Measuring the Impact of Automated Evaluation Tools on Alternative Text Quality: a Web Translation Study

Rodriguez Vazquez, Silvia

How to cite

RODRIGUEZ VAZQUEZ, Silvia. Measuring the Impact of Automated Evaluation Tools on Alternative Text Quality: a Web Translation Study. In: Proceedings of the 13th Web for All (W4A) Conference. Montreal (Canada). New York : ACM Press, 2016. doi: 10.1145/2899475.2899484

This publication URL: https://archive-ouverte.unige.ch//unige:83932
Publication DOI: 10.1145/2899475.2899484

Measuring the Impact of Automated Evaluation Tools on Alternative Text Quality: a Web Translation Study

Silvia Rodríguez Vázquez
Cod.eX Research Group - Department of Translation Technology
Faculty of Translation and Interpreting - University of Geneva
40, Bd. du Pont d'Arve - 1211 Geneva 4 - Switzerland
Silvia.Rodriguez@unige.ch

ABSTRACT
The number of Internet users has increased tenfold since the beginning of the century up to present, especially thanks to the improvements experienced in web accessibility and the growing number of languages which online content is available in. While translation professionals are making a considerable contribution to that digital information richness, little evidence exists regarding their involvement in the achievement of a more accessible web for all. In this paper, we present the main results of the first empirical study on web accessibility conceived around a translation task. The experiment sought to particularly investigate the quality of image text alternatives produced by French translators with the help of two evaluation tools: *aDesigner* and *Acrolinx*. The assessment of their alt text proposals, carried out by seven screen reader users, suggests that using both tools helps translators to create more appropriate text alternatives than when trying to do so with only one tool or without any automated support. A more in-depth analysis of the data gathered shows that *Acrolinx* offers better guidance than *aDesigner* for translators to render images accessible.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation]**: User Interfaces – *Evaluation/Methodology.* I.7.1 [**Document and Text Processing**]: Document and Text Editing – *Languages, Version control.* K.4.2 [**Computers and Society**]: Social Issues – *Handicapped persons/special needs*. K.7.4 **[The Computing profession]**: Occupations.

General Terms
Experimentation, Measurement, Documentation, Human Factors, Languages, Verification.

Keywords
Web translation, web accessibility, image text alternatives, evaluation tools.

W4A'16, April 11-13, 2016, Montreal, Canada
© 2016 ACM. ISBN 978-1-4503-4138-7/16/04…$15.00
DOI: http://dx.doi.org/10.1145/2899475.2899484

1. INTRODUCTION
In an increasingly globalized digital society, the multilingualism of the World Wide Web is far from going unnoticed. Behind are the early days of the Internet era, when users had to read English online content because there were few other alternatives. Today, websites adapt to the languages spoken by digital information and services consumers worldwide. According to Wikimedia Statistics, Wikipedia's articles are available in more than 230 languages. Almost half of Facebook users write in languages other than English. Twitter had its interface translated into more than 30 languages. The list of enterprises that decided to embrace web globalization could be endless. As Folaron proposes, the Web is now a 'space of translation' par excellence, where translation practices play a key role in keeping global and local networks alive, by maintaining a fluid communication among their users [13].

While the automated generation of translated content −through machine translation (MT) engines like the popular Google Translate− and crowdsourcing or volunteer-driven translation initiatives −such as GlobalVoices[1] or The Rosetta Foundation[2]− are contributing substantially to that flow of information across countries, languages and cultures, society still expects and trusts professional human translation services to be of higher quality, confidentiality and reliability. This is particularly applicable in the case of translation-related activities that go further than just manipulating and adapting textual content, such as website localization. Understood as the process of modifying an existing website to make it accessible, usable and culturally suitable to a specific target audience [37], localization's ultimate goal is the proper functioning of the translated site (*ibid*).

This process may involve making not only textual changes, but also technical and visual modifications to the original site. The latter means that translators' actions could both imply a threat to web accessibility or be considered of added-value. The images case can serve as an example to illustrate this assertion. If a given image of the source web page being localized is accessible −that is, a text alternative is available to provide the information or functionality conveyed through the image to a screen reader user−, one would predict that the translator will maintain that same mechanism in the target web page and correctly translate its content. However, the translator might also neglect the alt text existence, damage the page coding or be careless when suggesting a translation proposal. Similarly, the opposite can occur. If the image was not accessible, a translator with accessibility background or the adequate tools would be able to spot that

[1] http://globalvoicesonline.org/
[2] http://www.therosettafoundation.org/

accessibility barrier, amend it accordingly in the new translated page and inform the client or person responsible for the site.

Between December 2014 and January 2015, we conducted what was, to the best of our knowledge, the first web accessibility study with web translation professionals [29]. The objective was two-fold: (i) understanding the extent to which translators take into account web accessibility considerations during the web localization process, with a particular focus on image accessibility; and (ii) assessing the impact of using two different evaluation tools on the achievement of appropriate image text alternatives in the localized web product. In the preliminary findings presented in [29], we claimed that the use of such tools contributed to increase the visibility of alt texts as translatable elements. However, no indications were given as regards their appropriateness and how decisive the use of checking software had been in supporting translators to generate their final text alternatives proposals. This paper reports the main results obtained from an evaluation carried out with screen reader users to precisely cover that unexplored aspect. Its main contributions are the following:

- By extending prior work, this paper provides empirical evidence of the relevance of including accessibility testing in the web translation chain.
- Besides, a combination of a controlled language checker and a web accessibility conformance evaluation tool is put forward as an optimal solution to improve text alternatives' appropriateness and thus enhance image accessibility.

2. TRANSLATING THE WEB

In this section, we discuss current technology-mediated web translation workflows and we offer a glimpse of why accessibility evaluation should be integrated therein.

2.1 Computer-Assisted Translation

Since the late 1990s, translation practices have moved hand in hand with technological advancements. The advent and establishment of Computer-Assisted Translation (CAT) tools have enabled translation professionals to directly receive source HTML files −instead of decontextualized translatable strings in plain text format−, process them with this software, perform the translation task, and return the automatically-generated target language HTML files, sometimes without even touching the code. CAT tools isolate translatable content from the document markup, rendering it non editable [37]. In addition, they include built-in multilingual databases −known as translation memories (TMs)− which allow translators to reuse previously translated content when matches are found in the text being processed.

In the recent years, there has been an emerging demand of translators trained on web technologies, multimedia processing and desktop publishing techniques, driven by a continuously evolving Web, characterized by a complex interplay of (hyper)text, multimedia content and interactive elements that also need to be localized. When, along with the translation job, a localization engineering task is also requested, translators need to rely on other tools which are not necessarily integrated within their translation environment, such as web and other advanced text editors. Once the website (or pages) has been adapted, all files should be verified and tested [37].

2.2 Web Translation Quality Assurance

During the localization process, changes made might alter the website's layout or functionality, sometimes leading to encoding problems, broken links or truncated strings due to the new text length, which must be corrected [37]. In addition, translators need to make sure that all the source text content has been adapted to the target language. The latter can be done manually by loading the website in different browsers and visually inspecting each page translated, or automatically, through the use of quality assurance (QA) tools.

Current CAT tools feature language-related QA functionalities, such as looking for untranslated segments, as well as punctuation, formatting and terminology and inconsistencies. Stand-alone QA tools also offer the possibility, among others, of defining regular expressions to search for pattern matches or common typing mistakes, such as duplicated words or spacing errors [10]. Potential mistranslations can also be automatically flagged by comparing source and target sentences' length (*ibid*). Complementing CAT tools with linguistic intelligent authoring programs, through which controlled languages (CL) can be implemented, has been proposed as an alternative way to reinforce translation quality [35]. A CL is an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar and style [21]. Applying purpose-driven CL rules can partially relieve content authors and translators of going through time-consuming and hard-to-understand style guides and easily ensure, for example, that client-specific writing conventions are met.

However, all the above does not fully guarantee a quality web translation job. Previous work has suggested that two of the main obstacles experienced by screen reader users when particularly browsing the multilingual web are the presence of inaccessible language selectors and untranslated content [30]. We argue that this last flaw might be derived from the inability of CAT tools to extract all translatable text, or simply to the fact that not all textual content is visible on the screen and translators neglect its existence. Adding an accessibility testing phase to the translation workflow could help to solve such problems, as well as to increase awareness about screen reader users' needs and browsing behavior within the localization community. Since expertise in web accessibility matters has not been traditionally observed as a requirement for localization professionals, automated solutions could serve to bridge this gap.

3. WEB ACCESSIBILITY TESTING

The subsections below briefly review existing guidance and automated aids to perform web accessibility assessments, as well as to concretely tackle image accessibility-related barriers.

3.1 Automated Web Accessibility Evaluation

Automated web accessibility evaluation (WAE) solutions are quality assurance tools specifically designed to evaluate web content according to accessibility requirements, such as the Web Content Accessibility Guidelines (WCAG) 2.0 [40]. They facilitate the accessibility assessment of a high volume of websites in real-time and help to reduce the costs and human resources needed for other non-automated methods [17]. Nevertheless, the literature provides insight into some of their main disadvantages. Findings from a comparative analysis of six evaluation tools carried out by Vigo *et al.* [42] indicated that, while showing a high level of correctness, all tools covered less than 50% of the WCAG success criteria (SC) and caught less than 40% of

violations. If not complemented with user testing, as suggested in [45], the former may lead to wrong web accessibility assumptions. Another major drawback is the lack of comprehensive information about the reasoning behind the problems flagged throughout the automatic checking process [26]. A recurrent example found in the literature are the warnings concerning image accessibility, where the need of a text alternative is highlighted, but not enough support is usually offered to correct the problem detected.

3.2 Solutions for Rendering Images Accessible

Image accessibility strongly depends on (i) the presence of a text alternative to represent the meaning or purpose conveyed through the image, and (ii) the pertinence of that description itself. The most widely-used mechanism to introduce a text alternative in a web page is the use of an `alt` attribute within the `<img>` HTML element.[3]

3.2.1 Authoring Guidance

Scholars and official bodies have both proposed best practices for image description to support human verification of text alternatives. They often present general recommendations on how to insert the `alt` attribute in the HTML code and what value to provide it with depending on the nature of the image (informative vs. decorative), its context within the web page, and its ultimate purpose [11,20]. Petrie *et al.* [25] supported this work by adding up user input and suggested that more exhaustive guidelines needed to be developed. However, detailed guidance on linguistic appropriateness has been generally limited to subjective techniques; for instance, caring about spelling and grammar; using normal prose; and making text simple, succinct and accurate [9,44]. The Technical Specification ISO/IEC TS 20071-11:2012 [16], based on Tang's work and guidance tool *TATI*[4] [39], offers a more detailed question-guided procedure for providing informative text alternatives, but language-oriented hints are barely referred to. Moreover, dealing with such specialized documentation can be tiresome and sometimes even overlooked by web professionals due to lack of time or knowledge, as it has been claimed in previous studies [14,20].

3.2.2 Evaluation and Repair

Prior work on machine checking of image text equivalents have focused on image OCR or text pattern recognition techniques (e.g., dictionary-based word search, file type abbreviations, HTML code, number of characters) to automatically identify what should not be present in appropriate text alternatives [4,15,24]. The size of the image has also been used as a reference to detect non-accessible images [5], classifying as informative (and thus in need of `alt`) those bigger than 10 x 10 pixels, and then automatically giving them text alternatives based on web content analysis, OCR and human labelling (*ibid*). Other novel solutions include a game-based crowdsourcing method for image description [1,2]. However, this model relies on users labeling images with just isolated words, like the system proposed by Keysers *et al.* [18], which might not be regarded as sufficient by end users who expect more elaborated text alternatives. Vinyals *et al.* [43] solve this potential inconvenient with the automatic generation of natural sentences based on a neural and probabilistic network system. While extremely promising, to the best of our knowledge, it only produces text in plain English, thus leaving the multilingual web unattended.

The use of CL rules could prove advantageous in that regard. Rodríguez Vázquez and Lehmann [34] presented *Acrolinx*, a controlled-language based authoring tool, as an automated solution for text alternatives' quality checking. Following an error description formalism, a set of 40 style rules was developed. These were founded on a research-based analysis of French linguistic patterns frequently used in appropriate and non-appropriate text alternatives for images, representing descriptive, functional and uninformative content [34]. After each *Acrolinx* check, the user receives error repair guidance in the form of detailed information about the style rules that have been contravened, together with improvement suggestions, when available (*ibid*), as recommended by the W3C [40]. The advantage of such a customizable system is that more languages could be covered and other rule packages could be added to test different language-based web content.

4. WEB TRANSLATION EXPERIMENT

The aforementioned *Acrolinx* technology, customized for image accessibility evaluation purposes, was one of the tools proposed for the web accessibility study with 28 professional translators introduced in the first section of this paper. Participants were asked to (i) translate a website about a fake development campaign named 'Together Against Poverty' from English into French, including three web pages, and (ii) check them for image accessibility (the website contained 130 images). Detailed information about the participants' profile, the recruitment procedure and the experiment material can be found in [29]. For this factorial study, we chose a split-plot design to measure two independent variables: (i) web accessibility (WA) knowledge, and (ii) use of tools. Participants were divided in two different groups in order to manipulate the first independent variable. Investigating this factor goes beyond the scope of this paper, so it will not be considered during the interpretation of the data analyzed (sections 6 and 7).

The second independent variable had three levels, where the control condition was the translation of the website without the help of any evaluation tool. Participants were requested to submit their translation work (hereinafter referred to as translation version T1) to the researcher upon completion of the task. In the experimental condition, participants had to use two different checking software: *aDesigner,*[5] a general web accessibility evaluation tool, and *Acrolinx*.

The reasons that motivated the selection of *aDesigner* are the following: on one hand, it is a desktop application, like *Acrolinx'*s client for web pages verification, whose user interface is a priori intuitive and relatively simple to use; in addition, the results reporting format is very similar to *Acrolinx*'s, since errors detected are organized per rule violated and a description of the problem is also provided; on the other hand, as far image accessibility checking is concerned, *aDesigner* provides some clues about the alt text appropriateness of the images found [3], instead of just detecting if an `alt` attribute is present or not, which is a popular feature among other more up-to-date tools such as WAVE[6] or

[3] We are aware that the HTML5 `figure` and `figcaption` elements can now be used to associate a longer text alternative with an image, leaving the alt attribute just to label it. Nonetheless, these HTML elements are not yet accessibility supported by the majority of browsers [10].

[4] http://userlab.usask.ca/TATI/Instructions.php

[5] http://www.eclipse.org/actf/downloads/tools/aDesigner/

[6] http://wave.webaim.org/

FAE.[7] These tools use pattern recognition techniques to detect file names or alt texts longer than 100 characters, but do not flag uninformative text alterantives. *aDesigner,* in turn, would identify, for instance, 'banner', 'line', 'spacer' or 'image' as words within the text alternative and indicate that they probably lead to inaccessible images. *Acrolinx* CL rules also cover this assumption.

The order in which both tools were used during the experiment was counterbalanced, with a view to reduce bias due to random or confounding variables. As a result, 14 out of 28 participants (7 per group) handed in the second translation version (T2) once they had performed a check with *aDesigner*, and the other 14 followed the same procedure after using *Acrolinx*. This logic was also applied to collect translation version T3 (see Table 1).

Table 1. Experimental design of the web translation study

		Use of tools		
		Control	***Experimental***	
	Participants	**Translation Version**		
	Group	*T1*	*T2*	*T3*
WA knowledge	Group 1, *N=14* (with)	No tool *N=28*	aDesigner	Acrolinx
			Acrolinx	aDesigner
	Group 2, *N=14* (without)		aDesigner	Acrolinx
			Acrolinx	aDesigner

It should be noted that translation proposals were cumulative, that is, version T1 served as a starting point to produce T2. Similarly, the website translation proposal T2, checked with the first tool used and presumably amended according to the testing results it yielded, was the basis to generate the final version (T3). The rationale behind this decision was based on the belief that the use of several evaluation tools is not mutually exclusive but rather the opposite: combining different software's capabilities has been already put forward as a possible solution for tools' low effectiveness [42] and could thus lead to better accessibility results.

5. IMAGE ACCESSIBILITY EVALUATION

Once the translation experiment concluded, an evaluation study was conducted with the broader goal of gathering knowledge concerning the image accessibility level achieved by all participants in each translation version. To this end, a selection of image text alternatives were subjected to an assessment by an external panel of judges.

Through the analysis of the evidence obtained, we aim at testing the following hypotheses:

- **H.1** Using automated accessibility evaluation tools during the web translation process has a positive impact on the appropriateness of translated text alternatives for images.
- **H.2** When only one evaluation tool is used, a controlled-language tool with style-oriented rules for image text alternatives' checking like *Acrolinx* helps translators to achieve more appropriate text alternatives than a general web accessibility evaluation tool like *aDesigner*.
- **H.3** When two evaluation tools are used, a controlled-language tool with style-oriented rules for image text alternatives' checking like *Acrolinx* leads to more improvements than a general web accessibility evaluation tool like *aDesigner,* irrespective of the order in which they are used.
- **H.4** Using two different tools triggers more improvements in terms text alternatives' appropriateness than using only one.

The research hypotheses set forth are grounded on the preliminary conclusions withdrawn from an earlier exploratory study, whose results suggested that the application of CL rules specifically formalized for text alternatives verification could contribute to improve their adequateness [33]. Nevertheless, at that stage, *Acrolinx* rules had been applied by the author and no rule documentation was available. The present evaluation study seeks to support that claim with more solid empirical evidence. The subsections below summarize the methodology adopted to explore the quality of the text alternatives produced by the translators in each translation version submitted.

5.1 Materials and Task Design

The primary data retrieved from the experiment described in section 3 consisted of 84 versions of the same website (3 per translator), accounting for 252 web pages, with a total of 10,920 images. This numbers had to be reduced for a manual evaluation to be feasible. Randomly selecting a subset of web pages was not an option, since we wanted to examine the improvements, if any, made by individual translators throughout the different translation versions. Basing the sampling process on a selection of participants had to be dismissed if significant results were to be obtained. Another possibility was to carry out a manual evaluation according to image types; for instance, the impact of using several tools on the alt text quality could have been measured taking only into account functional images. However, Acrolinx CL rules had been developed per image purpose (decorative, informative or functional), and focusing just on one image type would have not allowed us to study, in future research work, the effectiveness of each rule developed, as well as the preferred linguistic patterns for describing each of these types of images. Therefore, we finally collected all `alt` values produced during the controlled experiment (28 translators × 130 images × 3 translation versions, N=10,920). A reduced sample was obtained after filtering all duplicates, reaching a total of 2,189 unique text alternatives.

The data generation method chose for the evaluation was a questionnaire, which was implemented using SurveyMonkey. In an attempt to increase the ecological validity of the study, evaluators received detailed information about the website from which all images had been extracted: what was the campaign about and who were its initiators. We deliberately decided not to send them the source English website for reference to avoid any potential bias. It is worth mentioning that most of the images were not originally accessible, that is, they either did not have an associated `alt` attribute or contained an inappropriate `alt` value. We believe that, if given access to this website, evaluators would have been tempted to assess the text alternatives based on their translation accuracy with respect to the source, and not on their adequateness in terms of accessibility. The structure of the website was also described, indicating that each page corresponded to one of the campaign partners. Similarly, evaluators were provided with exhaustive information about the macrostructure (header, body, footer) of each web page. These comprehensive explanations aimed at helping users to better picture the website from where the images had been retrieved.

[7] http://fae20.cita.illinois.edu/

Each image was presented to the evaluators in a separate page of SurveyMonkey for ease navigation purposes. At the top of each page, the evaluator would find the following data: (i) the website's page where a given image appeared, (ii) its relative location within that page (as per the macrostructure indicated above) and (iii) a neutral description of the image's context. Immediately after, blind users could read how many text alternatives they would need to assess, as well as the alt text list (see Figure 1).

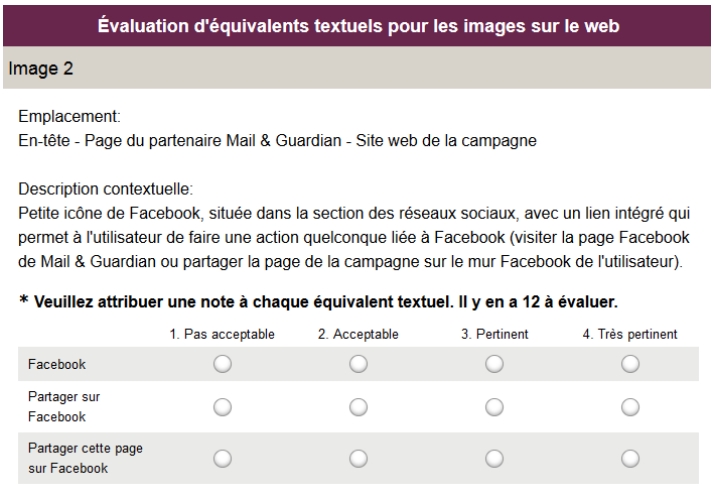

Figure 1. Text alternatives evaluation environment.

5.1.1 Evaluation Metric

As stated at the beginning of section 5.1, the study reported in this paper was designed to serve as a double evaluation exercise: on one hand, we expected data gathered to allow us to estimate if the use of tools helps translators to achieve quality text alternatives and thus a high level of image accessibility. On the other hand, we aimed at measuring the impact of the CL rules developed for *Acrolinx* [31]. Although this paper specifically addresses the former, an evaluation metric convenient for both purposes had to be chosen.

Accessibility evaluation. The literature is populated with multiple studies devoted to develop, test and review automatic and human web accessibility metrics, recently classified in two main groups: conformance-based metrics and accessibility-in-use metrics [41]. While the former are based on whether success criteria (SC) of given guidelines are met, the latter are founded on the premise that accessibility is a quality that differs from conformance (*ibid*). If we look at the WCAG 2.0 [8], image accessibility concerns guideline 1.1 and its first associated SC (1.1.1). Since we wanted to assess only one common web accessibility failure, selecting a conformance-based metric was not deemed appropriate. Furthermore, text alternatives quality goes beyond merely inspecting the code for accessibility conformance, rendering unsuitable binary scoring scales, such as the one used in the failure-rate metric [38]. Alt texts' length or complexity has always been a source of discrepancies. Previous work has highlighted that these are subjective parameters and that it is difficult to establish a clear baseline [25], so they were not consider relevant to determine text alternatives' appropriateness either.

Controlled language evaluation. The main advantage of CLs is that they make many aspects of text manipulation easier for both humans and computer programs [21]. Therefore, the metrics found in the literature to evaluate them strongly depend on the purpose for which the CL had been created. Traditionally, the most common applications of CLs have been the improvement of text readability, comprehensibility and machine translatability. A review of the related work reveals that the effect of CL rules on the latter has been measured through both automatic and human metrics [36]. Since the former mostly rely on reference translations to compute machine translation quality, we decided to consider the latter. Although human judgements are often regarded as subjective, the reliability of the results can be maximized by objectively defining the criteria that will be used by the selected group of evaluators *(ibid)*.

For the purposes of the current work, a tailor-made metric based on the appropriateness level achieved in the text alternative was designed to measure its quality (see Table 2). Notice that only one negative value is provided (1), thus just leaving room for a grading scale in the case of positive annotations. The final Likert-type scale includes four rating levels. It differs from other translation quality annotation metrics, such as the one presented in [23], in that it does not focus on the comprehensibility of the text or its linguistic richness. Instead, quality is based on the level of appropriateness reached with respect to the image context described. This approach is similar to Fischer's rating system [12], but applied to only one WCAG SC.

Table 2. Grading scale on four levels to assess alt text quality

	Score	Criteria
1	Not appropriate (*Pas acceptable*)	The text alternative is not appropriate for the image, according to the location and context described.
2	Acceptable (*Acceptable*)	The text alternative is acceptable for the image, according to the location and context described, but not all the information provided is necessarily pertinent.
3	Pertinent (*Pertinent*)	The text alternative provides minimal but sufficient and correct information about the image, according to the location described.
4	Very pertinent (*Très pertinent*)	The text alternative provides complete and precise information about the image, according to the location and context described.

5.2 Participants

A snowball sampling method was chosen to recruit screen-reader users willing to take part in the alt text assessment. Requirements were (i) having a full proficiency in French and good knowledge of English, and (ii) being experienced users of assistive technology and the Web. The first requirement made the recruitment process particularly challenging. Applying crowdsourcing techniques would have boosted the participation rate, but the effort made by users would have been similar. The reason is that, in order to get comparable results, each participant would have needed to annotate, at least, all alt texts corresponding to one of the image types. Still, following this procedure, alt texts produced by the same translator would have been annotated by different evaluators, thus not allowing us to correctly assess their overall performance.

From the 9 people who initially signed up, only 7 completed the task. One blind user indicated that VoiceOver for Mac OS was his preferred screen-reader. The rest were regular JAWS users, occasionally choosing NVDA when the former was not available. There were participants from three different French speaking countries: Switzerland, Canada and France. English was the most common second language spoken by all evaluators.

Within this group, we observed two main profiles:

- *Linguists*: Three participants (all female, aged between 37 and 40, $\bar{x} = 35.6$, $sd = 5.13$) had a translation background. They were all French native speakers. All acknowledged to use braille displays for a better work performance. Two of them self-reported to have some basic knowledge on web accessibility.
- *Web specialists*: Four participants (all male, aged between 29 and 41, $\bar{x} = 32.4$, $sd = 6.65$) reported to have a rather technical background, with three of them working as web accessibility consultants. The fourth evaluator was an experienced web developer who had some knowledge on the mater but never worked in accessibility-related projects. All had French as their mother tongue but two, who were German native speakers living on a long-term basis in a French speaking country.

5.3 Procedure

Upon acceptance of the task, evaluators were sent by e-mail a MS Word file with all the instructions needed to perform the evaluation in French. The document included the following contents: an introduction to the study; a detailed description of the website's purpose and each page's macrostructure (see section 5.1), information about how the questionnaire was organized and an explanation of the score values in our rating scale. We also informed them that the image text alternatives they were about to assess were extracted from a website translated by multiple translators, hence the presence of alt text not only in French, but also in English. Still, they were requested to assign scores on the premise that, in a real life situation, they would be browsing a French website.

We had estimated that the assessment exercise could take between 5 and 12 hours, so we also provided evaluators with tips about how to enable the cookies before starting the SurveyMonkey questionnaire. This would allow them to take breaks during the task or work on it throughout several days without losing their responses to the questions already completed. The link to the questionnaire was both included in the e-mail and at the end of the instructions file. Each screen-reader user received a monetary compensation of CHF 100 as an acknowledgement for the work done.

6. RESULTS

Time spent by evaluators (hereinafter also referred as "judges") on the task was consistent with our initial estimation ($\bar{x} = 9.28$, $sd = 4.75$). Upon data collection, a within-subjects analysis was performed in order to study if there was a significant effect of the use of tools (*independent variable*) on the quality level of the text alternatives produced by the translators in each translation version. Discussing which linguistic patterns are recommended to write appropriate text alternatives depending on the image meaning and purpose, as inferred from the scores assigned by the judges to each alt text, goes beyond the scope of the present paper and will not thus be addressed.

Before the statistical analysis, it was crucial to take into account two important characteristics of the data gathered:

- The score per alt text (*dependent variable*) was based on a Likert-like scale (ordinal data) so, in principle, its distribution cannot be assumed normal.
- There is a correlation between the observations made (repeated measures per *judge*, *image type* and *translator*).

After considering the above, we decided to use a repeated measure one-way analysis of variance (ANOVA) on the 4-level score scale, presented in sub-section 5.1.1, with 3 random factors (*judge*, *image type* and *translator*), in order to test the hypotheses stated in section 5. The approach adopted assumes that the tests will be robust, despite the non-normal distribution of the data, due to the vast amount of observations (130 images × 28 translators × 3 translation versions × 7 judges, N=76,440). This model was complemented with a post-hoc analysis, for which we applied a Tukey's HSD (honest significant difference) correction. The overall analysis was performed using the R statistical software (library lme4 and multcomp) [27].

The analysis provides strong evidence that there is an effect of using tools on the scores collected ($\chi^2 = 1764$, df=3, $p < 0.001$)[8], irrespective of the number or type of tools chosen. The multiple comparisons of means (see Table 3) of the scores gathered per each condition of the independent variable "use of tools" (*none, aDesigner, Acrolinx, both*) show that the difference between the scores obtained for text alternatives produced in translation version T1 (without any tool) and those of translation version T3 (both tools) is highly significant ($p < 0.001$). Additionally, we found a significant difference in the scores of alt texts amended after using one tool (*aDesigner* or *Acrolinx*, translation version T2) compared to those obtained for alt texts in T1 ($p < 0.001$). Finally, there is a significant difference between scores assigned to text alternatives checked with *aDesigner* and those collected for *Acrolinx* ($p < 0.001$), when only one tool was used (T2).

Table 3. Tukey's test results for independent variable conditions *none, aDesigner, Acrolinx, both*

Comparison of Independent Variable Conditions	Estimate	Std. Error	*p*-value
both tools (T3) > no tool (T1)	0.2977	0.0073	0.001
Acrolinx (T2) > no tool (T1)	0.2235	0.0096	0.001
aDesigner (T2) > no tool (T1)	0.0830	0.0096	0.001
Acrolinx (T2) > *aDesigner* (T2)	0.1405	0.0126	0.001

In order to check the robustness of our results to the non-normality of the scores, a non-parametric Kruskal-Wallis test (K-W) was also performed. The inconvenience of this second approach is that it does not take the correlation structure of the data into account. The test confirms the global significant effect of the use of tools on the judge's scores ($p < 0.001$). The non-parametric post-hoc analysis (Nemenyi test) results also support the strongly significant differences found in scores between the following pairs: Both tools (*T3*) > no tool (*T1*), *Acrolinx* (*T2*) > no tool (*T1*), and *Acrolinx* (*T2*) > *aDesigner* (*T2*); $p < 0.001$. However, a weaker significance was observed in the pairwise comparison of the scores for alt texts produced without any automated solution (T1) with the scores obtained for alt texts verified with *aDesigner* (T2) ($p = 0.041$). Overall, the results from the repeated measure one-way ANOVA are confirmed.

As previously mentioned in the paper, translation versions were cumulative. Therefore, to test hypotheses **H.3** and **H.4**, the order

[8] For the purposes of this paper, we have used an alpha level of .05 for all statistical tests.

in which both tools were used during the translation experiment was also considered in a second analysis. The translation versions T3 (for which input from both tools was presumably taken into account by translators) were thus coded according to the last tool used, namely: (i) "both-last-aDesigner" when T2 was checked with *Acrolinx*, and (ii) "both-last-Acrolinx" when T2 was checked with *aDesigner*. The difference between scores for alt texts verified with the help of two tools (T3) and those run through only one tool (T2) is very significant (see Table 4), regardless of the tool order. Furthermore, a significant difference was observed when comparing scores within the last experimental condition (T3), with regard to the tool used for the last automatic check.

Table 4. Tukey's test results for independent variable conditions *aDesigner, Acrolinx, both-last-aDesigner, both-last-Acrolinx*

Comparison of Independent Variable Conditions	Estimate	Std. Error	*p*-value
both-last-aDesigner (T3) > *Acrolinx* (T2)	0.1159	0.0103	0.001
both-last-Acrolinx (T3) > *aDesigner* (T2)	0.1728	0.0103	0.001
both-last-aDesigner (T3) > both-last-Acrolinx (T3)	0.1664	0.0145	0.001

7. DISCUSSION

The results presented in the previous section provide enough evidence to accept **H.1** *Using automated accessibility evaluation tools during the web translation process has a positive impact on the appropriateness of translated text alternatives for images*. The judges' scores show that alt texts' quality significantly improves when both tools, *Acrolinx* and *aDesigner*, are introduced in the translation workflow to support translators in their effort to render images accessible. We argue that this gain in text alternatives' quality is due to the proper changes made to inappropriate alt texts produced during the first translation version which, in turn, were probably the result of a word-by-word translation of the original English text. This is reflected on the data through (i) the higher proportion of positive scores (2-3-4) obtained in T3 (see Figure 2a, bars *Both* vs. *none*), and (ii) the increased number of pertinent and very pertinent text alternatives achieved (see Table 5, score values 3 and 4).

Although the tests performed to study the analysis of variance point to a significant improvement in the alt texts quality when translators use just *aDesigner*, it can be readily seen from the stacked charts that the use of *Acrolinx* leads to a considerable higher decrease in the number of non-appropriate alt texts, thus supporting **H.2** *When only one evaluation tool is used, a controlled-language tool with style-oriented rules for image text alternatives' checking like Acrolinx helps translators to achieve more appropriate text alternatives than a general web accessibility evaluation tool like aDesigner.* In addition, *Acrolinx* contributes overall to produce more pertinent alt texts for images than *aDesigner* (4,846 and 2,889 respectively, if we add up scores for values 3 and 4; see Table 5). We believe the latter might be motivated by the detailed explanations of the errors flagged by the tool and the improvement suggestions offered [34]. Nevertheless, a closer examination of the screen recorded translation sessions would be needed to validate this statement. Data gathered during the evaluation also suggests that using both tools could yield similar good quality results in terms of image accessibility to those obtained when just using *Acrolinx* (see Figure 2a). We hypothesize that this is due to the specificity of the tool as regards image accessibility checking, since it provides detailed support for the text alternative repair task, featuring language-based recommendations. For instance, *Acrolinx* would (i) flag "Facebook" as an inappropriate alt text for a Facebook icon with an embedded link that allows the user to share a given web page on his Facebook wall, and (ii) suggest to replace it with "Share this page on Facebook". We argue that alt appropriateness issues are currently more popular in the Web than just the existence of `<img>` elements with no `alt` attribute which, according to the literature, is a problem increasingly solved through the introduction of support for accessibility in web authoring tools [28]. This was also the case in our test website.

Table 5. Perceived alt text quality by all judges. Total scores are shown per different levels of the factor *Use of tool*

	Alt text appropriateness score			
Use of tool	**1. Not appropriate**	**2. Acceptable**	**3. Pertinent**	**4. Very pertinent**
none	14006	5298	4194	1982
aDesigner	7155	2696	1976	913
Acrolinx	5252	2642	3028	1818
Both	10841	5377	5952	3310
Both-last-Acrolinx	6290	2611	2539	1300
Both-last-aDesigner	4551	2766	3413	2010

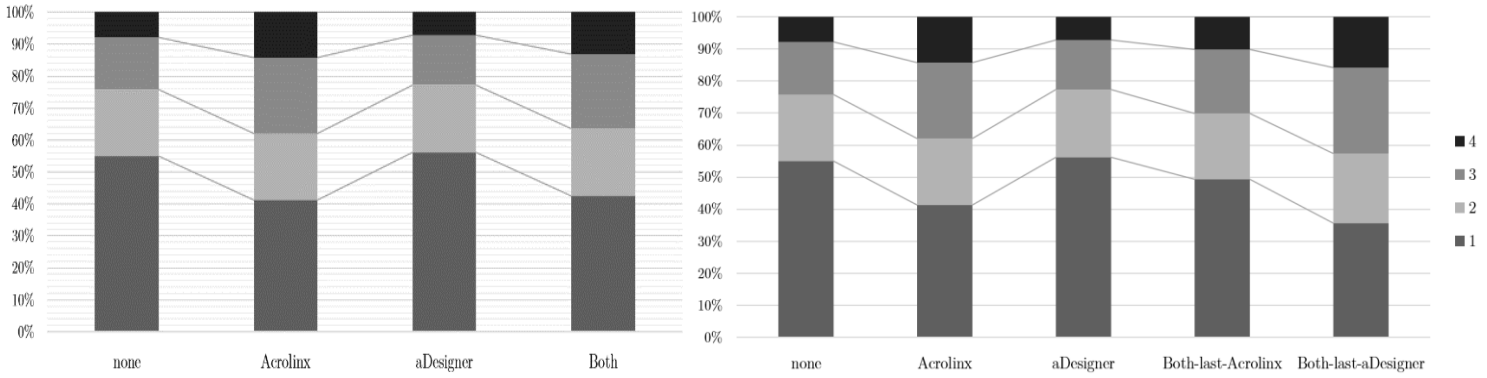

Figure 2. (*a, stacked chart, to the left*); Score proportions per independent variable conditions *none, aDesigner, Acrolinx, both;* and (*b, cumulative stacked chart, to the right*) Score proportions per all independent variable conditions, where translation version T3 scores are divided according to the last tool used (*both-last-Acrolinx, both-last-aDesigner*).

While *Acrolinx* has proved to be more useful than *aDesigner* when presented to the translators in the first place, results seem to indicate that, when used as a second evaluation tool, Acrolinx is not as efficient (see Figure 2b and Table 5). If the order followed is *Acrolinx-aDesigner*, improvements achieved regarding alt text quality are greater than if the reverse order is applied. Hence, we reject **H.3**. We hypothesize that the second tool used has less influence on the translators' work, even if the difference in the improvements made is significant (see Table 4). One of the reasons might be the lack of time: translators had 90 minutes to use both tools and they probably devoted more than half of the time to solve the barriers identified through the first tool. By tackling language-related quality issues ahead with the help of *Acrolinx*, they obtained a higher number of improvements, which were quickly complemented at the end of the session with *aDesigner* suggestions. Since they mostly concerned the inexistence of `alt` attributes, they were easy to implement. This might explain why the final image accessibility level achieved in the last translation version was higher. Conversely, after having already used a tool like *aDesigner*, translators might have found *Acrolinx*'s recommendations too time-consuming and more difficult to apply, thus ignoring some of them. Still, as shown in Table 4, using two different tools triggers more improvements in terms text alternatives' appropriateness than using only one, irrespective of the order in which tool checks were run. This goes in line with the conclusions withdrawn by Vigo *et al*. in previous work [42]. Consequently, we can accept our last hypothesis (**H.4**).

Our methodology had two major shortcomings. On one hand, image contextual descriptions were given to evaluators instead of providing them with fully functional independent websites to replicate a real case browsing scenario. This might have influenced, for instance, how English untranslated content was annotated by the judges, since they were often presented as well with French alternative solutions proposed by other translators. Still, as explained in section 5.1, this approach would have required a higher investment in time and resources and would not have enabled us to collect enough data to perform further analysis on individual translators' performance. In spite of the limited number of judges, we consider that the reliability of the results was not compromised. The intra-class correlation (ICC) coefficient is estimated at 0.33 − 0.39 if participant B7 is not taking into account (see Figure 3). These results reflect a fair agreement between judges. We believe that a higher inter-rater agreement could have been reached if the task would have been shorter or the judges' profile was closer. Still, according to the data analyzed, the outliers seem not induce errors or have too much influence on the conclusions of the tests, which happen to be uniform along the various statistical methods we have applied.

On the other hand, time constraints might have also had an effect on translators' performance. By looking at the results, we estimate that, should translators' have received a similar job in a professional context, they would have invested more efforts on complying with the image accessibility requirements pointed at by the tools. This assumption is supported by the comments left by some translators in the post-task questionnaire, where they acknowledged to have only focused on one of the three web pages of the website or just one type of images. In this sense, it is worth mentioning that the analysis presented in this paper is based on the data collected from all 28 translators. However, five of them did not make any modifications to their first translation proposal after using the two evaluation tools. Moreover, another three translators only applied changes according to one of the tools used. A more in-depth investigation of these cases might help to provide further insight into our research findings.

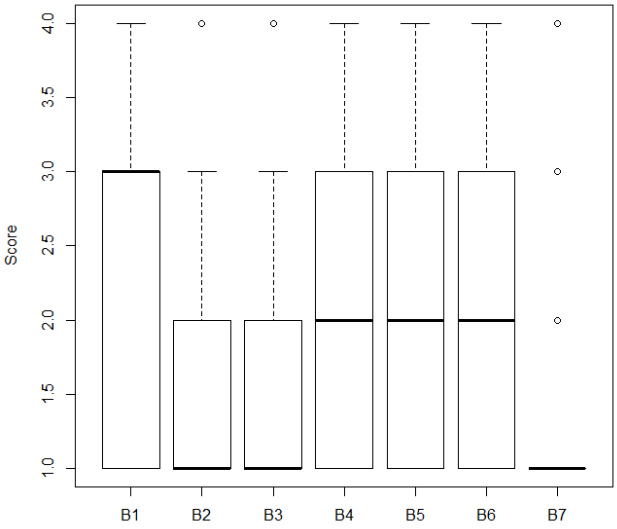

Figure 3. (*boxplot*) Distribution of scores per judge.

8. CONCLUSIONS AND FUTURE WORK

The evaluation study described above has shown that, if equipped with the appropriate tools, web translation professionals can significantly contribute to image accessibility in localized websites, irrespective of their previous knowledge on the subject. In the near future, we plan to examine if the latter had an effect in the overall quality of the text alternatives produced by translators from the experimental group, who had followed a webinar on web accessibility prior to the controlled experiment. If confirmed, empirical evidence would be available to support the belief that translators and web localization engineers should be educated in web accessibility if the broader goal of a universal web is to be met. The global outcome of this research raises again the question of accountability with regard to accessibility in the multilingual web, already explored in prior work [32]. As highlighted in section 5.1, the source English website had a poor image accessibility level. Introducing accessibility testing in the web localization workflow could not only benefit users with special needs in the target audience for whom we are translating the website, but also those from the source culture, language or country, simply by instructing web translators to report accessibility barriers found during the localization process.

Results obtained illustrate that the *Acrolinx* technology, based on accessibility-oriented CL rules, offers more relevant guidance to translators on how to improve the appropriateness of the text alternatives for images they have translated than *aDesigner*. The more comprehensive analysis of the impact of each CL rule developed presented in [31] confirms this conclusion. We thus believe that, if directly integrated in popular CAT tools −such as Alchemy Catalyst or SDL translation products, which already support *Acrolinx* plug-ins−, the alt text checking rule set could facilitate a smooth image accessibility testing within one single working environment. This approach would allow to bring accessibility concerns closer to professional translators who have not been trained on the matter. What is more, as discussed in previous sections, this image accessibility evaluation methodology could prove efficient for translators if complemented with a general web accessibility evaluation tool like *aDesigner* to

achieve optimal results. It would be interesting to conduct a similar experiment with other web professionals (developers, designers, webmasters), in order to see if *Acrolinx* yield comparable results.

We foresee to further investigate if the variability of the scores correlates with the background and accessibility expertise of the judges recruited for the study. Finally, we expect to compare this first evaluation outcome with (i) results from a second human evaluation carried out with sighted experts, and (ii) results collected from the automated test reports saved by each translator for each tool. This would provide added-value to our research in two fronts: first, we would be able to understand if all improvements were directly motivated by the tools' feedback or whether some of them were founded on the own translators' initiative; second, it would allow us to better study the effectiveness of each tool [6] with regard to image accessibility checking by analyzing their correctness (how well they reduce false positives) and their completeness (how well they reduce false negatives).

9. ACKNOWLEDGMENTS

We will always be indebted to the seven screen reader users who donated some of their precious time to our evaluation study. We would also like to express our gratitude to Marc-Olivier Boldi, from the Research Center for Statistics of the University of Geneva, who actively contributed to the data analysis sections, and Véronique Bohn, who proofread all the evaluation materials written in French. Finally, we want to thank Markel Vigo, whose guidance and feedback has been extremely helpful before and during the preparation of this paper.

10. REFERENCES

[1] von Ahn, L. and Dabbish, L. 2004. Labeling Images with a Computer Game. In *Proceedings of CHI 2004*, Vienna, Austria, April 2004.

[2] von Ahn, L., Ginosar, S., Kedia, M., Liu, R. and Blum, M. 2006. Improving Accessibility of the Web with a Computer Game. In *Proceedings of CHI 2006*, Montréal, Québec, Canada, April 2006.

[3] Asakawa, C. 2005. What's the Web Like if You Can'T See It? In *Proceedings of W4A 2005*. Chiba, Japan, May 2005.

[4] Bigham, J., Kaminsky, R., Ladner, R., Danielsson, O. and Hempton, G. 2006. WebInSight: Making Images Accessible. In *Proceedings of ASSETS 2006*, Portland, Oregon, USA, October 2006.

[5] Bigham, J. 2007. Increasing Web Accessibility by Automatically Judging Alternative Text Quality. In *Proceedings of IUI 2007*. Honolulu, USA, January 2007.

[6] Brajnik, G. 2004. Comparing accessibility evaluation tools: a method for tool effectiveness. *Universal Access in the Information Society*. 3 (3-4), 252-263. Springer, Berlin.

[7] Bredenkamp, A., Crysmann, B. and Petrea, M. 2000. Looking for Errors: A Declarative Formalism for Resource-Adaptive Language Checking. In *Proceedings of LREC 2000*, Athens, Greece, May 2000.

[8] Caldwell, B., Cooper, M., Reid, L. and Vanderheiden, G. (eds). 2008. Web Content Accessibility Guidelines 2.0. World Wide Web Consortium (W3C) Recommendation. http://www.w3.org/TR/WCAG20/.

[9] Craven, T. 2006. Some features of alt texts associated with images in Web pages. *Information Research*, 11(2).

[10] Debove, A., Furlan, S. and Depraetere, I. 2011. A Contrastive Analysis of Five Automated QA Tools. In Depraetere, I. (ed). *Perspectives on Translation Quality*, 161–92. Text, Translation, Computational Processing (TTCP) 9. De Gruyter Mouton, Germany.

[11] Faulkner, S. (ed). 2014. HTML5: Techniques for providing useful text alternatives. (W3C) Working Draft. http://www.w3.org/TR/html-alt-techniques/

[12] Fischer, D. and Wyatt, T. 2011. The case for a WCAG-based evaluation scheme with a graded rating scale. In *Proceedings of the W3C WAI Symposium on Website Accessibility Metrics*. Article 7. http://www.w3.org/WAI/RD/2011/metrics/paper7/.

[13] Folaron, D. 2012. Digitalizing translation. *Translation Spaces*. John Benjamins, 1 (2012), 5-31.

[14] Harper, S. and Chen, A. 2012. Web Accessibility Guidelines: A Lesson from the Evolving Web. *World Wide Web*. 15 (1): 61–88.

[15] Hu, J. and Bagga, A. 2003. Identifying Story and Preview Images in News Web Pages. In *Proceedings of ICDAR 2003*, Edinburgh, Scotland, UK, August 2003. IEEE Computer Society Washington, DC, USA.

[16] ISO/IEC TS 20071-11:2012. Guidance for alternative text for images. 2012. Switzerland: International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).

[17] Ivory, M. and Hearst, M. 2001. The State of the Art in Automating Usability Evaluation. *ACM Computing Surveys*, 33(4), December 2001, 470–516.

[18] Keysers, D., Renn, M. and Breuel, T. 2007. Improving Accessibility of HTML Documents by Generating Image-Tags in a Proxy. In *Proceedings of ASSETS 2007*, Tempe, Arizona, USA, October 2007.

[19] Korpela, J. 2010. Guidelines on alt texts in IMG elements. https://www.cs.tut.fi/~jkorpela/html/alt.html.

[20] Lazar, J., Dudley-Sponaugle, A., and Greenidge, K. 2004. Improving Web Accessibility: A Study of Webmaster Perceptions. *The Compass of Human-Computer Interaction* 20 (2): 269–88.

[21] Nyberg, E., Mitamura, T. and Olaf-Huijsen, W. 2003. Controlled Language for Authoring and Translation. In Somers, H. (ed), *Computers and Translation. A Translator's Guide*, 245–81. John Benjamins.

[22] O'Brien, S. 2012. Translation as Human-Computer Interaction. *Translation Spaces*, John Benjamins, 1 (2012), 101-122.

[23] O'Brien, S., and Roturier, J. 2007. How Portable Are Controlled Language Rules? A Comparison of Two Empirical MT Studies. In *Proceedings of MT Summit XI*, 345–52. Copenhagen, Denmark. September, 2007.

[24] Olsen, M., Snaprud, M. and Nietzio, A. 2010. Automatic Checking of Alternative Text on Web Pages. In Miesenberger et al. (eds), *ICCHP 2010, Part I*, 425-432. Springer, Berlin.

[25] Petrie, H., Harrison, C. and Dev, S. 2005. Describing images on the Web: a survey of current practice and prospects for the future. In *Proceedings of HCII 2005*, Las Vegas, Nevada, USA, July 2005.

[26] Petrie, H., Power, C., Swallow, D., Velasco, C.A., Gallagher, B., Magennis, M., Murphy, E., Collin, S. and Down, K. 2011. The value chain for web accessibility: challenges and opportunities. In *Proceedings of ADDW 2011*, Sun SITE Central Europe.

[27] R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[28] Richards, J., Montague, K. and Hanson, V. 2012. Web Accessibility as a Side Effect. In *Proceedings of ASSETS 2012*, Boulder, Colorado, USA, October 2012.

[29] Rodríguez Vázquez, S. 2015.Unlocking the Potential of Web Localizers as Contributors to Image Accessibility: What Do Evaluation Tools Have to Offer? In *Proceedings of W4A 2015*, Florence, Italy, May 2015.

[30] Rodríguez Vázquez, S. 2015. Exploring Current Accessibility Challenges in the Multilingual Web for Visually-Impaired Users. In *The 24th World Wide Web (WWW) Conference 2015 Companion Volume*. ACM.

[31] Rodríguez Vázquez, S. 2015. A controlled language-based evaluation approach to ensure image accessibility during web localisation. *Translation Spaces*. John Benjamins. 4 (2): 187-215.

[32] Rodríguez Vázquez, S. and Bolfing, A. 2013. Multilingual Website Assessment for Accessibility: a Survey on Current Practices. In *Proceedings of ASSETS 2013*, Bellevue, WA, USA, October 2013.

[33] Rodríguez Vázquez, S., Bouillon, P. and Bolfing, A. 2014. Applying Accessibility-Oriented Controlled Language (CL) Rules to Improve Appropriateness of Text Alternatives for Images: An Exploratory Study. In *Proceedings of LREC 2014*, Reykjavik, Iceland, May 2014.

[34] Rodríguez Vázquez, S. and Lehmann, S. 2015. Acrolinx: a Controlled-Language Checker Turned into an Accessibility Evaluation Tool for Image Text Alternatives. In *Proceedings of W4A 2015*, Florence, Italy, May 2015.

[35] Rösener, C. 2010. Computational Linguistics in the Translator's Workflow−Combining Authoring Tools and Translation Memory Systems. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing*, Los Angeles, California, June 2010. ACL, 1-6.

[36] Roturier, J. 2006. An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-Translated Technical Documentation for French and German Users. *PhD Thesis*. Dublin City University, Ireland.

[37] Sandrini, P. 2008. Localization and Translation. *MuTra Journal*, 2 (2008), 167-191.

[38] Sullivan, T., and Matson, R. 2000. Barriers to use: usability and content accessibility on the Web's most popular sites. In *Proceedings of the ACM Conference on Universal Usability (CUU'00).* Arlington, USA, November 2000.

[39] Tang, L. 2012. Producing informative text alternatives for images. *PhD thesis*. University of Saskatchewan, Saskatoon.

[40] Velasco, C. and Abou-Zahra, S. (eds). 2014. Developers' Guide to Features of Web Accessibility Evaluation Tools. W3C First Public Working Draft.

[41] Vigo, M. and Brajnik, G. 2011. Automatic web accessibility metrics: where we are and where we can go. *Interacting with Computers*. Elsevier, 23(2), 127-155.

[42] Vigo, M., Brown, J. and Conway, V. 2013. Benchmarking Web Accessibility Evaluation Tools: Measuring the Harm of Sole Reliance on Automated Tests. In *Proceedings of W4A 2013*, Rio de Janeiro, Brazil, May 2015.

[43] Vinyals, O., Toshev, A., Bengio, S. and Ehran, D. 2014. Show and Tell: A Neural Image Caption Generator. *CoRR*, abs/1411.4555. http://arxiv.org/abs/1411.4555.

[44] Web Accessibility in Mind, WebAIM. 2013. Alternative Text. http://webaim.org/techniques/alttext/

[45] Yesilada, Y., Brajnik, G., Vigo, M. and Harper, S. 2015. Exploring perceptions of web accessibility: a survey approach. *Behaviour & Information Technology.* 34:2, 119-134.