi Rainbow's Term Extraction Utility*[14]; and *JATE (Java Automatic Term Extraction)*[15]. The features they used were the following:

1. Bilingual term extraction
2. Source and target context comparison
3. Term validation
4. Bilingual dictionaries compilation
5. Context extraction
6. Support various file formats
7. Rank terms by frequency
8. Support for many languages
9. Specify the minimal number of occurrences
10. Show linguistic information
11. Specify the maximum number of translations
12. Stopword list option
13. Choose the minimum and maximum number of words per term
14. Term statistics

[7] http://www.sdl.com/cxc/language/terminology-management/multiterm/extract.html

[8] http://www.dail-software.com/shop/en/9-terminology-extractor-simpleextractor-v112.html

[9] http://termsuite.github.io/

[10] https://www.sketchengine.co.uk/

[11] http://labs.translated.net/terminology-extraction/

[12] http://terminus.iula.upf.edu/cgi-bin/terminus2.0/terminus.pl

[13] http://www.nzdl.org/Kea/

[14] http://okapi.sourceforge.net/Release/Utilities/Help/termextraction.htm

[15] https://github.com/ziqizhang/jate

Table 3, adapted from Costa et al. [2016], summarizes their findings. As may be observed, only *MultiTerm* includes all 14 features taken into account in the comparison. *TermSuit*, *KEA*, *Okapi Rainbow* and *JATE* have 9 of 14 features, and *SimpleExtractor* and *Terminus* have 8. This study may serve as a good starting point for researching translator needs in terms of TET tools and proposing an evaluation methodology similar to the one proposed by Zaretskaya [2016] for CAT tools.

	SDL Multiterm	**Simple Extractor**	**TermSuit**	**Sketch Engine**	**Translated**	**Terminus**	**Kea**	**Rainbow**	**JATE**
Bilingual extraction	yes	-	yes	yes	-	-	-	-	-
Source and target context comparison	yes	-	-	-	-	-	-	-	-
Terms validation	yes	yes	-	yes	-	yes	yes	yes	yes
Bilingual dictionaries compilation	yes	-	yes	-	-	-	-	-	-
Context extraction	yes	yes	yes	-	yes	yes	yes	yes	yes
Support various file formats	yes	yes	yes	yes	-	yes	yes	yes	yes
Rank terms by frequency	yes	yes	yes	-	-	yes	yes	yes	yes
Support for many languages	yes	-	yes	yes	-	yes	yes	yes	yes
Specify the minimal number of occurrences	yes	yes	-	yes	-	yes	yes	yes	yes
Show linguistic information	yes	-	yes	-	-	yes	-	-	-
Specify the maximum number of translations	yes	-	yes	-	-	-	-	-	-
Stopword list option	yes	yes	-	-	yes	-	yes	yes	yes
Choose the minimum and maximum number of words per term	yes	yes	-	-	-	-	yes	yes	yes
Term statistics	yes	yes	yes	yes	yes	yes	yes	yes	yes

Table 3: TET tools feature comparison [Costa et al., 2016].

Another group of users that can benefit from the usage of technologies are interpreters. Costa et al. [2014a] offers an overview of the most relevant features that interpreters would require from terminology management tools both prior to and during interpreting. They compared eight terminology tools for interpreters: Intragloss[16], InterpreBank[17], Intraplex[18], SDL Multiterm[19], AnyLexic[20], Lingo[21], UniLex[22] and The Interpreter's Wizard[23]. The features used for the comparison were those identified as needed by interpreters and are as follows:

- Manages multiple glossaries

[16] https://intragloss.com/

[17] www.interpretbank.de

[18] www.fourwillows.com

[19] www.sdl.com/products/sdl-multiterm/desktop.html

[20] www.anylexic.com

[21] www.lexicool.com/soft_lingo2.asp

[22] www.acolada.de/unilex.htm

[23] the-interpreters-wizard.topapp.net

- Number of possible working languages
- Number of languages per glossary allowed
- Number of descriptive fields
- Handles documents
- Unicode compatibility
- Imports from...
- Exports to...
- Embedded online search for translation candidates
- Interface's supported languages
- Remote glossary exchange
- Well-documented
- Availability
- Operating System(s)
- Other relevant features

The authors include a summary table aimed at helping interpreters to choose the tool that best suits their needs. However, they conclude that "there is a pressing need to design terminology management tools tailored to assist interpreters in the preparation stage, before their interpreting service or during it" and that it is also needed to assess and identify the exact needs of interpreters.

As a follow up of this work, Costa et al. [2014b] focus on computer assisted interpreting and offer an overview of available tools for interpreters. These range from terminology management tools to others such as unit converters and note taking tools.

3 Data collection

One of the aims of the EXPERT project was also to investigate how data repositories can be built automatically in a way that makes them useful to multiple corpus-based approaches to translation. Such new resources can also be compiled taking into account the user perspective, or using potential end-users.

Although some of the data gathered during the EXPERT project cannot be publicly released due to Intellectual Property Rights and Confidentiality Agreements, the experiments run by Parra Escartín and Arcedillo [2015a,b] and Parra Escartín and Arcedillo [2015c] constitute a great resource for investigating the impact of Machine Translation in the translation industry and professional translators. Similarly, experiments run by ESR1 with the aim of comparing Post-Editing difficulty of different Machine Translation Errors in English→Spanish and English→German data[24], constitute an ad-

[24] See Subsection 5.1 for a more detailed report on this work.

ditional resource for similar studies.

Besides running experiments with real users, another way of engaging the users in the data collection process consists of including them as data providers. ESR12 has collected a data set of 1,000 semantically related sentences and their machine translations. The sentences were extracted from the FLICKR images data set used in previous SemEval tasks and consists of sentence pairs annotated by humans in terms of similarity (ranging from 1 to 5), and a French translation created by the state-of-the-art Statistical Machine Translation system Moses [Koehn et al., 2007]. The main objective was to build a data set where the machine translations (Sentence A – MT output, and Sentence B – MT output) of semantically similar sentences in the source language are assigned a quality rating and a semantic relatedness value similar to that between Sentence A and Sentence B.

Two types of human annotations were done:

1. *Translation quality annotation*: for each sentence in the data set, a quality score ranging from 1 to 4 was assigned through manual evaluation. A professional translator was hired to annotate the sentences.
2. *Translation similarity annotation*: for each sentence pair and its respective French machine translations, a similarity score between the translations was assigned. This data was gathered by means of crowdsourcing. A google form[25] was used and invitations to participate were sent to the French department at the University of Sheffield and other Universities in France.

The resulting data set will be useful for tasks such as Quality Estimation and detection of Semantic Text Similarity across languages. Moreover, the annotations about the semantic similarity of the machine translations can be used to explore whether or not semantic similarity carries over through machine translation.

4 Language technology, domain ontologies and terminologies

The translation industry has experienced a massive technification in the past 10 to 15 years. With the appearance of Computer Assisted Translation (CAT) tools, the translation workflow changed radically and embraced new technologies to boost translators' productivity and improve the overall quality of translations thanks to terminological tools, quality checkers and alike.

[25] https://script.google.com/macros/s/AKfycbzodNMqEqKmeRsorLNzwtpBq8l9n1CWHNJOOpgi3xI/dev

A key component of the translation workflow is Translation Memories (TMs) because they help translators both retrieve past translations that are either identical or similar to new segments to be translated. In the case of similar segments, translators can use these past translations as a starting point and correct the differing fragment of the segment to produce the right translation. This way of working, also called fuzzy-match post-editing due to TM leveraging, has one constraint: it is computed based on Levenshtein's Edit Distance (ED) or some flavor of it and thus focuses only on the surface forms of the translated text disregarding synonyms and other semantic or syntactic variants of the same sentence. Gupta and Orăsan [2014] explored ways of improving the existing fuzzy matching algorithms by incorporating paraphrases in the process. Their strategy consists of creating new TM segments by means of paraphrasing the existing ones. They evaluated their system and obtained positive and encouraging results in terms of TM retrieval, and this lead to a further study in which they carried out a human centered evaluation of their proposed system. This study is reported in Gupta et al. [2015b].

Moreover, the work by Gupta et al. [2015b] constitutes the first work "on assessing the quality of any type of semantically informed TM fuzzy matches based on post-editing time or keystrokes". The authors posed three main research questions:

1. How much improvement can paraphrases yield in terms of TM retrieval?
2. What is the quality of the retrieved TM matches with paraphrases?
3. What is the impact of using paraphrases on the work of human translators?

As the authors aimed at successfully integrating their methodology in real translation workflows, instead of creating a new matching algorithm that could be hard to integrate in the existing CAT tools used by professional translators, their strategy consists of analyzing the existing TM, retrieving all TM segments that can be paraphrased, creating new TM segments with the paraphrases and concatenating them to the original TM. In this way, the new enlarged TM can be used without further processing in any CAT tool. When incorporating paraphrases in a TM and subsequently running the ED TM leveraging process, two cases can occur:

(a) Thanks to the paraphrase, the fuzzy match score is increased and the top-ranked match is reinforced.

(b) Thanks to the paraphrase, a fuzzy match which was lower in the ranking gets a higher score and is re-ordered to the top.

While in the first case paraphrasing only reinforces the fuzzy match retrieved, in the second case, it supersedes the existing edit distance model, as in this case the segment presented to the translator for post-editing would be different.

In their experiments, the language pair used was English→German and the translators were Bachelor or Masters students of translation without any technical translation background or expertise. To avoid adding an extra layer of complexity, the corpus used was from a more general domain: Europarl, [Koehn, 2005]. They filtered out the corpus so that all segments were between 7 and 40 words long, and subsequently took a random test set consisting of 9,981 segments. The remaining 1,565,194 segments were used as their original TM.

Table 4 shows the number of segments in the test set found in the TM and their respective fuzzy match bands. *ED* refers to the number of segments retrieved when using only Edit Distance, whereas *+Paraphrasing* refers to the segments retrieved by ED once the paraphrasing approach had been used to enlarge the TM. *% Improvement* refers to the improvement in retrieval obtained when incorporating the paraphrases and *Rise to top* reports on the number of segments that rose to the top as a consequence of using the paraphrasing approach. As the table shows, there is an obvious improvement in terms of TM retrieval when incorporating the paraphrases into the workflow.

	100	**[85, 100)**	**[70, 85)**	**[55, 70)**
ED	117	98	225	70,3
+Paraphrasing	16	30	98	311
%Improvement	13,67	30,61	43,55	44,23
Rise to top	9	14	55	202

Table 4: Retrieval of fuzzy matches from Europarl used as a TM [Gupta et al., 2015b].

In order to further test whether the paraphrasing strategy is actually useful for translators, the authors carried out three different experiments:

1. *Post-Editing time and keystrokes*: Ten different translators were asked to post-edit TM fuzzy matches. All fuzzy matches were computed by simple ED, but two different TMs were used: one with paraphrases, and another one without. Half of the fuzzy matches came from the original TM and half of them originated from a TM including paraphrases.
2. *Subjective evaluation with two options*: Seventeen translators were presented with two different fuzzy matches, one originated from the original TM, and another originating from a TM including paraphrases, and they were asked to choose which segment was better to post-edit.
3. *Subjective evaluation with three options*: Seven translators were asked to choose whether an ED distance fuzzy match or one originating from

the paraphrasing strategy was better, or whether both options seemed equal.

Table 5 summarizes the number of segments included in each of the test sets used for the experiments. They were extracted randomly from the set of segments including paraphrases that as a consequence of this, rose to the top of the ranking. They were created in such a way, that a translator would finish post-editing them in one sitting.

	100	**[85, 100)**	**[70, 85)**
Set 1	2	6	6
Set 2	5	4	7
TOTAL	7	10	13

Table 5: Number of segments per fuzzy match band and test set.

The results of all three experiments showed very positive results. In terms of post-editing time and amount of keystrokes, on average, the paraphrases allowed the translators to save 9.18% of time when compared with simple edit distance. Translators that post-edited the paraphrased segments edited 25.23% less than those who were only presented with segments retrieved with simple ED. On average, these segments required 33.75% more keystrokes and 10.11% more post-editing time.

In the case of the subjective evaluations, the results also seemed to favor the use of paraphrases. In the one with two options, 334 replies chose the paraphrased fuzzy match over the simple ED fuzzy match, which was chosen only 176 times[26]. In the subjective evaluation with three options, in 99 cases the paraphrased segments were tagged as better, in 38 cases the simple ED matches were tagged as better, and in 73 cases there was a tie[27].

The positive results retrieved both in retrieval (c.f. Table 4), as well as in this user evaluation, seem to indicate that adding paraphrases to the TMs in the translation workflow increases the fuzzy matches retrieved. This can be considered a major contribution of the EXPERT project with a direct impact on real translation scenarios, as the adoption of this technique may result in a productivity boost.

Parra Escartín [2015] also explored ways of increasing the TM fuzzy match retrieval by fulfilling the translators' wishes. As a follow up of an internal survey carried out at Hermes, a shallow, language-independent method was implemented and tested in a pilot study where the requests of translators were integrated in the translation workflow.

Her strategy consists of generating new TM segments using formatting and punctuation marks, and the newly generated segments were modifica-

[26] In total, 510 replies were gathered (30 segments * 17 translators).

[27] In total, 210 replies were gathered (30 segments * 7 translators).

tions of the original ones, or fragments of them. In some cases, the strategy simply consisted of splitting an already existing segment in two[28].

The combination of the existing TMs with the newly generated segments decreased the number of segments not started (i.e. the fuzzy match retrieval improved) in all cases, and increased the number of segments translated with the fragment assembly functionality. Moreover, the results showed that "the bigger the TM with new fragments, the higher the number of segments that benefit from fragments".

5 Learning from and informing translators

5.1 Learning from translators

Clear end-users of MT systems are professional translators. In fact, post-editing tasks are more and more common in the translation industry and that is why now it is more crucial than ever to take translators into consideration and include them in the assessment of MT systems. At the same time, if new technologies and tools are introduced in the translation industry, we shall also have to bear in mind their end-users to avoid mistrust and to allow for the enhancement of such technologies and tools.

In this respect, ESR1 has also researched the Post-Editing difficulty of different Machine Translation errors in English → Spanish and English → German. This has been collaborative work between ESR1 and the Research Group from Saarland University, another partner in the EXPERT consortium where ESR1 did a 3-month secondment. This research focused on analyzing two of the three types of Post-Editing (PE) efforts distinguished by Krings [2001]: the temporal effort and the technical effort. While the temporal effort refers to the time a translator takes to post-edit a given segment (i.e. post-editing time), the technical effort refers to the number of edits carried out by the translator when post-editing a given segment. In their study, they use the post-editing effort measure, which is based on the fuzzy match algorithms used in CAT tools and which serves as a reference as to the amount of editing made within the segment.

For their experiments, they used the English → German and English → Spanish corpora annotated with MT errors released within the QTLaunchPad project[29] [Burchardt et al., 2013]. In order to be able to measure the impact of the different types of errors, sentences containing only one error type were selected. The errors were marked in the sentences by means of curly braces to indicate to the post-editors what they had to change in each sentence. 200 sentences were included in the English → German data set,

[28] For a more detailed account of the methodology used by Parra Escartín [2015], see *Deliverable D6.3. Improved corpus-based Approaches* [Parra Escartín and Sepúlveda Torres, 2016a].

[29] http://www.qt21.eu/launchpad/

and 163 sentences in the English → Spanish one. In total, the German sentences account for 1941 words (10 words per sentence), and the Spanish ones for 2347 words (14.4 words per sentence).

19 German and 24 Spanish translation students participated in the experiment. All of them were either in their final bachelor year (12 German and 18 Spanish participants), or studying a masters in translation (7 German and 6 Spanish participants). The experiment was conducted in *MateCAT* and its logs were used to analyze the results.

The results show that the post-editing time is higher for Spanish, while the post-editing effort is lower. However, the higher post-editing time seems to be caused by the sentence length. The average time-per-word (average time taken to post-edit one word) is used to compare the results across languages. Aiming at establishing whether it is more suitable to talk about types of errors or temporal and technical PE effort, the authors also investigated the correlation between PE time and PE effort. The results showed that time-per-word is more related to PE effort than PE time.

Finally, as far as the PE difficulty that the different types of MT errors represent, the results varied across languages and the PE difficulty indicators used (PE time and PE effort). This suggests that PE difficulty of different MT errors varies significantly across languages.

5.2 Informing translators

5.2.1 Quality Estimation

One of the main objectives of the EXPERT project additionally consisted of informing end-users about the quality of machine translations prior to post-editing. This task is usually referred to as Quality Estimation (QE, or MTQE), and it is defined as the prediction of the quality of an automatically translated sentence. QE systems can be trained with a diverse range of data including previous post-edits of the training set. As post-edits are gathered during any MT Post-Editing (MTPE) task, their collection is straightforward and only requires the usage of real projects and professional translators.

Two ESRs (ESR6 and ESR7) have worked directly in MTQE and have made significant progress. ESR6 investigated ways of collecting and extracting useful human feedback for the improvement of statistical machine translation systems. She explored both (i) the improvement of the QE system and (ii) the incorporation of the QE scores in MT systems. On the other hand, ESR7 addressed the challenge of finding the best features for document-level QE, including studies on discourse phenomena and document-wide information. Additionally, ESR7 also focused on finding appropriate labels for measuring the MTQE at document level. These labels go beyond the simple aggregation of sentence-level quality scores.

Scarton et al. [2015b] assess whether popular automatic machine translation evaluation metrics can be used to provide labels for quality estimation at document and paragraph levels. In their work, the authors highlight the fact that such metrics disregard the discourse structure of the text and how this is a major limitation in terms of MTQE at document-level. To further understand the limitations of such metrics, ESR7 designed experiments with human annotators and proposed a way of quantifying differences in translation quality that can only be observed when sentences are judged in the context of entire documents or paragraphs.

Using paragraphs as whole documents, the authors explored two different strategies to collect labels for MTQE using human annotators: the annotation of cohesion and coherence using a scale (*SUBJ experiment*), and a two-stage post-editing task (*PE1* and *PE2 experiments*) in which the context was excluded and included.

In the *SUBJ experiment*, the annotators were asked to assess the quality of paragraphs in terms of cohesion and coherence. Cohesion was defined as "the linguistic marks (cohesive devices) that connect clauses, sentences or paragraphs together", and coherence as the mechanism that "captures whether clauses, sentences or paragraphs are connected in a logical way". The proposed scale ranged from 1 to 4 in each case and is summarized in Table 6.

	Cohesion	**Coherence**
1	Flawless	Completely coherent
2	Good	Mostly coherent
3	Disfluent	Little coherent
4	Incomprehensible	Incoherent

Table 6: Scale used by Scarton et al. [2015b] for annotating the cohesion and coherence of machine translated paragraphs.

The *PE1* and *PE2 experiments* aimed at objectively assessing the quality of MT by means of post-editing. This was done in two rounds. In the first one, the annotators were given the sentences without a context and in the second round the same post-edited sentences were given to the annotators facilitating them in the contexts in which they appeared and they were asked to fix any issues arisen upon reviewing the sentences in context.

The annotators were all students of translation studies familiarized both with MT and with post-editing tools.

The results of the *SUBJ experiment* showed a very low inter-annotator agreement (measured in terms of Spearman's ρ rank correlation). Cohesion ranged from 0.09 to 0.43, and coherence from 0.05 to 0.28, having 0.58 as an outlier. This seems to indicate that the quality assessment of paragraphs in terms of cohesion and coherence is a difficult and very subjective task.

This may have also been due to the fact that the definitions provided to the translators were too vague and, additionally, previous knowledge may have played an important role during the annotation. Nevertheless, the results obtained show that involving humans in the process of MT evaluation is not trivial, and that definitions shall be carefully thought of and established to avoid big inter-annotator disagreement due to factors beyond subjectivity (e.g. vague definitions).

As far as the *PE1* and *PE2 experiments*, the Spearman's ρ rank correlation was calculated comparing the HTER scores obtained comparing the MT output against the first round of post-editing (PE1), and the first post-editing (PE1) against the second one (PE2). No major changes were expected during the second post-editing round (only those required by the context), and this was confirmed by the low HTER scores obtained. The correlations varied from 0.22 to 0.56 in the case of PE1 and MT, and from -0.14 to 0.39 in the case of PE1 and PE2. This shows that annotators strongly disagreed regarding what changes were necessary during the second post-editing round. A manual inspection of the changes made actually revealed that some annotators made unnecessary changes disregarding the task instructions. Nevertheless, the results obtained were very promising and revealed issues that rely on a wider context beyond the sentence level and the need for QE tags for paragraph and document level.

Results obtained by Scarton et al. [2015b] encouraged testing another approach for assessing the translation quality at document-level. In Scarton and Specia [2016], the authors present a reading comprehension corpus compiled with the aim of developing a new quality label at document level. The corpus consists of a range of texts translated by different MT systems (*Google Translate*[30], *Bing Translator*[31], *SYSTRAN*[32], a *MOSES*[33] baseline system, and a mixed MT where sentences from all the MT systems were randomly used) and a human translator, and it was used in a reading comprehension experiment. The underlying hypothesis was that if the readers of the target language were able to reply to a set of manually written reading comprehension questions, the document translation was of good quality. When they failed at the task, the MT output was deemed to be of a bad quality.

The corpus used is an extension of the CREG corpus[34] [Ott et al., 2012]. The authors took the original German documents and translated them into English. The reading comprehension questions were answered by paid volunteers who were fluent in English and staff members and students of the

[30] https://translate.google.com/

[31] http://www.bing.com/Translator

[32] http://www.systransoft.com/

[33] http://www.statmt.org/moses/, Koehn et al. [2007]

[34] http://www.uni-tuebingen.de/en/research/core-research/collaborative-research-centers/sfb-833/section-a-context/a4-meurers.html

University of Sheffield, UK.

The corpus was divided into sets of 6 documents and two scenarios with a differing order of presentation of the MT systems were prepared: *MOSES*, *Google Translate*, *Bing Translator*, *Systran*, the mixed version and the human translation (Scenario 1), and mixed, *Systran*, human translation, *MOSES*, *Bing Translator* and *Google Translate* (Scenario 2). 19 sets were included in each scenario.

The reading comprehension questions were manually classified using the classes in Meurers et al. [2011], focusing on question forms and comprehension types [Day and Jeong-suk, 2005].

- Question Forms (directly defined by the question structure and by the expected answer):
  - Yes/no questions
  - Alternative questions
  - True/false questions
  - Wh-questions
- Comprehension Types (one needs to read the text and identify the answer):
  - Literal questions
  - Reorganisation questions
  - Inference questions

Scarton and Specia [2016] offer an extensive analysis of the test takers agreement, which was calculated using the Fleiss' Kappa metric and the Spearman's ρ correlation coefficient. The results show a "fair" or "moderate" agreement according to Fleiss' Kappa in the majority of the scenarios. The authors also prove that the worse results were not related with the presence of more difficult questions on an specific scenario and also were not correlated with the increase in number of words. The test takers also had a low agreement when answering questions on the human translated documents. On average, *MOSES* was the system that showed the highest agreement, followed by *Bing Translator*. The worst agreement was found for *Systran*.

Although the low inter-annotator agreement prevented the authors from drawing any definite conclusions, they conclude by stating their "hypothesis is that reading comprehension questions can provide valuable information about the quality of an entire machine translated document" and that further research is needed on this matter.

5.2.2 Machine Translation Evaluation

Another important issue to be taken into account when dealing with MT are its evaluation metrics. BLEU [Papineni et al., 2001] and TER Snover et al. [2006] are very popular metrics used along with Edit Distance in the translation industry. In fact, and despite already existing criticism [Callison-Burch et al., 2006], BLEU continues to be the predominant metric used in MT research. Scarton et al. [2015a] explored the correlations of high BLEU and RIBES [Isozaki et al., 2010] –which is more sensitive to reordering– scores with human judgments and showed that a higher score does not correlate better with human judgments. They also carried out a segment-level meta-evaluation with the aim of identifying those segments where their system's high BLEU and RIBES improvements had been deemed substantially worse than the baseline translations.

Gupta et al. [2015a] propose a new evaluation metric based on dense vector spaces and recurrent neural networks, in particular Long Short Term Memory networks. They tested their metric with WMT-14 data and their new metric scored best for two of the five language pairs. As this metric performs well and it does not require the usage of additional external resources, it would be suitable for usage outside of Academic circles as an alternative metric to traditional ones such as BLEU and TER.

Parra Escartín and Arcedillo [2015a,b] and Parra Escartín and Arcedillo [2015c] also explored the role of MT evaluation metrics in real translation settings. To this aim, they engaged professional translators working at Hermes in translation, fuzzy-match editing and MT output post-editing. From the perspective of professional translators and project managers, traditional MT evaluation metrics are difficult to understand. The translation industry would thus welcome an MT evaluation metric with which they are already familiarized and which they understand. Professional translators and project managers use the source-side fuzzy match score to analyze new projects and allocate time and resources. Moreover, this score also determines fare discounts when appropriate due to a productivity increase. As part of the research on MT evaluation taking into account user aspects, ER2 and her colleague Manuel Arcedillo (Hermes) tested whether a target-side Fuzzy Match Score (FMS) can be used as an alternative evaluation metric for MT output [Parra Escartín and Arcedillo, 2015a,b]. Their results show that a target-side FMS correlates with productivity as well as, or even better than, BLEU and TER.

Another important outcome of the research on evaluation including user aspects is reported in Parra Escartín and Arcedillo [2015c], in which the authors study the minimum score at which machine translation evaluation metrics report productivity gains in a machine translation post-editing task and establish the minimum MT quality thresholds for English → Spanish MTPE tasks from the technical domain. They conclude that productivity gains can be achieved with BLEU scores above 45, TER values below 30 and FMS values above 75. The authors also compared equivalent segments from MTPE and TM matching samples and their findings show that when equivalent text editing is involved, MTPE segments require a longer time to post-edit even if the MT/TM output was not modified in the end.

6 Hybrid corpus-based approaches

Prior to EXPERT, "hybrid corpus-based solutions considered each approach individually as a tool, not fully exploiting integration possibilities". The proposed EXPERT solution was to "fully integrate corpus-based approaches to improve translation quality and minimize translation effort and cost."

Several ESRs have worked on this topic within the EXPERT project. In what follows, we present a summary of their work in the field of Automatic Post-Editing (APE) mainly. Automatic Post-Editing is the task which consists of automatically attempting to correct Machine Translation (MT) output so that it is more similar to a human reference. Sometimes, these systems are used as a post-processing module in MT tasks to improve the overall performance of an MT system. One approach towards APE consists of training an MT system with raw MT output (used as the source language), and its corresponding human post-edits (used as the target language). Such APE system thus "translates" an MT output into its corresponding corrections.

Pal et al. [2016a] present an APE system based on a bidirectional recurrent neural network (RNN) model. It consists of an encoder that encodes an MT output into a fixed-length vector from which a decoder provides a post-edited (PE) translation. The results obtained showed statistically significant improvements over the original MT output (+3.96 BLEU), a phrase-based APE system (+2.68 BLEU) and a hierarchical one (1.35 BLEU). The original MT system was Google Translate (English–Italian), and the baseline system was already obtaining an impressive 61.26 BLEU score, which makes it more difficult to beat.

Besides the automatic evaluation, Pal et al. [2016a] run a human evaluation of the output in which four professional translators and native speakers of Italian were asked to rank the output of the MT output and the APE system. As it was not possible to evaluate the whole test set, a subset of 145 randomly sampled sentences were selected. The translators were then presented with the source sentence and the two outputs (*Google Translate* or the *APE system*), and they had to select the best system of the two, or annotate the sentence as 'uncertain' when both sentences were equally good or bad. The two system outputs were presented randomly so that the translators did not know which system they were voting for in each case.

The winning system resulted to be the *APE system* trained by Pal et al.

[2016a]. It received 285 (49.13%) votes, while *Google Translate* got only 99 (17.07%). The remaining votes (196, 33.79%) were classified as ‘uncertain’. A 0.330 Cohen’s κ coefficient [Cohen, 1960] was obtained when computing inter-annotator agreement.

The human evaluation also revealed that their system drastically reduces the preposition insertion and deletion error in the Italian *Google Translate* output, and additionally handles the improper use of prepositions and determiners and reduces word ordering errors to some extent. These observed improvements, together with improvement in automatic evaluation metrics, give reason to believe that the system proposed by Pal et al. [2016a] could be useful for PE tasks in real industrial settings and would reduce human PE effort.

In Pal et al. [2016c], ESR2 and his co-authors test with positive results an Operation Sequential Model (OSM) combined with a phrased-based statistical MT (PB-SMT) system. Their APE system is trained on monolingual data between MT outputs (TL$_{MT}$) produced by a black-box MT system and their corresponding post-edited version (TL$_{PE}$). When evaluated against the official test set of the APE shared task, their system achieves an improvement of +1.99 absolute points and +3.2% relative improvement in BLEU over the raw MT output. In the case of TER, they achieve -0.66 absolute points and -0.25% relative improvement.

They report plans on integrating their APE methods in their Computer Assisted Translation (CAT) tool *CATaLog*, also co-authored by ESR2 [Nayek et al., 2015, Pal et al., 2016b]. Their ultimate goal is to provide better suggestions for post-editing. They also plan to carry out user studies to investigate how integrating APE into a CAT tool impacts human post-editing performance.

7 Conclusion

In this Deliverable, we have reported on the main advances in terms of user evaluation within the EXPERT project. As reported here, users have played an important role in the research carried out in the project and several user evaluations have been done. User requirements have been researched extensively (cf. Section 2), and the advances made during the project have also undergone user studies to measure their impact in real settings. Users have been a key player in some of the data collection efforts carried out (cf. Section 3), and have also been involved in many of the experiments done by the EXPERT researchers. As we have seen, there have been different proposals to improve the TM leveraging process for real translation projects (cf. Section 4), and efforts have been made to both learn from translators (cf. Section 5.1) and informing translators (cf. Section 5.2). Finally, users were also involved in hybrid corpus approaches investigating how to improve

Automatic Post Editing systems that can subsequently be integrated in real translation workflows (cf. Section 6).

The results of the user evaluation studies carried out also highlight the great importance that they should be given in applied research contexts where real users can benefit from the advances done in research projects. The combination of academic and industrial partners within EXPERT has favored the emergence of synergies and has helped researchers understand that at the other end real people will be using the results of their research and the benefits of including them in the development cycle of new technologies and tools. Although not all user studies could be reported here because some are still in the planning stage, this Deliverable is proof of the potential bridges that can be established and how both worlds can learn and benefit from each other.

References

Aljoscha Burchardt, Arle Lommel, and Maja Popovic. Tq error corpus. Technical Report Deliverable D 1.2.1, QT Launchpad Project, 2013.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, 2006. URL http://www.aclweb.org/anthology/E06-1032.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. doi: 10.1177/001316446002000104. URL http://dx.doi.org/10.1177/001316446002000104.

Hernani Costa, Gloria Corpas Pastor, and Isabel Durán Muñoz. A Comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68–76, Dublin, Ireland, August 2014a. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/W14-4809.

Hernani Costa, Anna Zaretskaya, Gloria Corpas Pastor, and Miriam Seghiri. Nine terminology extraction Tools: Are they useful for translators? *MultiLingual 159*, 27(3):14–20, April/May 2014b.

Hernani Costa, Gloria Corpas-Pastor, Miriam Seghiri, and Anna Zaretskaya. Nine terminology extraction tools - are they useful for translators? *Multilingual*, April/May 2016. URL https://multilingual.com/all-articles/?art_id=2327.

Richard Day and Park Jeong-suk. Developing reading comprehension questions. *Reading in a foreign language*, 17(1):60, 2005.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frédéric Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. *The Matecat Tool*, pages 129–132. 2014.

Rohit Gupta and Constantin Orăsan. Incorporating paraphrasing in translation memory matching and retrieval. In *Proceedings of the European Association of Machine Translation (EAMT-2014)*, pages 3–10, 2014.

Rohit Gupta, Constantin Orăsan, and Josef van Genabith. Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, 2015a.

Rohit Gupta, Constantin Orăsan, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. Can translation memories afford not to use paraphrasing? In *Proceedings of the 2015 Conference on European Association of Machine Translation (EAMT-2015)*, Antalya, Turkey, 2015b.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 944–952, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1870658.1870750.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *In Conference Proceedings: the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

Hans P. Krings. *Repairing texts : empirical investigations of machine translation post-editing processes.* Kent State University Press, Kent, Ohio, 2001.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. Evaluating answers to reading comprehension questions in context: Results for german and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Edinburgh, Scottland, UK, 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-2401.

Tapas Nayek, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. Catalog: New approaches to tm and post editing interfaces. In *Proceedings of the 1st Workshop on Natural Language Processing for Translation Memories. Workshop on Natural Language Processing for Translation Memories (NLP4TM), located at RANLP 2015, September 11, Hissar, Bulgaria*, pages 36–43, 2015.

Niels Ott, Ramon Ziai, and Detmar Meurers. Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam, 2012. URL http://purl.org/dm/papers/ott-ziai-meurers-12.html.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. A neural network based approach to automatic post-editing. In *Proceedings of the The 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin (Germany), August 2016a. ACL.

Santanu Pal, Marcos Zampieri, Sudip Kumar Naskar, Tapas Nayak, Mihaela Vela, and Josef van Genabith. Catalog online: Porting a post-editing tool to the web. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016b. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Santanu Pal, Marcos Zampieri, and Josef van Genabith. Usaar: An operation sequetnial model for automatic statistical post-editing. In *Proceedings of the ACL 2016 First Conference on Machine Translation (WMT16)*, Berlin, Germany, 11–12 August 2016c. ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, September 2001.

Carla Parra Escartín. Creation of new tm segments: Fulfilling translators' wishes. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 1–8, Hissar, Bulgaria, September 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-5201.

Carla Parra Escartín and Manuel Arcedillo. A fuzzier approach to machine translation evaluation: A pilot study on post-editing productivity and automated metrics in commercial settings. In *Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 40–45, Beijing, China, July 2015a. ACL.

Carla Parra Escartín and Manuel Arcedillo. Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings. In *Proceedings of the MT Summit XV*, Miami, Florida, USA, October 2015b. International Association for Machine Translation (IAMT).

Carla Parra Escartín and Manuel Arcedillo. Living on the edge: productivity gain thresholds in machine translation evaluation metrics. In *Proceedings of the Fourth Workshop on Post-editing Technology and Practice*, pages 46–56, Miami, Florida, USA, November 2015c. Association for Machine Translation in the Americas (AMTA).

Carla Parra Escartín and Lianet Sepúlveda Torres. Deliverable D6.3. Improved corpus-based Approaches. EXPERT Project, 2016a.

Carla Parra Escartín and Lianet Sepúlveda Torres. Deliverable D6.4. Improved Hybrid Translation Tool. EXPERT Project, 2016b.

Carolina Scarton and Lucia Specia. A reading comprehension corpus for machine translation evaluation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3652–3658, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Carolina Scarton, Liling Tan, and Lucia Specia. Ushef and usaar-ushef participation in the wmt15 qe shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 336–341, Lisbon, Portugal, September 2015a. Association for Computational Linguistics. URL http://aclweb.org/anthology/W15-3040.

Carolina Scarton, Marcos Zampieri, Mihaela Vela, Josef van Genabith, and Lucia Specia. Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation. In *EAMT 2015*, pages 121–128, Antalya, Turkey, 2015b.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August 2006.

Anna Zaretskaya. Deliverable D2.1: User requirement analysis. Technical report. Deliverable in the EXPERT project, 2015a. URL http://expert-itn.eu/sites/default/files/outputs/expert_d2.1_20150210.pdf.

Anna Zaretskaya. The use of machine translation among professional translators. In *Proceedings of the EXPERT Scientific and Technological Workshop*, Málaga, Spain, May 2015b. EXPERT: EXPloiting Empirical appRoaches to Translation.

Anna Zaretskaya. A quantitative method for evaluation of cat tools based on user preferences. In *Proceedings of the AELFE XV International Conference*. University of Alcalá, June 2016.

Anna Zaretskaya, Gloria Corpas Pastor, and Miriam Seghiri. Translators' requirements for translation technologies: Results of a user survey. In *Proceedings of the AIETI7 Conference. New Horizons is Translation and Interpreting Studies (AIETI)*, Málaga, Spain, 2015.