Paper presented at Machine Translation Summit XIV. 2-6 September 2013, Nice, France

User Evaluation of Advanced Interaction Features for a Computer-Assisted Translation Workbench

V. Alabau and **J. González-Rubio** and **L.A. Leiva**
D. Ortiz-Martínez and **G. Sanchis-Trilles** and **F. Casacuberta**
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
Camí de Vera s/n, 46021 Valencia (Spain)
{valabau, jegonzalez, luileito}@dsic.upv.es
{daormar, gsanchis, fcn}@dsic.upv.es

B. Mesa-Lao and **R. Bonk** and **M. Carl** and **M. García-Martínez**
Center for Research and Innovation in Translation and Translation Technology (CRITT)
Copenhagen Business School
Dalgas Have 15, 2000 Frederiksberg (Denmark)
{bm.ibc, rbo.ibc, mc.ibc, mgm.ibc}@cbs.dk

Abstract

This paper reports on the results of a user satisfaction survey carried out among 16 translators using a new computer-assisted translation workbench. Participants were asked to provide feedback after performing different post-editing tasks on different configurations of the workbench, using different features and tools. Resulting from the feedback provided, we report on the utility of each of the features, identifying new ways of implementing them according to the users' suggestions.

1 Introduction

Machine translation (MT) technology has been playing an increasingly important role within translation over the past six decades. Nowadays its impact is undisputedly extensive and has reached an unprecedented level that deserves careful consideration as a crucial factor which affects human translators in the first place.

The use of MT systems for the production of post-editing drafts has become a widespread practice among many Language Service Providers (LSPs). This is confirmed by an extensive market study (TAUS, 2009) in which industry practices were surveyed in regard to translation automation in 129 LSPs. 40% of the surveyed LSPs reported that they are already using MT, while 89% of the remaining 60% reported that they were planning to integrate MT in their translation processes within the following two years.

The reasons for this increase in the adoption of MT technology are diverse. Apart from the productivity gains in the translation industry reported by several studies (de Almeida and O'Brien, 2010; Plitt and Masselot, 2010; Guerberof, 2012), there are many other reasons behind such a recent MT adoption. Some of these reasons could be a greater availability of resources and tools for the development of MT systems, a change in the expectations of MT users, as well as a successful integration of MT systems in already well-established computer-assisted translation (CAT) workbenches.

Traditionally post-editing workflows only take into account the human component in a serial process (Isabelle and Church, 1998). First the MT system provides complete translations which are then proofread by a human translator. In such a serial scenario, there is no actual interaction between the MT system and the human translator, making it impossible for the MT system to benefit from overall human translation skills and preventing the human translator from making the most out of the adaptive ability of some MT systems.

An alternative to this traditional workflow is represented by the interactive machine translation (IMT) approach (Langlais and Lapalme, 2002; Casacuberta et al., 2009; Barrachina et al., 2009). In the IMT approach, a fully-fledged MT engine is embedded into a post-editing workbench allowing the system to look for alternative translations whenever the human translator corrects the MT output. MT technology is used to produce full target sentences (hypotheses), or portions thereof, which can be interactively accepted or edited by a human translator. The system continues search-

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 361–368.

ing for alternative renditions as the translator edits the text. The MT engine then exploits the changes made by the translator to produce improved outputs, and provides the user with fine-tuned completions of the sentence being translated.

IMT can be seen as an evolution of the statistical MT (SMT) framework (Koehn, 2010b). Within the IMT framework, a state-of-the-art SMT system is used in the following way. For a given source sentence, the SMT system automatically generates an initial translation. A human translator checks this machine translation, correcting the first error. The SMT system then proposes a new completion or suffix, taking the correct prefix into account. These steps are repeated until the whole input sentence has been correctly translated.

The present study reports on a user evaluation of an IMT workbench being implemented as part of the CASMACAT project[1]. Research was devised so as to investigate user satisfaction while post-editing MT outputs using a translation workbench featuring different tools and resources. The ultimate aim of testing these different configurations was to assess their potential and decide which of them can be successfully integrated into the second prototype of the CASMACAT workbench for the benefit of the human translator. This study also aimed at fine-tuning some the IMT features tested in light of the feedback provided by the users.

Improving and maximizing the potential of a post-editing workbench is one of the priorities set by both the industry and researchers when addressing the technological challenges faced by human translators. The motivation behind this research ultimately comes from a desire to know how such tools can be of greater support to translation professionals, and how technology can even empower them to make an unrestrained choice of the translation methods, strategies and tools they feel comfortable with and which bring out the best of their skills (Mesa-Lao, 2012).

2 Background research

Human translator interaction with MT technology draws back to the emergence of the first effective MT systems (Vasconcellos and León, 1985). Traditionally this human-computer interaction involves the human translator as a post-editor (proofreader) of MT outputs, but rarely involves the human translator guiding the decisions of a MT system. Recent seminal efforts on building interactive MT systems include Langlais et al. (2000) and Barrachina et al. (2009). Both studies develop research systems looking into a tighter integration of human translators in MT processes by developing a prediction model that interactively suggests translations to the human translator as she types. Similar work was carried out by Koehn (2010a), displaying different translations to human translators and letting them choose the one that better suited their needs for post-editing.

An important contribution to IMT technology was pioneered by the TRANSTYPE project, where data driven MT techniques were adapted for their use in an interactive translation environment. Langlais et al. (2002) performed a human evaluation on their interactive prototype emulating a realistic working environment in which the users could obtain alternative renditions as they were typing to fix MT outputs. In this study, post-editors' productivity decreased by 17%, but they appreciated such an interactive system and declared that it could help them to improve their productivity after proper training.

In line with the aims of the TRANSTYPE project, Barrachina et al. (2009) also worked with the IMT approach by using fully-fledged MT systems to produce MT hypotheses. Translators could choose new suggestions from the SMT system as they were correcting MT outputs. Each corrected output was used by the system as additional information to achieve future improved suggestions. Further research has also been carried out as part of the TRANSTYPE2 project (Casacuberta et al., 2009). In this project, post-editors' performance tended to increase as they became acquainted with the system over a 18-month period.

A slightly different approach was studied in Koehn (2010a), where monolingual users evaluated a translation interface supporting predictions and the so-called "translation options". On Arabic-English and Chinese-English, using standard test data and current SMT systems, 10 monolingual users were able to translate 35% of Arabic and 28% of Chinese sentences correctly on average, with some of the participants coming close to professional bilingual performance on some of the texts.

[1] CASMACAT: *Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation.* Project co-funded by the European Union under the Seventh Framework Programme Project 287576 (ICT-2011.4.2).

3 Workbench Features

For the purpose of the evaluation we decided to implement a web-based prototype supporting IMT features. Web-based applications present several advantages. Firstly, they provide a powerful and mature environment to implement dynamic interfaces with advanced visual features. Secondly, they can be easily deployed worldwide reaching virtually anyone. For this purpose, we leveraged the MATECAT post-editing interface (Bertoldi et al., 2012), which is an open source web application. On top of their interface, we implemented the visualization of the advanced features, connected to our IMT servers. Figure 1 shows the implemented CASMACAT interface with some features that we believe are desirable in any IMT-based workbench.

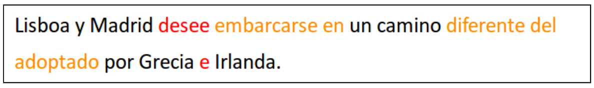
In the following subsections we present a short description of the main features that were implemented in the prototype. Such features are different in nature, but all of them aimed at facilitating the post-editing process.

3.1 Intelligent Autocompletion

IMT with intelligent autocompletion takes place every time a keystroke is detected by the system (Barrachina et al., 2009). In such an event, the system produces a (full) suitable prediction according to the text that the user is writing. This new prediction replaces the remaining words of the original sentence at the right of the text cursor.

3.2 Confidence Measures

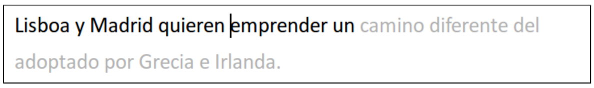
Current MT systems are still far from perfect. It would be thus desirable to improve their use by adding information on the reliability of the output produced. A way to do so would be by highlighting chunks of translated text that, according to the system knowledge, are not reliable enough (González-Rubio et al., 2010). In the CASMACAT workbench, we use confidence measures to inform post-editors about the reliability of translations under two different criteria. On the one hand, we highlight in red those translated words that are likely to be incorrect. We use a threshold that maximizes precision in detecting incorrect words. On the other hand, we highlight in orange those translated words that are dubious for the system. In this case, we use a threshold that maximizes recall.



Lisboa y Madrid desee embarcarse en un camino diferente del adoptado por Grecia e Irlanda.

3.3 Prediction Length

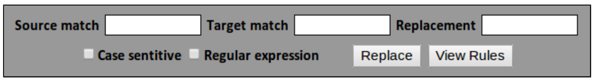
Providing the user with a new prediction whenever a key is pressed has been proved to be cognitively demanding (Alabau et al., 2012). For this reason it was decided to limit the number of predicted words that are shown to the user by only predicting up to the first word with a low CM according to the system. In our implementation, pressing the Tab key allows the user to ask the system for the next set of predicted words, painting in gray the remaining words in the suggested translation.



Lisboa y Madrid quieren emprender un camino diferente del adoptado por Grecia e Irlanda.

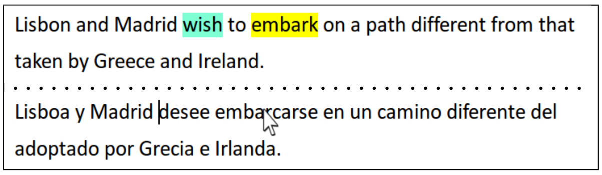
3.4 Search and Replace

Most of the computer-assisted translation tools provide the user with intelligent search and replace functions for fast text revision. The CASMACAT workbench also features a straightforward function to run search and replacement rules on the fly. Whenever a new replacement rule is created, it is automatically populated to the forthcoming predictions made by the system, so that the user only needs to specify them once.



3.5 Word Alignment Information

Alignment of source and target information is an important part of the translation process (Brown et al., 1993). In order to display the correspondences between both the source and target words, this feature was implemented in a way that every time the user places the mouse (yellow) or the text cursor (cyan) on a word, the alignments made by the system are highlighted.

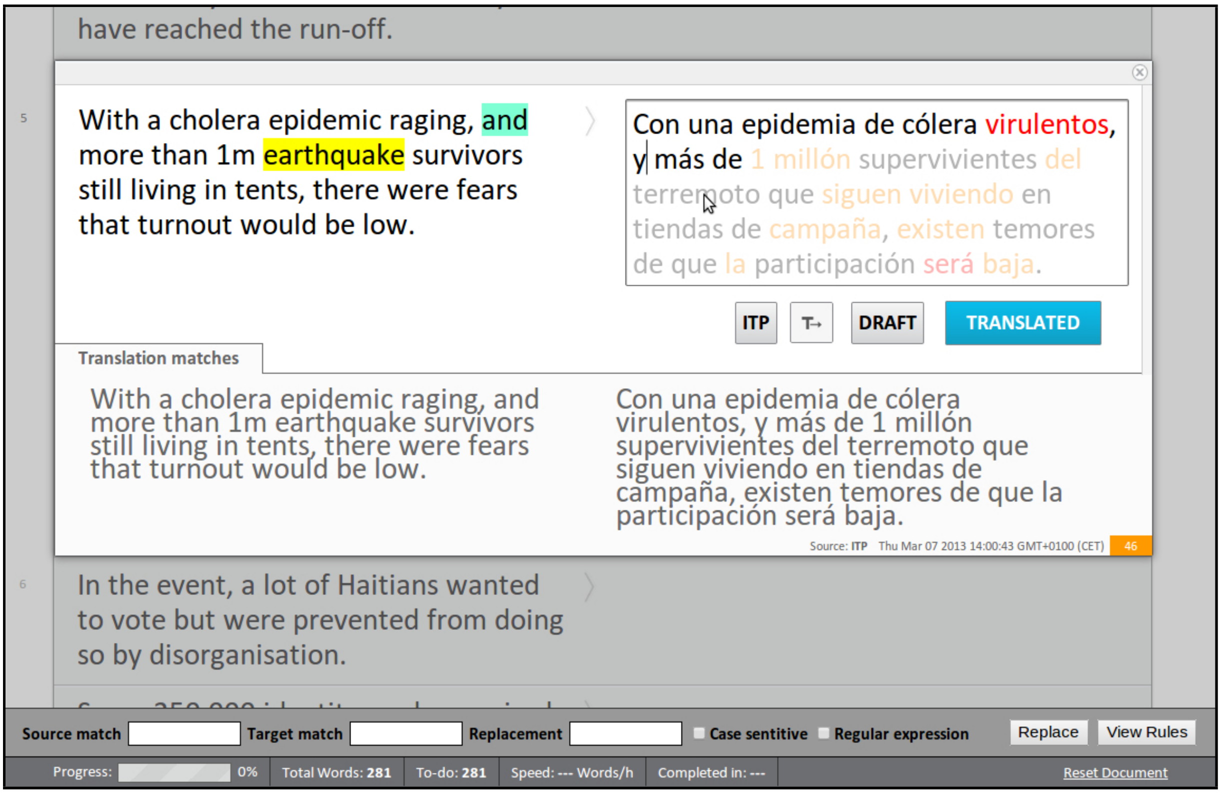


Lisbon and Madrid wish to embark on a path different from that taken by Greece and Ireland.

Lisboa y Madrid desee embarcarse en un camino diferente del adoptado por Grecia e Irlanda.

3.6 Prediction Rejection

With the purpose of easing user interaction, our prototype also supports a mouse wheel rejection feature (Sanchis-Trilles et al., 2008). By scrolling



have reached the run-off.

5

With a cholera epidemic raging, and more than 1m earthquake survivors still living in tents, there were fears that turnout would be low.

Con una epidemia de cólera virulentos, y más de 1 millón supervivientes del terremoto que siguen viviendo en tiendas de campaña, existen temores de que la participación será baja.

Translation matches

With a cholera epidemic raging, and more than 1m earthquake survivors still living in tents, there were fears that turnout would be low.

Con una epidemia de cólera virulentos, y más de 1 millón supervivientes del terremoto que siguen viviendo en tiendas de campaña, existen temores de que la participación será baja.

Source: ITP Thu Mar 07 2013 14:00:43 GMT+0100 (CET) 46

6

In the event, a lot of Haitians wanted to vote but were prevented from doing so by disorganisation.

Progress: 0% | Total Words: 281 | To-do: 281 | Speed: --- Words/h | Completed in: ---

Figure 1: Screenshot of our workbench with all its features enabled.

the mouse wheel over a word, the system invalidates the current prediction and provides the user with an alternate translation in which the first new word is different from the previous one.

4 User Evaluation

The main goal of this research was to measure user satisfaction when performing post-editing tasks using different workbench features (see Table 1). In this context, we were interested in knowing whether translators find the use of such features useful while post-editing MT outputs.

4.1 Workbench Configurations

For this purpose, we defined four different configurations of the workbench (see Table 1). Each of them differs in the set of features that are included (see section 3). System 1 (S1) was a baseline system for IMT including only basic intelligent autocompletion. Systems 2 to 4 (S2–S4) included intelligent autocompletion together with some of the advanced features described above.

4.2 Participants Profile

A group of 16 users (10 females and 6 males) aged between 21 and 34 volunteered to perform the evaluation of the different systems. All participants had a degree in translation studies and were regular users of computer-aided translation tools (i.e., SDL Trados and MemoQ), but they had never used IMT technology to post-edit. When asked about previous experience in post-editing of MT outputs, 55% of claimed to have previous experience in post-editing assignments. This difference in post-editing experience was not considered a bias in the sample of the study, since the aim was not to measure productivity but user satisfaction.

4.3 Questionnaires

A system usability scale (SUS) questionnaire was used to collect quantitative data on user satisfaction. Users had to asses each system in a typical five-level Likert scale, with five denoting the highest satisfaction, right after performing a post-editing task in each of the four different systems. In addition to the Likert scale, each questionnaire also included a text area for users to submit additional comments and feedback on the feature being tested. A final overall questionnaire was also filled out in order to know which of the four configurations of the workbench was most preferred.

	Systems			
Workbench features	**S1**	**S2**	**S3**	**S4**
basic intelligent autocompletion (IMT)	*			
IMT + confidence measures		*		
IMT + prediction length control			*	
IMT + search and replace function				*
IMT + word alignments				*
IMT + prediction rejection				*

Table 1: List of the workbench features included in each of the four evaluated systems (S1 to S4).

4.4 Source Texts

The source texts compiled for this user evaluation were short pieces of news that are likely to appear in any general scope newspaper, extracted from the News Commentary corpus[2]. No expert knowledge was thus required in order to successfully perform the post-editing task. The language pair involved was English to Spanish.

4.5 Procedure

Each system was tested using a different data set consisting of 20 segments each; two pieces of news per system. Before performing the evaluation, participants were asked to fill out an introductory questionnaire in order to collect data about their profile as professional translators, as well as their previous experience in post-editing. The evaluation always involved System 1 in the first place, since it was considered as a baseline prior to testing the advanced IMT features implemented in the other systems. The evaluation of Systems 2, 3, and 4 was done in a randomized order in order to minimize the effect of any ordering on user satisfaction (i.e., due to learning or fatigue effects). The presentation of the different source texts was also randomized along the different systems so as to avoid the potential effect of text difficulty on the evaluation of the system. No time constraints were imposed on the participants involved in the evaluation.

5 Results

From the submitted questionnaires, an overview of user satisfaction for the different systems is shown in Figure 2 following the above described five-level Likert scale, where 5 denotes the highest satisfaction. For each of the evaluated systems, we display the average of the satisfaction scores given by the users (blue box), the 95% confidence interval for the average satisfaction score (black whisker), and the actual distribution of user satisfaction scores (gray pattern). The baseline system (System 1) was given an average satisfaction score of 2.4. In comparison, System 2 was given a slightly worse satisfaction score (2.1) while both System 3 (3.3) and System 4 (2.9) scored clearly above the baseline. Moreover, the confidence intervals for System 1 and System 3 do not overlap.

Overall, the most popular workbench feature among participants was the one implemented in System 3 (with prediction length control). Participants seemed to favor the idea of editing chunks of information while having such a visual aid; i.e., showing in black the text that has been already post-edited and showing in gray the text that stills needs revision. As stated by one participant, *"[...] This feature guided me in the post-editing process, having a greater control of what I had actually edited in the text. I didn't have the feeling that the system was making too many changes at a time and I felt more in control of the editing process"*. System 2, featuring confidence measures (red for wrong and orange for dubious translations), recorded the lowest user satisfaction scores. However, some participants reported in the open-ended questionnaire that this feature seems to be very promising if a more reliable implementation was deployed. *"I could definitely benefit from this type of visual aid, but the system stills need to make better predictions. Many times the words marked by the system as wrong were actually, while wrong translations remained in black. In the end I had to double-check most of the sentences to make sure that words marked in black were actually acceptable translations"*, stated one participant.

None of the participants rated the baseline sys-

[2] `http://www.statmt.org/wmt12`

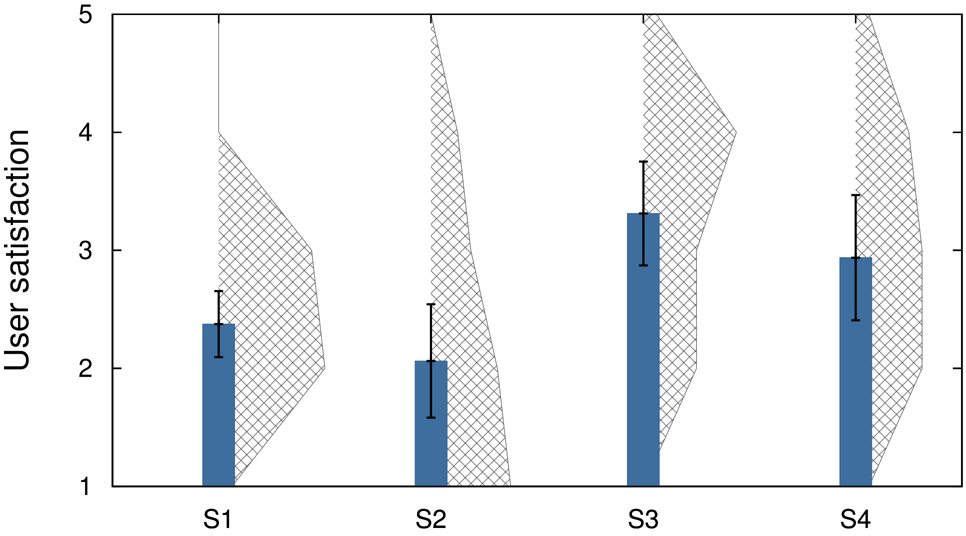


Figure 2: Average user satisfaction reported by the users for each system (S1–S4). We additionally display in black the 95% confidence interval for the average satisfaction score, and in gray the actual distribution of the satisfaction values given to each system.

tem above 3 and actually 50% of them were dissatisfied with the translations produced. These poor results could be attributed to the fact that System 1 was used as a baseline (featuring basic intelligent autocompletion with no advanced features), and therefore it was always evaluated in first place. Users would have certainly benefited from a warm-up session to become acquainted with IMT previous to the formal evaluation.

In line with previous findings by Barrachina et al. (2009) and Casacuberta et al. (2009), the more the participants became familiar with the system, the less the system was perceived as being cumbersome. Feedback recorded in the open-ended questionnaire showed that on-demand word alignment, implemented in the System 4, was very positively perceived by the users as a real aid to spot sources of mistranslations.

Another important finding was the fact that, as most of the participants were experienced touch-type translators, some of them reported that it would have been faster for them to type longer strings of text instead of having to interact with the help of IMT system. In this regard, some of them suggested an extra feature for enabling and disabling IMT depending on the segment that is being post-edited.

In addition to the general findings described above, the feedback provided by the users contains valuable information that can be used to guide the future development of the CASMACAT workbench. Next sections describe the lessons learned about each of the features and tools included in the workbench.

5.1 Confidence Measures

The clarifications made by the users revealed that the main problem of this feature stems in the tendency of the system to classify as incorrect words that, from the translator point of view, are clearly correct. For example, proper names are usually classified as incorrect since they tend to appear few times, if any, in the training data. Such errors are infrequent, so they do not penalize much the performance of the confidence measure as evaluated in most automatic measures. However, these errors are quite annoying for the users who then distrust the confidence information provided by the system.

Users also provided us with feedback on how to display the confidence measures computed by the system. All participants agreed that the color selection was adequate, allowing for an easy identification of potential wrong translations in red and dubious in orange. However, they had mixed opinions regarding the usefulness of showing both wrong and dubious equivalents. Five users considered confidence measures for dubious equivalents (words in orange) a source of visual noise. They pointed that it is only useful to highlight confidence measures for clearly wrong equivalents (words in red). The rest of the participants preferred both thresholds (wrong and dubious equivalents) to be displayed. As a consensus, it seems that translators should be provided with both options and let them decide which of these options, if not both, they want to use.

5.2 Prediction Length

In contrast to the criticism received by the system including confidence measures, the system retraining the prediction length did yield positive satisfaction ratios, even though the length of the prediction is set according to the same confidence measures. Users stated that this feature eased their interaction with the system, by reducing the stress involved in deciding upon the acceptability/correctness of the (sometimes quite different) completions provided by the system.

Some users commented that the limitation imposed by this feature to the autocompletions was a good indicator of what had actually been edited in the text. Nevertheless, this was not the intended purpose of this feature, but this seems to suggest that users would find useful a specific feature targeted to identifying already edited words. For instance, already edited words could be highlighted in green or a special symbol could be used to display the last position of the caret.

5.3 Search and Replace

Although the evaluation did not present enough sentences to the users so that the search and replace feature could be actually assessed, it was perceived positively. Translators agreed in that it is indeed a must in any professional workbench. So far, our search and replace module operates on the autocompletions provided by the system by dynamically applying replace rules. However, since the traditional search and replace feature is perceived as so valuable, future work will be addressed to find different ways of integrating it into the CASMACAT workbench.

5.4 Word Alignment Information

Word alignment information was considered to be quite useful. However, user opinions were mixed regarding the utility of the different visualization options. One frequent comment was that the alignment information triggered by the cursor position can be considered a source of distraction during the translation process as aligned words kept changing as the user edited the MT output. Therefore we conclude that word alignment information should only be displayed on user demand. For instance, it could be shown only when the user presses a given keyboard shortcut.

5.5 Prediction Rejection

This feature also received positive reviews by most of the participants on this user evaluation. Nonetheless, some users reported that the implemented interaction mechanism was somehow unexpected. They would have expected the rejection operation to affect only the word under the cursor, instead of operating on the whole of the remaining sentence to the right of the cursor. Users suggested that this prediction rejection feature should be limited to single words (i.e. looking for alternative equivalents) instead of triggering further changes at the sentence level. Some users also commented that, instead of having to jump from prediction to prediction before finding the right one, a drop-down list would be preferable. Such an implementation of this feature could show several predictions at a time, making the interaction with the system faster. This suggestion, however, challenges the TRANSTYPE2 findings (Langlais and Lapalme, 2002), where drop-down lists were perceived as too overwhelming by the participants in the study. Further research is still needed on how best we can present predictions to the user.

6 Summary and Conclusions

This user evaluation of features for an IMT-based workbench has proved to be successful in addressing the actual benefits of automating interactivity between the MT system and the human translators. In this sense, the surveyed translators provided us with valuable feedback from real users in order to fine tune some of the tested features. One of the key findings of this user satisfaction study is the lack of agreement from most of the translators about which features they want to see implemented in a workbench to make post-editing a more rewarding task. This is certainly a crucial issue that needs further consideration by both human translators and tool developers. Overall, the workbench configuration that translators seem to be more satisfied with is the one featured in System 3 (with prediction length control). Further research is still needed with different user profiles as well as with more hours of interaction with the different features of the workbench.

Acknowledgments

Work supported by the European Union 7th Framework Program (FP7/2007-2013) under the CasMaCat project (grants agreement n^{o} 287576), by Spanish MICINN under grants TIN2009-14205-C04-02 and TIN2012-31723, and by the Generalitat Valenciana under grant ALMPR (Prometeo/2009/014).

References

Alabau, Vicent, Luis A. Leiva, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. User evaluation of interactive machine translation systems. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 20–23.

Barrachina, Sergio, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Bertoldi, Nicola, Alessandro Cattelan, and Marcello Federico. 2012. Machine translation enhanced computer assisted translation. First report on lab and field tests. Available from: http://www.matecat.com/wp-content/uploads/2013/01/MateCat-D5.3-V1.2-1.pdf.

Brown, Peter F, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Casacuberta, Francisco, Jorge Civera, Elsa Cubel, Antonio Luis Lagarda, Guy Lapalme, E. Macklovitch, and Enrique Vidal. 2009. Human interaction for high quality machine translation. *Communications of the ACM*, 52(10):135–138.

de Almeida, G. and Sharon O'Brien. 2010. Analysing post-editing performance: correlations with years of translation experience. In *Proceedings of the 14th annual conference of the European Association for Machine Translation (EAMT)*. St. Raphael, France, 27-28 May.

González-Rubio, Jesús, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2010. On the use of confidence measures within an interactive-predictive machine translation system. In *Proceedings of the 14th annual conference of the European Association for Machine Translation (EAMT)*. St. Raphael, France, 27-28 May.

Guerberof, Ana. 2012. *Productivity and quality in thepost-editing of outputs from translation memories and machine translation*. Ph.D. thesis, Tarragona: Universitat Rovira i Virgili.

Isabelle, Pierre and Ken Church. 1998. Special issue on: New tools for human translators. *Machine Translation*, 12(1/2).

Koehn, Philipp. 2010a. Enabling monolingual translators: post-editing vs. options. In *NAACL HLT 2010 - Human Language Technologies: Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 537–545.

Koehn, Philipp. 2010b. *Statistical Machine Translation*. Cambridge University Press.

Langlais, Philippe and Guy Lapalme. 2002. TransType: development-evaluation cycles to boost translator's productivity. *Machine Translation*, 17(2):77–98.

Langlais, Philippe, George Foster, and Guy Lapalme. 2000. TransType: unit completion for a computer-aided translation typing system, applied natural language processing. In *Applied Natural Language Processing (ANLP)*, pages 46–51.

Mesa-Lao, Bartolomé. 2012. The next generation tranlator's workbench: post-editing in CasMaCat v.1.0. In *Translating and the Computer Conference Proceedings, 34*.

Plitt, Mirko and Franois Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Sanchis-Trilles, G., Daniel Ortiz-Martínez, Jorge Civera, Francisco Casacuberta, Enrique Vidal, and Hieu Hoang. 2008. Improving interactive machine translation via mouse actions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

TAUS. 2009. LSPs in the MT loop: current practices, future requirements [report]. Available from: http://www.translationautomation.com/reports/lsps-in-the-mt-loop-current-practice-future-requirements.

Vasconcellos, Muriel and Marjorie León. 1985. Spanam and engspan: machine translation at the pan american health organization. *Computational Linguistics*, 11(2-3):122–136.