In-context online localisation tools: the case study of Reverso Localize

Cristina Peron, Lucía Morado Vázquez
Département de Traitement Informatique Multilingue
Faculty of Translation and Interpreting
University of Geneva, Switzerland
peroncristina89@gmail.com; lucia.morado@unige.ch

Abstract
Traditional computer-assisted translation tools have been widely used to translate HTML content, and specific localisation tools have also been developed to allow the process of website localisation to be done entirely online with the help of Machine Translation (MT) and by offering an in-context experience. The aim of our work was to evaluate one of the latter tools (Reverso Localize) according to three EAGLES (1996) criteria, namely functionality, usability and efficiency. We carried out an experiment with model-users who localised a website into Italian and post-edited the MT output. The analysis of our results showed that the tool was very responsive and that its two best performing and appreciated functionalities were the interaction system and the in-context post-editing scenario. On the other hand, the usability between other studied functionalities such as the MT output still showed room for improvement. Consequently, we can infer that under specific circumstances this type of tool can represent a real alternative to applying a professional localisation process, but their limitations should also be taken into consideration, and the technical breach that they represent should not be underestimated either.

Keywords: localisation, website localisation, in-context localisation, online localisation tools, machine translation, in-context post-editing, free localisation software, reverso

1. Introduction
The rise of the World Wide Web has made distances significantly shorter and has had a considerable impact on market dynamics. Nowadays, a company can sell its products anywhere in the world and its website is very often the first form of introduction to, and interaction with, a potential client (De Bortoli and Maroto 2003). Although English still plays a crucial role in digital communication, several studies have demonstrated that users tend to have a better opinion of a company – that is, they trust and appreciate it more – when they can browse its website in their mother tongue, whether they can understand English or not (Tong in De Bortoli and Maroto 2003). This is the reason that both multinational companies and small and medium enterprises, seeking to expand their business abroad, are increasingly investing into creating multilingual corporate websites (Valdés 2008). Moreover, according to a study by T. Schewe, a company’s marketing strategy and its localisation choices are strictly connected (Sandrini 2005). Therefore, website localisation can be considered as “a function of the international marketing strategy” (Sandrini 2005: 4).

Given its significant impact on a company’s or an organisation’s image, website localisation is a complex process that implies special attention from both a technical and a cultural point of view. As well as traditional computer-assisted translation (CAT) tools, a new line of tools that promises to render localisation an accessible task to web developers has started to emerge. In this study, we evaluate Reverso Localize (from now on RL), an online tool for in-context web localisation, developed by the Reverso-Softissimo team, that promises to enable web owners to have their site localised in several minutes without requiring professional localisers to be involved.

The structure of this paper is as follows: section 2 introduces RL, the tool being studied and offers a general overview of its use; section 3 presents the methodology put into place to evaluate the tool and offers the results of the experiments carried out; our conclusion is presented in section 4.

2. Reverso Localize: motivation and purposes of the research
Reverso Localize is an online localisation platform that uses MT and allows the user to post-edit the raw MT output in context. It was developed by Reverso-
Softissimo as part of the European project Flavius – which started in November 2009 – and was completed and became operational in November 2012. The Reverso-Softissimo team has since been delivering frequent updates and improvements; hence, we must state that all the data contained in this article refers to the platform’s development status as of April 2013, when we carried out our research.

Our choice of this platform was especially motivated by the nature of the tool, which allows the post-editor to work in context. In addition, we aimed at testing the real advantages that this type of tool can represent to the web localisation process as well as its weaknesses and limitations. At the time of our experiment, the free version of the tool allowed the localisation of a website up to 3,500 KB in size (around 30 pages of formatted text) into a maximum of two languages. Twelve languages were available as both source and target and the user could choose between two different translation scenarios: mirror site or file translation. In both cases, a built-in MT engine produced a raw translation of the source text. However, in-context post-editing (PE) was only available when choosing the mirror site option, and in Figure 1 explain how the mirror site worked and were accessible from the platform:

As explained in Figure 1, a localisation project can be set up quickly by typing the URL of the website to be localised, selecting the language combination and adjusting the settings. The project is managed from the dedicated dashboard. Prior to the translation task, the original content is spellchecked. RL produces a machine-translated version and it then creates a mirror version of the website, that is, a machine-translated copy that can be post-edited in context (see Figure 2). At this point, a post-editor can be invited, or the user himself can decide to proofread the MT output. In the first case, the user and the post-editor work together on the same project within the RL interface, and the user can monitor the PE completion status and the edits. Once the PE is finished, a link provided by RL has to be copy-pasted into the source code of the original website in order to publish and index its localised version. On the platform blog, the user can find instructions on how to publish static and dynamic (Wordpress and OverBlog) sites.

The key aspects of RL philosophy are its ease of use

![Figure 1. The mirror site.](image)

so this option was therefore chosen for our study. As indicated by RL developers, the mirror site was mainly aimed at novice users; on the other hand, the file translation was better adapted for more professional localisation processes. The instructions and its speed. Indeed, the software is intended for both professional and non-professional users and the company’s challenge is to allow them to localise a website without any prior technical skills and within a very short timeframe. Since this seemed to match
the current market needs, the main goal of our research was, on the one hand, to estimate the platform’s internal performance and, on the other hand, to evaluate user satisfaction. Based on our results, we then assessed the positive aspects of the tool and those that could be improved upon from the point of view of system functionality and system-user interaction. Having said this, a more general goal of ours was to spark a discussion on this new type of tool and lay the groundwork for the creation of an evaluation method.

3. Experimental research

3.1 Methodology

Although the evaluation of translation tools is generally thought of as a very important research field, a dedicated, standard and recognised testing method still has to be developed (Quah 2006). According to Quah, this is due to the plethora and variety of available tools: new systems, featuring diverse functionalities and aiming at different goals are constantly being launched into market. Hence, it is not possible to design a global method that could be applicable to every type of system and, at the same time, that would be detailed enough to be perfectly adjustable to each tool. Since no method met all our requirements – that is, there was no specific method for evaluating a localisation platform with a built-in machine translation engine – we adopted the one developed by the EAGLES (Evaluation of Natural Languages Processing Systems) working group in 1996. This method applies to all language technologies and is based on ISO (International Organisation for Standardisation) 9126 standard. In addition, EAGLES 7-step recipe (EAGLES 1999), describing the seven fundamental steps of an effective evaluation of a natural language processing tool, served as a model for our study.

As recommended by the 7-step recipe, we first defined the motivations and the objectives of our study and then we designed the structure and the parameters of the evaluation, which took a field experiment as its starting point. Eight volunteers participated in it: four users, who worked on the main localisation project and four post-editors, whose task involved adapting the raw MT output. The experiment consisted of localising the Geneva Youth Hostel website (http://www.yh-geneva.ch/) from French into Italian, applying the mirror site scenario. The chosen website was built with the Joomla! content management system (version 1.5). Users worked on the whole French content of the website, whereas post-editors only focused on the machine-translated content of the homepage. All participants were students at the Faculty of Translation and Interpreting (FTI), University of Geneva, Switzerland. Therefore, they were not – or not yet – localisation or post-editing professionals.
The next stage required by the 7-step recipe was to define our own evaluation criteria. We decided to evaluate the platform according to the six criteria proposed by EAGLES (1996) that, in turn, follow the experimental task for each group, both users and post-editors worked on the same project.

As shown in Figure 4, prior to the main experimental

<table>
<thead>
<tr>
<th>EAGLES criterion</th>
<th>Objective evaluation</th>
<th>Subjective evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Functionality</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notification and interaction</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Spellchecker</td>
<td>Precision</td>
<td></td>
</tr>
<tr>
<td>Localisation system</td>
<td>• Localisation issues</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Omissions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Silence</td>
<td></td>
</tr>
<tr>
<td>Machine translation and post-editing systems</td>
<td>Evaluation of the raw MT output and the post-edited version through SAE J2450 standard</td>
<td>Questionnaires</td>
</tr>
<tr>
<td>Mirror site</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>Usability</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Mouse click count</td>
<td>Screen recording analysis</td>
<td></td>
</tr>
<tr>
<td>• Tab count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Keystroke count</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Efficiency</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Internal speed</td>
<td>Screen recording analysis</td>
<td></td>
</tr>
<tr>
<td>• External speed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Evaluation methodology based on the 3 EAGLES criteria.

high quality principles established by the ISO 9126 standard: functionality, usability and efficiency. Each criterion was evaluated both objectively and subjectively; which means that our results were derived from three different data analysis methods: first, an objective analysis/extraction, carried out following the principles of systematic test (EAGLES 1996); second, an objective analysis of the material obtained in the scenario test (ibid.); and third, an analysis of the impressions and opinions of our participants (see Figure 3).

3.2 Workflow, tools and Content
The experiment reproduced a real-life localisation scenario using the platform. It was divided into two parts, respectively devoted to the two main parties involved in a localisation project using RL: users and post-editors. Although we designed a different task, all participants had to fill out a general background questionnaire aimed at collecting data on their profile. Data obtained from this questionnaire allowed us to determine that all participants shared similar characteristics in terms of previous experience in web localisation. They were then asked to access the Geneva Youth Hostel website using the Mozilla Firefox web browser and get a first glimpse at its structure and content.

Following this, they had to log into the RL website, read through the home page content and basic information about RL and watch an introductory video in English about the platform. Then, the user needed to take the following steps: log into their account on the platform, type the Gmail address and the related password provided by the researchers; set up a localisation project and adjust the project settings; analyse the source text spelling report,
automatically generated by the platform spellchecker; examine the localisation output inside the mirror site; invite a post-editor through the RL invitation system and let them modify the raw Italian MT output. The last task was dedicated to publication and consisted of linking the localised version to the original site. When this experiment was carried out, RL did not provide specific instructions on how to publish a localised version of a Joomla!-based website, as was our case. To overcome this, we drafted a new instruction sheet, taking a document about static sites that was available on the platform as a reference. As a result, users worked offline – as if on a static website – in order to link the localised version to the website code. As the user clicked on the Publish button, RL created a link leading to the localised version of the site. The user then had to download the source code of the original content, open the HTML file with an advanced text editor (Notepad++ in our case) and replace the French code portion with the newly created link, as well as changing “Français” for “Italiano”. They also had to replace the French flag icon with the Italian one that was stored in a folder on their desktop. At the end of the process, it was possible to browse the localised version of the website offline.

As for post-editors, after watching the introductory video, they had to log into the Gmail account where they received their invitation. The automatically-
generated email prompted them to click on a link in order to access the localisation project and start proofreading the translated content inside the mirror site. Once there, they were asked to correct the raw MT output in context by clicking on the pencil symbol of any incorrect segment (Figure 5) and editing text and/or links inside a dedicated post-editing window (Figure 6).

When working on RL, the user is constantly notified of any updates and/or editing tasks through email and on-screen notifications displayed on the platform interface. Participants could choose to follow the project status via one of those two notification channels. Once the experiment was completed, they all had to fill in a task questionnaire aimed at collecting their opinion on the platform and its performance.

3.3 Analysis and results
As explained in the methodology description, the three chosen evaluation criteria were verified in both an objective and a subjective manner.

3.3.1 Objective evaluation
The objective analysis of Functionality focused on the performance of the following components and content types: the spellchecker, the localisation system, the raw MT output and the post-edited text. As for the first one, RL provided a spellchecking report of the source text, where errors were classified by type and displayed together with some context and a correction suggestion. We rated the precision of this functionality by calculating the number of real errors contained in the corpus. The spellchecker detected 96 French mistakes on a total 8319 words (data collected on April 14th 2013), only 37 of which were real errors, while the remaining 59 were noise: therefore, the tool scored a 38.5% precision rate.

In order to evaluate the functionality of the localisation system, we concentrated on three types of errors found on the mirror site: localisation problems, omissions (elements that were not present in the localised version) and silences (elements that should have been localised but remained in the original language). We recognised three localisation problems: a character encoding issue (Figure 7, No. 1), two overlapping elements (Figure 7, No. 2) and a truncation issue (Figure 7, No. 3).

As visible in area No. 1, both the character (&) and its entity reference (amp;) were displayed in the localised version. In area No. 2, the overlapping of the magnifying glass and the search bar was due to tabs being longer in the Italian version than in French, therefore taking up almost all of the available space. Finally, in area No. 3, the word truncation issue could have been due to the fact that the Italian segments inside the box were shorter than the original phrases, which probably caused the following text to move back up. RL, however, did not offer a way to fix those three spatial issues to either users or post-editors. As for omissions, we only detected one such error that, nevertheless, affected the whole website: tooltips that are usually displayed on mouseover were not visible in the mirror site and they were neither translated nor shown once the localised site was published. Some recurring localisation problems, instead, were in fact silences. Firstly, it was not possible to edit the dynamic elements such as banner texts or Flash animations as
localisation processes may have been used. In some cases, the tool did not have direct access to the original code of such elements; in a regular localisation process, they are usually modified separately with the aid of specific tools, such as advanced image editors. Secondly, a number of website pages that were only accessible via a link on another page (an estimated 30% of total content) were left entirely untranslated. In all likelihood, the reason for this is that RL could only extract data down to a certain depth of the website tree structure. As a result, users browsing the Geneva Youth Hostel localised website were likely to visit one or more subpages containing French text. According to Schiller (2006), such problems can have a high impact on the site credibility, damaging the company-customer relationship and even leading the user to exit the website and never visit it again. This localisation issue was therefore rather serious, especially if we consider that the user could not do anything about it.

As a reference for evaluating MT quality, we chose the SAE (Society for Automotive Engineering) J2450 standard that defines the following seven categories of errors, seen as unacceptable in a translated text: wrong term (WT), syntactic error (SE), omission (OM), word structure or agreement error (SA), misspelling (SP), punctuation error (PE) and miscellaneous error (ME) (SAE J2450 2001). Errors in each category can be classified as Major ( _M) or minor ( _m).

As shown in Table 1, we detected 48 errors over 28 segments, corresponding to a total of 180 words. Wrong terms were the most represented category (63% of errors), followed by miscellaneous errors (19%), among which we included wrong prepositions and articles, as well as untranslated words. We point out that the system did not make any punctuation or

<table>
<thead>
<tr>
<th></th>
<th>WT</th>
<th>SE</th>
<th>OM</th>
<th>SA</th>
<th>SP</th>
<th>PE</th>
<th>ME</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>_M</td>
<td>19</td>
<td>6</td>
<td>Ø</td>
<td>1</td>
<td>Ø</td>
<td>Ø</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>_m</td>
<td>11</td>
<td>1</td>
<td>Ø</td>
<td>1</td>
<td>Ø</td>
<td>Ø</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>TOTAL</td>
<td>30</td>
<td>7</td>
<td>Ø</td>
<td>2</td>
<td>Ø</td>
<td>Ø</td>
<td>9</td>
<td>48</td>
</tr>
<tr>
<td>% on total word count</td>
<td>16,6</td>
<td>3,8</td>
<td>Ø</td>
<td>1,1</td>
<td>Ø</td>
<td>Ø</td>
<td>5</td>
<td>26,5</td>
</tr>
</tbody>
</table>

Table 1. Errors in the raw MT output according to SAE standard.
spelling errors, and we did not identify any omission. Nevertheless, we also note that only 5 segments out of 28 were completely error-free. According to the reference standard (SAE J2450 2001), the error rate is measured with the following formula:

\[(\text{Error No. } 1 \times \text{error weight}) + (\text{Error No. } 2 \times \text{error weight}) + \ldots \]

source text word count

In our particular case, the tool received a score of 0.9, that is, 90% of errors. It is therefore obvious that the raw MT output could have never been published without “full post-editing” (Allen 2003). If we consider the significant editing effort required by the post-editor, this could reopen the debate on the usefulness of machine translation. We will deal with this specific point in the subjective part of the evaluation. We should also state that the MT engine is an external element of RL and therefore it can be changed and improved.

In order to compare MT and PE performance, we also examined the four versions post-edited by our volunteers, taking the SAE standard once again as a reference. The post-edited texts, made up of a total 180 words, contained an average 6 errors, nearly half of which were categorised as minor. It is clear that MT had a great impact on the total error count of the post-edited versions: 12 errors out of 25 were connected to MT. For example, as we observed after examining screen recordings, a post-editor failed to see a specific error twice (but corrected the exact same mistake in other occasions), possibly because he had grown accustomed to seeing the wrong terms displayed on the mirror site: these missed corrections cost him two major wrong term errors, since the raw MT was not acceptable.

For processing MT errors that occurred more than once, post-editors could use the system’s relaunch feature in order to apply changes to the whole content of the website. However, this functionality only worked for 100% matches, while other segments – the most part in our case – could only be edited manually. In the same way, we could state that the PE system was quite rudimentary, as it did not provide any functionality for shifting words or for automatically searching and fixing errors.

In order to objectively evaluate the Usability criterion, we made use of BB FlashBack (Express version), a tool developed by Blueberry Software and used for recording activity on a computer screen and extracting useful data for further investigation (e.g., mouse movement, mouse click count, keystroke count and shortcuts used, as well as the time spent on a specific action). To develop the usability criterion, we relied on three aspects related to ergonomics: mouse click count (including scrolling), keystroke and tab count. We calculated the first by working out the total click count, the number of clicks made by each participant during the most important tasks and those required for scrolling and tab switching. The number of clicks performed was similar among all participants and indicated a certain effort level, which was significantly low for some tasks (e.g., interaction between the user and the post-editor while working on the same project) and much more intense for others (project setup, use of notification system, post-editing). This being said, we noticed that the click count was much higher in the experiment with post-editors, while the PE window was the feature that had the biggest impact on the total. Indeed, the PE system did not provide any advanced post-editing features (an easy way to shift words, regular expressions, etc.); this forced the post-editor to make extensive use of the mouse. What is more, a mouse click was also required for validating each translation, as the Enter key could not be used for this. As a result, the quality of the raw MT output and a rudimentary post-editing system were the two elements that had an impact on ergonomic comfort, as low-quality translations required a greater effort from post-editors even from the physical point of view. In our case, we observed that it took them an average of 100 mouse clicks to post-edit a text of 180 words.

During the analysis of recordings, we also calculated the number of tabs opened while working on RL; indeed, having to shift among tabs is likely to have an impact on the total click count, on the amount of time dedicated to a task and on the global ergonomic comfort of the platform. Users opened two RL tabs on average, while post-editors opened five. This was especially due to how RL worked. Indeed, every time that a user ran a certain task, such as accessing the mirror site, relaunching a translation, opening the updated version after relaunching or accessing the latest edits through an email notification, a new tab was automatically opened, even if there was already a tab dedicated to that action. This was detrimental to the platform usability for two reasons. Firstly, working with multiple tabs of the same program could be misleading for the user, this happened to one participant who struggled to find his way back to the tab that he was working on – one of six open tabs – for more than a minute. Secondly, as shown on Table 2, tabs also had an impact on the number of mouse clicks, therefore influencing ergonomic comfort in both ways.
To complete the objective evaluation of Usability, we extracted an XML file containing keystroke and shortcut logging information from BBFlashback. This file gave us an interesting overview of how participants worked, as well as of the RL performing mechanism. We noticed that, for instance, users often resorted to using arrow keys when navigating around the page code, during the publication task. As for post-editors, they mainly used the backspace key for deleting wrong MT solutions and letter keys for editing them, as well as a great number of arrow keys for navigating around the text. These data indicated that tasks having the highest impact on ergonomic comfort were PE and publication.

In order to evaluate the tool’s Efficiency, we relied once again on the video recordings collected using BBFlashback. We determined, specifically, the tool’s internal speed, namely the average time that it took for RL to complete a given task, and its external speed, that is, how long it took participants to carry out a particular task in the platform. The former was mainly linked to data extraction speed.

At each new project setup, RL carried out four tasks: extracting the data of the original site, spellchecking the source content, translating into the target language and finally displaying the post-editing and the publication windows. Data extraction usually took a short period of time: project setup (including spellchecking, translation and mirror site creation) took an average of 1:13 minutes, while an average of 10 seconds was necessary to create a link that was used to publish the localised version. Looking at these data, one can safely state that the tool’s internal speed was undoubtedly one of its major highlights, making it possible for a user to setup a ready-to-post-edit localisation project in slightly more than a minute. If, however, RL failed during project setup – as happened in one of our experiments – the only solution was to start over again, which inevitably reduced ergonomic comfort.

To determine the tool’s external speed, we took into account the main tasks carried out by the participants and the time spent on each of them. The collected data about external speed were in line with those on data extraction and we can state that they were not high at all: an average of 1:34 minutes was required for setting up a localisation project, and users needed 34 seconds on average to interact with post-editors who, following this, could work on the project without any additional support. What is more, looking at the total duration of the experiment, we can safely say that all users managed to setup and publish a localised website in less than half an hour (excluding the PE work). The time spent on PE, however, must also be taken into account: it took post-editors 25:51 minutes on average to revise a text that was 180-words long. According to Vasconcellos and León (in O’ Brien 2010), the daily output of a professional can range between 4,000 and 10,000 words, when doing “light” post-editing. Since our volunteers were not professional post-editors, we rather chose to refer to other data that set an estimated daily post-editing output of 3,000 to 9,000 words (O’ Brien 2010).

Therefore, assuming that they worked 8 hours per day at the pace recorded in our experiment, our post-editors could had processed around 3,320 words per day, a number that is barely acceptable according to O’Brien’s estimates.

The publication task, which took 8:43 minutes on average, contributed to increasing the total duration of the experiment. Indeed, we noticed that the participants struggled to complete this assignment. This is no surprise, as having to tinker around with code, deleting and copy-pasting code portions in the editor window, is obviously not a straightforward process. However, we cannot blame RL for this difficulty, since that method was specifically designed by the researchers to tackle a scenario that was not addressed in the platform instructions. Although RL states on its website that no technical knowledge is needed to use the tool, we ask ourselves if it would actually have been possible for a non-professional to figure out how to localise Joomla!-based websites and several other types of sites, for which no instructions were available at the time of the experiment. We can therefore conclude that, even though the platform itself is a responsive tool, working on it is not always a quick job, as it greatly depends on the type of task being performed and the characteristics and nature of the original website.

**3.3.2 Subjective evaluation**

The subjective evaluation, based on the results of task questionnaires filled out by participants at the end of the experiment, aimed at collecting the opinion of model users about the platform, and their first impressions. Its first part was the same for users and post-editors and focused on the tool’s usability and efficiency, as well as on their general satisfaction level. The questions regarding functionality obviously differed, depending on the respondent role: users were asked to rate the spellchecker and the
localisation and publication systems, while post-editors had to give their opinion on the MT results and the post-editing system. Question types were rather varied: Likert scales (where participants expressed their level of agreement or disagreement on an agree-disagree scale for a series of statements) were the most frequent type used, but we also included yes/no, ranking and open questions. Answers to almost all of the questions – excluding the open ones – were measured on a 1-6 scale, 6 being the highest or best score.

As far as Usability is concerned, RL scored a total average score of 4.2. Participants were particularly satisfied with the very low learning effort required, the effectiveness of instructions and the interface, which they described as simple and easy to use. However, they showed some perplexity about whether they could have completed all the tasks without the instruction sheet provided, and they gave the corresponding question a lower score (3.75).

To evaluate Efficiency, users and post-editors were asked if the time spent on localisation and post-editing was acceptable and if they thought that RL could allow a localiser or post-editor to save time. The final score for this criterion was 4.4. According to participants, the time spent on their tasks was acceptable, or even short for one of them. All users agreed that RL allowed localisers to save time, every one of them giving positive scores and an average score of 5; post-editors, however, gave a much lower average score of 3. This result can be explained by linking it to the Functionality criterion and, particularly, to the respondents’ answers on the raw MT output. There, they stated that the MT slowed their work progression and that, in general, they would have preferred to translate from scratch. In other words, they thought that the time spent working on the platform was by and large acceptable, but the tool did not allow for a significant time-saving.

Among functionalities, the best rated one was the mirror site (average: 5.5), which was described as useful, easy to use and innovative, and received only positive scores. The “Notifications and interaction” functionality also scored well, especially for allowing users and post-editors to interact with each other, collaborate on the same project and see the edits in real time. The notification system, on the contrary, was criticised on some points. As already said, every time that a user completed an action, he/she was notified both via email and on-screen within the RL interface. The user could not configure nor modify those default parameters. However, half of participants declared that, if they could have chosen one or the other, they would have only kept the on-screen notification system activated. From these answers, we can infer that participants may not have liked receiving an email notification for something that they could have checked within the RL interface itself. All the more so as checking one’s inbox meant having to open multiple browser tabs and losing concentration.

The localisation system received an average score of 4. Participants found it efficient (causing few issues in the localised version) and well-structured. However, users were less enthusiastic when talking about the publication task; although they did not consider it too challenging, they stated that it would have taken them longer – or several attempts – to complete it without the instruction sheet designed by the researchers. Another element that scored quite poorly was the page footer that contained the platform logo, flags for changing the language and a string saying: “This website has been automatically translated by the Reverso Localize platform”. Post-editors pointed out that their role was omitted by this sentence and they found it unfair that it gave all the credit to Reverso MT system when, in fact, its output was edited by a human.

The spellchecker received a 3.8, which was slightly below the pass score. Users noted that the real error-noises ratio was reasonable, the relevance of correction suggestions was acceptable and that, by and large, this latter functionality could be quite useful.

Finally, MT scored an average of 2.8. Users who evaluated the performance of the system gave it an average of 1.75 and claimed unanimously that MT slowed down their work rate, the majority of them saying that they would have rather translated the source text from scratch. We should state that comparing MT+PE against translation from scratch did not fall under the scope of our research and therefore we did not do any empirical test in that sense; but it would be an interesting path to follow in the future, especially after having heard our participants opinions on that topic.

4. Conclusions

On the basis of our study and the collected data, we will now try to draw some conclusions regarding the localisation platform. On the whole, we can state that the intended goal was accomplished: non-
professional users had localised/post-edited a website in less than half an hour using RL.

Let’s now have a look at the most effective features of the system and those that could be improved. Among the evaluated functionalities, the post-editing environment (that is, the mirror site) was the best performing as both users and post-editors found it useful, straightforward and, above all, innovative. Indeed, the in-context post-editing scenario brought by RL introduced a groundbreaking concept in the panorama of existing CAT/localisation tools on the market. As for the interaction system, it was equally appreciated by participants thanks to its high cost-effectiveness.

The localisation system still needed to be improved on some points: although users appreciated its approach, they also identified some irreversible errors that hinder the tool’s performance. A user-system interaction option should be introduced in order to let users fix interface issues (for instance, the platform could incorporate a field and window resizing feature or a method for translating possible silences). What is more, the publication task was not yet exhaustively documented on the platform; instructions were not available for every website type, therefore affecting non-professionals’ ability to work independently. The spellchecker was another tool with mediocre performance. Despite being well-structured and easy to use, it only achieved a 38.5% precision rate. On the other hand, technically speaking, in RL the spellchecker is an external element, which means that RL can easily incorporate another spellchecker with a better performance if required.

MT and PE systems were, however, the less performing features of the platform. The reason being that the target text required a thorough and continuous intervention from the post-editors, who lacked access to functionality that could increase their work rate and complained that it took too many mouse clicks to edit each segment. Participants gave their lowest – and well below the pass – scores to the MT system. The internal functions of the software should be improved, for instance, allowing users to customise the dictionaries. In a similar way, it would be important to reinforce the PE system with a word-shifting feature, as well as automatic search patterns and error correction. As matters stand now, both the MT and PE systems seemed to have a negative impact not only on the tool’s global performance, but also on its usability and efficiency.

As for Usability, we can draw two conclusions. Due to its structure and the way it was designed, the platform implied a considerable effort in terms of necessary mouse clicks and tab shifting. On the other hand, RL required a low learning effort as instructions were clear and the interface was straightforward.

The Efficiency of the tool can be considered one of its main advantages. The platform turned out to be noticeably fast; its external speed was quite good for some tasks, such as project setup and interaction, but it was not as optimal on others, such as PE and publication.

Even though the evaluated tool was an innovation in its field, and it was therefore difficult to compare it with other existing software, or to collect data about future trends, we can conclude that RL came as a real novelty, featuring key points like the mirror site, the interaction and localisation systems, and the cost-effectiveness of the program.

The downsides to the program were the MT output, the PE system and some aspects of its usability. If those last aspects were to be improved, RL would become a very competitive program, taking into account that it is free and easily accessible. However, as with everything web-related it is likely to change at a fast pace. Indeed, as of the date of the publication of this article, several updates and modifications have been made to the software, compared to when the study was performed. An interesting direction for future investigation would be to evaluate the tool once again, in order to measure the impact of the updates and compare the new results with those of this study. It would also be useful, then, to increase the number of participants and repeat the experiment with different parameters of the tool, to see if the trends differ.

Hence, RL is quite unique in its kind, and it integrates various technologies. Consequently, we could not base our study on a specific model or standard designed for such a tool, nor could we take previous works as a reference, or compare it to similar platforms. This work constitutes, therefore, an introductory investigation, the results of which could, in the future, be compared to those of other types of tools. Finally, it could also become a blueprint for evaluation, because it has allowed for the narrowing down of the functionalities that are vital for these tools.
Acknowledgements

This work was carried out in 2013 under the supervision of Prof. Pierrette Bouillon and its extended version was defended as the Master thesis of Cristina Peron at the University of Geneva.

We would like to acknowledge all the other individuals that collaborated with this research: the Reverso Localize team for their assistance and support; the experiments’ participants who donated their time; and Miss Mary Sue Giachetti, for her help in the editing process.

References


