Cite as: Rossetti, A. and O'Brien, S. (2019). Helping the helpers: Evaluating the impact of a controlled language checker on the intralingual and interlingual translation tasks involving volunteer health professionals. In McDonough Dolmaya, J. and Del Mar, M. (eds.) Special Issue of Translation Studies. Social Translation: New Roles, New Actors.

Helping the helpers: Evaluating the impact of a controlled language checker on the intralingual and interlingual translation tasks involving volunteer health professionals

Dr Alessandra Rossetti (corresponding author)

Dublin City University, Ireland

alessandra.rossetti2@mail.dcu.ie

Prof Sharon O'Brien

Dublin City University, Ireland

sharon.obrien@dcu.ie

Cochrane is a non-profit organization which mainly relies on volunteer health professionals for the production, simplification, evaluation, and multilingual dissemination of high-quality health content. The approach that Cochrane volunteers adopt for the simplification (or intralingual translation) of English health content is non-automated and involves the manual checking and implementation of plain language guidelines. This study investigated whether and to what extent the introduction of a controlled language (CL) checker—which would make the simplification approach semi-automated—increased authors' satisfaction and machine translation (MT) quality. Twelve Cochrane authors completed a standardized questionnaire and answered follow-up questions on their level of satisfaction and preferences. Forty-one Cochrane evaluators assessed the quality of the Spanish MT

outputs of simplified texts. Authors showed a preference for the introduction of a CL checker. Differences in MT quality scores were slight.

Keywords: controlled language checker; text simplification; volunteer satisfaction; machine translation quality; health information; non-profit organization

1 Introduction and Related Work

Volunteers play an increasingly important role in the production, editing, and (multilingual) dissemination of health content, particularly online. Websites that are widely consulted for health-related purposes often rely on contributions from volunteers. For instance, Wikipedia, which has become an important source of online health information, is supported by the Wikimedia Foundation—a non-profit organization—and is written, maintained and edited collaboratively by volunteers (Laurent and Vickers 2009; Heilman et al. 2011). Translators without Borders (TwB) is another example of a non-profit organization relying on volunteers to translate critical information (including medical content) during crises. TwB also partnered with Wikipedia for the *100x100 project*, which involves the translation of 100 of Wikipedia's highest ranking medical articles into 100 developing world languages (TwB 2018a). Epistemonikos is an online database that collates and translates high-quality health evidence by relying on volunteer domain experts (Rada, Pérez and Capurro 2013). Gigliotti (2017) discusses the widespread reliance on volunteer translators at non-profit organizations and goes on to analyse the quality of the translations produced at TwB, The Rosetta Foundation, PerMondo, and Translations for Progress. These initiatives can be defined as *cause-driven* since they contribute to the development of non-profit and humanitarian agendas (McDonough Dolmaya 2012).

Often the volunteers involved in the authoring, editing and translation activities are also non-professional writers/translators, i.e. individuals who have not received formal training in the areas of linguistics/translation (Pérez-González and Susam-Saraeva 2012). It is

not uncommon then for volunteers/non-professional writers/translators to be provided with guidelines and instructions, particularly when they are health domain experts, who might find the usage of layperson terms challenging (Zethsen 2009). For example, Simple English Wikipedia contributors are expected to adhere to an editorial policy including recommendations on plain language writing (Den Besten and Dalle 2008). Similarly, TwB volunteers, who can be either professional translators or simply individuals fluent in two languages, are expected to comply with a code of conduct according to which "[t]ranslators shall always thoroughly and faithfully render the source language message, omitting or adding nothing, giving consideration to linguistic variations in both source and target languages, and conserving the tone and spirit of the source language message" (TwB 2018b, no page number).

Technology can support non-professional volunteers. Pym (2011, 6) argues that technology "enables new social locations for translation, particularly in the non-professional sphere." He also points out that volunteer and professional translators might collaborate, for instance, with volunteers post-editing MT outputs that are later revised by professionals (ibid.). A few examples can help illustrate how this kind of collaboration can take place. With regard to simplification or intralingual translation, Leroy, Kauchak and Mouradi (2013) developed a writer support tool to help health practitioners simplify texts for patients and save time. Regarding interlingual translation, Abekawa and Kageura (2007) describe a translation aid system for volunteers translating from English into Japanese, which facilitated the consultation of references, among other tasks. Translation at Epistemonikos often involves the adoption of machine translation (MT) systems and the subsequent collaborative editing of the MT output (Rada, Pérez and Capurro 2013). As of 2013, 99.6% of Epistemonikos articles had been translated with MT into languages other than English (ibid.).

TwB is also increasingly relying on MT technology to speed up the translation process, e.g. into Kurdish languages (TwB 2016, 2019).

In summary, the important role played by non-professional volunteers for the production, editing, and (multilingual) dissemination of health content has been recognized, and writing/translation technology is increasingly supporting their contributions. However, to the best of our knowledge, no studies have investigated the impact that technologies for intralingual translation—such as authoring support tools and controlled language (CL) checkers—can have on the satisfaction of non-professional volunteers, and on the machine translatability of the simplified texts that they produce.

2 The Case of Cochrane

For our study, we focused on Cochrane[i], a non-profit organization addressing the demand for simplified and multilingual medical content online through intralingual and interlingual translation tasks mainly conducted by volunteer health practitioners, who act as content creators, editors, translators, and quality evaluators. This organization was chosen because of its non-automated approach to intralingual translation (which might have benefited from the introduction of technological assistance), and its increasing reliance on MT technology for simplified texts. This section will delve into the intralingual and interlingual translation workflows at Cochrane, while also outlining the goals of our study.

Cochrane provides online high-quality health-related information by producing Systematic Reviews of empirical evidence answering a specific research question—mainly on the impact of treatments and interventions (Chandler et al. 2017). The specialised medical language that characterizes these Systematic Reviews might be difficult for the lay public to comprehend. Therefore, each Cochrane Systematic Review is accompanied by a plain language summary (PLS)[ii], which both summarizes and simplifies its content (The Cochrane Collaboration 2013). Numerous readers would consult only the PLS rather than the entire

Systematic Review (Maguire and Clarke 2014).

To produce PLS, Cochrane mainly relies on volunteer authors with a health background (i.e. health domain experts), who are therefore involved in the workflow of intralingual translation (i.e. simplification of content to render it accessible to lay readers) (Muñoz-Miquel 2012; Kajzer-Wietrzny, Whyatt, and Stachowiak 2016). Different sets of written guidelines are available to the authors of PLS, which deal with both content (e.g. which information in the Systematic Review to include or exclude), and language/style (e.g. sentence length or terminology). These guidelines can be found in a variety of documents, such as the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins and Green 2011), or the *Standards for the Reporting of Plain Language Summaries in New Cochrane Intervention Reviews* (PLEACS) (The Cochrane Collaboration 2013). Our analysis of Cochrane PLS guidelines has shown that they are characterized by vagueness and contradictions, for instance on the length of PLS and on the use of acronyms. Moreover, this simplification approach is non-automated, because there is no automatic checking for adherence to the PLS guidelines—volunteers are asked to manually check and implement guidelines, which can represent a difficult and time-consuming task (Temnikova 2012), particularly for volunteer health practitioners acting as new agents in the field of linguistics or translation.

Against this background, the first goal of this study was to investigate the impact that introducing Acrolinx[iii] into Cochrane's non-automated simplification/intralingual translation workflow would have on the satisfaction of volunteer authors. Acrolinx is a CL checker developed at the German Research Center for Artificial Intelligence. A CL is a set of rules applied to text production, which aims at making the language more readable and translatable, and a checker is the software that checks for adherence to those rules (O'Brien 2010). Acrolinx automatically and consistently flags readability and translatability issues in a

text, while also providing suggestions on how to solve them (Rodríguez Vázquez 2016). We therefore hypothesized that, by availing of this tool, Cochrane volunteer authors would be more satisfied with the overall simplification workflow. We refer to the approach involving Acrolinx as semi-automated because, even though readability and translatability issues are automatically and consistently flagged, the author needs to manually apply the edits.

Cochrane has also been broadening its international audience, thus adopting strategies for the multilingual dissemination of its content (Von Elm et al. 2013). In particular, PLS are being translated into a variety of languages, including Chinese, Portuguese and Russian (Birch et al. 2017). Due to the non-profit nature of Cochrane, the interlingual translation workflow has largely relied on volunteer health practitioners. Only a small number of Cochrane volunteers are professional translators. Moreover, even when translations are conducted by professionals, a final revision of the accuracy of translated content from volunteer health practitioners/domain experts is common practice (Birch et al. 2017). Cochrane has observed the potential impact of publishing translations. In 2016, 66% of all visits to cochrane.org were conducted through browsers set to a language other than English, and Internet users from Spanish- and French-speaking countries represented a large percentage of this audience (ibid.).

To make the translation effort more sustainable and to increase the number of available translations, Cochrane has been considering the integration of technologies such as CL checkers and MT systems into their workflow (Von Elm et al. 2013). Customized MT systems are being developed for Cochrane content (Bojar et al. 2017). However, the impact that simplification with a CL checker might have on the quality of the Spanish MT output of PLS obtained with freely available MT systems has not been investigated. The second goal of this study was therefore to determine whether introducing the Acrolinx CL checker into the workflow of PLS production would increase the machine translatability of Cochrane PLS.

To summarize, our goal was to test the impact of the Acrolinx CL checker on the satisfaction of Cochrane volunteer PLS authors, and on the machine translatability of the PLS. In the following sections, we will describe methodology, participants, experimental materials, experimental design and tasks within our study. Section 7 will present the main results, while Section 8 will outline conclusions and areas for future work.

3 Methodology

Authoring Study

In this study, we defined satisfaction as the "extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations" (ISO 9241-11:2018, 3.1.14). To measure satisfaction, we asked twelve participants recruited from the pool of Cochrane volunteer authors (Section 4) to revise their PLS previously produced with Cochrane guidelines by following Acrolinx automatic suggestions on readability and translatability (Section 6). Subsequently, these authors were asked to complete the System Usability Scale (SUS) (Brooke 1996) focusing on their interaction with both Cochrane PLS guidelines and the Acrolinx CL checker.

The SUS is a Likert scale questionnaire composed of ten statements (Appendix A). We slightly modified the wording of the statements, i.e. the word "system" in the original SUS was respectively replaced with "Acrolinx" and "Cochrane PLS guidance". Slight wording changes have been shown not to alter the reliability of the SUS (Lewis and Sauro 2009). This questionnaire was selected because it is technology-agnostic and has been adopted for various tools and applications, from websites (Sauro and Lewis 2011) to punched voting cards (Byrne, Greene, and Everett 2007), to CL checkers (Miyata et al. 2017). Moreover, several studies have proven the reliability and validity of the SUS (Bangor, Kortum and Miller 2008).

The SUS provides one single score—the higher the score, the higher the users' satisfaction. However, it does not provide diagnostic information (Chaparro et al. 2014). Therefore, authors also answered follow-up questions on their future preferences for a specific approach. In particular, they were asked if, in the future, they would use: (i) both Cochrane PLS guidance and Acrolinx; (ii) Acrolinx only; (iii) Cochrane PLS guidance only; (iv) other types of authoring support; or (v) no support at all. Finally, authors were also asked about the reasons for their preferences.

MT Evaluation Study

The MT evaluation experiment was conducted with forty-one participants recruited from the pool of Spanish-speaking Cochrane volunteer domain experts (Section 4). Each Spanish-speaking Cochrane volunteer was presented with two English PLS (one produced following Cochrane PLS guidelines, henceforth *non-automated PLS*, and the other revised with Acrolinx, henceforth *semi-automated PLS*) and their Spanish MT outputs, produced by Google Translate, and segmented at the sentence level. Although Google Translate is a generic MT engine, it was selected because it is not unrealistic to expect lay people and experts to use freely available MT systems for assimilation purposes (Section 4).

The presentation order of the sentences followed the structure of the text, rather than being randomized (Doherty 2017). Each MT output was evaluated against its source PLS. Evaluators were asked to assign two scores (from 1, the lowest, to 4, the highest) to each machine-translated sentence. One score was based on fluency and the other score on adequacy (Miyata et al. 2017). A 4-point scale was selected to avoid mid-point bias. Evaluators were instructed to provide their intuitive reaction to each sentence, rather than pondering their decision (Linguistic Data Consortium 2002).

Most evaluators did not have a background in linguistics/translation and might have found it difficult to understand what was meant by fluency or adequacy. Therefore, drawing

upon Castilho et al. (2017), for each pair of sentences, participants were presented with the two following questions, dealing with adequacy and fluency, respectively: “How much of the information contained in the English source sentence appears in the Spanish target sentence?”; and “Indicate the extent to which the Spanish target sentence is in grammatically well-formed and fluent Spanish”.

Human evaluation of MT has several drawbacks, including subjectivity and time commitment (Fiederer and O’Brien 2009). However, due to the importance of accuracy in the health field, the MT output of Cochrane texts requires human validation. Therefore, involving human evaluators created a more realistic scenario. Furthermore, fluency and adequacy are very common in MT evaluation studies, less time consuming than, for example, error annotation, and more accessible as tasks for non-linguist evaluators.

4 Participants

Authoring Study

We recruited authors from the pool of Cochrane volunteer contributors with a health background (Section 2). Twelve Cochrane authors who had written a PLS in the past agreed to revise their PLS by using the Acrolinx CL checker (Section 6). They were volunteer contributors to various Cochrane Review Groups, such as Cochrane Vascular Group or Stroke Group. Most authors were native speakers of English (n=7). The remainder reported Dutch (n=2), German (n=1), Portuguese (n=1) and Russian (n=1) as their native languages. All authors were health domain experts and worked either as academics (n=9) or health practitioners (n=3). Authors differed in the number of PLS that they had produced in the past by following Cochrane PLS guidelines, from only one to ten PLS. There was also wide variability in the time elapsed between the production of PLS with the non-automated approach and the use of Acrolinx—from one month to three years. In other words, our

sample of authors was heterogeneous in terms of native language, familiarity with plain language writing, and memory of Cochrane PLS guidelines.

We tried to recruit a larger and more homogeneous sample of Cochrane authors by distributing our call for participations via a variety of channels, and by sending multiple reminders. However, this was not possible—it is important to remember the Cochrane contributors are academics and health professionals with busy schedules. As a result of the variability in authors' background characteristics, our results cannot be generalized (Section 8).

MT Evaluation Study

Forty-one Cochrane volunteers participated in the MT evaluation task. Like authors, evaluators were recruited from the pool of Cochrane volunteer contributors. All evaluators were native speakers of Spanish and working or training to work in the health field. They were health practitioners (n=22), academics (n=8), both academics and health practitioners (n=2), students or trainees (n=8), and pharmaceutical consultants (n=1). Unlike other MT evaluation studies, evaluators were not professional translators, so their scores were unlikely to be influenced by the perception of MT as a threat (Fiederer and O'Brien 2009). Nonetheless, it is worth noting that eight participants also reported some experience or training as translators. For instance, one participant had translated pharmaceutical/medical documents.

Prior to conducting the evaluation tasks, and as part of a pre-task questionnaire (Section 6), participants were asked to complete the Cambridge English test[iv] for a quick assessment of their level of English proficiency (Parra Escartín et al. 2017). This assessment was needed for the purpose of the adequacy evaluation of the MT output (Section 3), which requires source language competence (Castilho et al. 2018). The Cambridge English test provides a score which is then converted into one of the six CEFR (Common European

Framework of Reference for Languages) levels, from A1 (lowest proficiency) to C2 (highest proficiency), each associated with different reception, production and interaction skills in a foreign language (Council of Europe 2011).

Evaluators varied in terms of English proficiency—Table 1 shows the number of evaluators who were assigned to each CEFR level by the Cambridge English test. Accordingly, during data analysis (Section 7), we calculated mean adequacy scores from all evaluators first, and then only from evaluators who were assigned a B1 CEFR level or higher, in order to identify and remove potential biases due to low English language proficiency. B1 was selected as a threshold because, unlike those at A1-A2 levels, users at B1 level "[c]an understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc." (Council of Europe 2011, 5).

Finally, regarding MT use, the vast majority of participants (n=36) reported using MT to understand information in an unknown language. Most of them reported using MT either always or frequently (n=28), while the remainder indicated that they used MT rarely (n=8).

Level of English	Number of evaluators
A1	1
A2	8
B1	13
B2	5
B2 - C1	4
C1	2
C1 - C2	5
C2	3

Table 1: Level of English of MT evaluators

5 Experimental Materials

Since twelve Cochrane volunteers took part in our authoring study (Section 4), the following experimental materials were available to us: (i) a corpus of twelve non-automated PLS and their Spanish MT outputs; (ii) the corresponding corpus of semi-automated PLS (i.e. revised

by using Acrolinx) and their Spanish MT outputs.

Due to the inconsistencies in Cochrane PLS guidance (Section 2), the length of the experimental materials produced with the non-automated approach varied between 300 and 700 words (M=437 words, SD=120.95). The set of semi-automated PLS roughly contained the same average number of words (M=436.41 words, SD=132.85).

6 Experimental Design, Tasks and Procedures

Authoring Study

We adopted a within-subject design—each author engaged with both Cochrane PLS guidance and the Acrolinx CL checker on the same PLS. More precisely, Cochrane authors who agreed to take part in this study were asked to:

(i) Complete a pre-task questionnaire to determine if they were eligible (i.e. if they had authored a PLS in the past) and to collect background information;

(ii) Complete the SUS on their satisfaction with the non-automated approach (i.e. Cochrane PLS guidance);

(iii) Install the TeamViewer[v] free software on their computers, which allowed for remote access to other computers;

(iv) By using TeamViewer, access a computer owned by one of the researchers, where Acrolinx was installed as a Microsoft Word plugin, and where their past PLS was ready for them to edit/revise using this CL checker. However, before the main editing task, authors conducted a warm-up task on a sample PLS (also provided on the researcher's computer) to familiarize themselves with Acrolinx. Figure 1 shows a participant conducting the warm-up task.

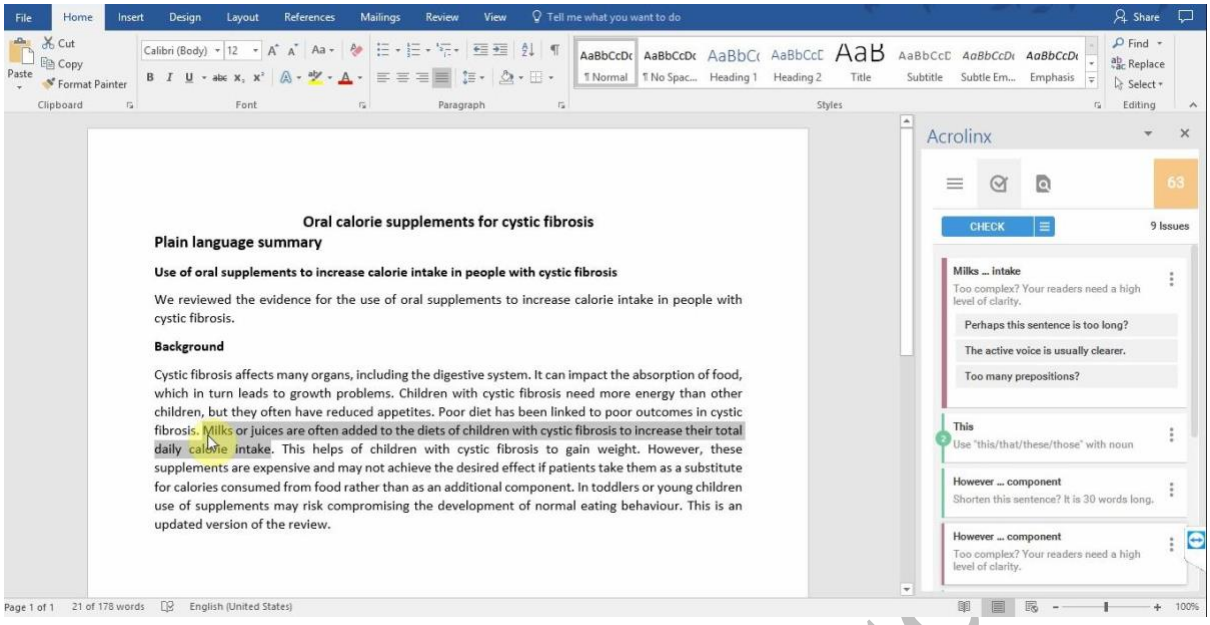

Figure 1: Warm-up task with Acrolinx

(v) Edit their past PLS in Microsoft Word by following Acrolinx suggestions. However, authors were instructed to use their common sense in deciding whether to apply a change recommended by Acrolinx; and

(vi) Complete the SUS on their satisfaction with Acrolinx and answer the remaining preference questions.

Authors were provided with instructions for each of the tasks described above. No time limit was set for the warm-up task or the main editing task. This decision was taken to maximize the ecological validity of our study since Cochrane authors do not produce PLS under time pressure[vi]. With regard to the characteristics of Acrolinx, we used the Microsoft Word Sidebar Edition which shows readability and translatability issues in a sidebar in Word, along with examples and suggestions on how to edit them in the text. Examples of Acrolinx recommendations are: "Simplify word", or "Use 'this/that/these/those' with noun". Customizing Acrolinx CL rules for Cochrane content was not possible. However, to make the two simplification approaches as comparable as possible, we deactivated any Acrolinx CL rules which contravened Cochrane guidelines (e.g. on the use of hyphens after prefixes).

MT Evaluation Study

A within-subject design was also adopted for the MT evaluation study, where each participant conducted two evaluation/scoring tasks (Section 3). In particular, Cochrane evaluators were asked to:

(i) Complete a pre-task questionnaire to test their eligibility (i.e. that they were native speakers of Spanish and had a background in the health field). Additional questions were asked about their profiles, particularly on their level of English and on their use of MT; and

(ii) Conduct the evaluation tasks of two Spanish MT outputs.

Both the pre-task questionnaire and the two MT evaluation tasks were presented to participants online, on Google Forms. Each sentence in each MT output was evaluated in terms of fluency and adequacy by either three or four participants, depending on the number of volunteers recruited. Dyson and Hannah (1987) recommend at least three evaluators.

To compensate for fatigue effect, the order in which non-automated and semi-automated PLS were presented to evaluators was counterbalanced (MacKenzie 2013). Overall, authors were divided into twelve groups and different texts were assigned to each group. In groups 1-6, the semi-automated PLS was presented first, while in groups 7-12 the non-automated PLS was presented first. Moreover, the two PLS (and corresponding MT outputs) assigned to each evaluator dealt with two different health-related topics.

7 Results and Discussion

Volunteer authors' satisfaction and preferences

Table 2 shows the SUS scores that the twelve authors assigned to Cochrane PLS guidelines and Acrolinx. Results from the ten Likert-type items that compose the SUS were combined into a single score prior to being analysed because, taken individually, they do not relate to

any specific characteristic of a system (Brooke 2013). In addition to the mean and the standard deviation (SD), the median was also included in the analysis because it is less influenced by extreme values (Doane and Seward 2011).

Simplification approach	SUS score (mean)	SUS score (SD)	SUS score (median)
Non-automated (Cochrane PLS guidelines)	62.29	26.53	70
Semi-automated (Acrolinx)	75.41	14.49	78.75

Table 2: Descriptive statistics of the SUS scores assigned by the twelve authors

The mean and the median rates in Table 2 indicate that using Acrolinx to check the readability and translatability of a PLS (i.e. making the simplification approach semi-automated) resulted in a higher level of authors' satisfaction. However, a paired t-test conducted using the statistical software package *Stata 12.1* showed that the difference was not statistically significant, $t(11)=1.2549$, $p=0.2355$.

To complement the SUS results, authors were also asked which type of authoring support they would use when producing PLS in the future. Figure 2 reports the authors' responses. The horizontal axis indicates the number of participants who selected each option.

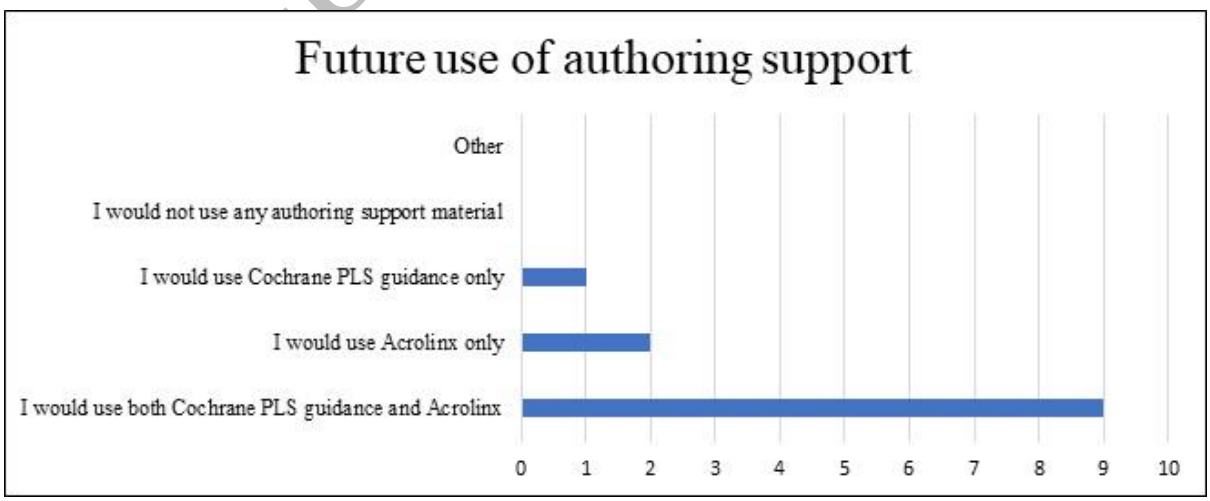

Figure 2: Future use of authoring support for PLS

Most participants (n=9) reported that, in the future, they would use both Cochrane PLS guidance and Acrolinx, i.e. they would welcome the integration of a CL checker into the simplification workflow[vii]. Only a small percentage of participants reported that they would use either Cochrane guidance only or Acrolinx only. No participant reported that they would produce a PLS without using any support. This result is not surprising considering that all the participants were either health professionals or academics in the health field who were likely to be more familiar with specialised medical language than with plain language (Section 4).

Two authors (P05 and P15) answered that they would like to use only Acrolinx in the future. P05 provided as a reason the fact that Acrolinx readability suggestions were specific. However, this participant also specified that they struggled to increase the Acrolinx readability score of their PLS because the software did not recognize medical terminology. They added that this issue could be solved by creating an Acrolinx term set for the medical field (which is indeed possible). P15 reported that they found the software very useful to improve both the readability and style of the text. Moreover, P15 appreciated Acrolinx suggestions on translatability and reported that they had already been using an authoring support tool called Grammarly, thus suggesting that Cochrane PLS guidance alone might not be enough.

This was the first time that participants had interacted with Acrolinx. Therefore, there was some agreement on the need to practice with Acrolinx before becoming familiar with the tool. For this study, participants were assigned a warm-up task before conducting the main simplification task on their PLS. Nonetheless, the time that participants dedicated to the warm-up varied greatly.

MT quality

To analyse the quality of the Spanish MT outputs, we calculated the average adequacy and fluency score (across all text sentences) per evaluator (Koehn and Monz 2006). The average

adequacy score per evaluator ranged from 2.47 to 4, while the average fluency score per evaluator ranged from 1.04 to 4. For each evaluator, an independent-samples t-test was conducted in *Stata 12.1* to determine if the average score that each evaluator assigned to the non-automated PLS was significantly different from the average score that they assigned to the semi-automated PLS at the 0.05 significance level. For each evaluator, Table 3 reports the average adequacy scores assigned to both texts. Similarly, Table 4 presents the average fluency scores per evaluator. Statistically significant differences are signalled with an asterisk. Moreover, we calculated a grand mean of fluency and adequacy scores for both corpora to get an overall picture of the evaluations.

Groups	**Evaluators**	**Non-automated PLS**	**Semi-automated PLS**
		Mean (SD) adequacy score	**Mean (SD) adequacy score**
1	E13	3.73 (0.44)	3.89 (0.31)
	E25	4 (0)	4 (0)
	E01	3.43 (0.66) (*)	3.94 (0.22) (*)
	E37	4 (0)	4 (0)
2	E14	3.88 (0.47)	4 (0)
	E02	4 (0)	3.96 (0.2)
	E26	3.88 (0.47)	3.68 (0.69)
3	E15	3.76 (0.53)	3.72 (0.45)
	E03	4 (0)	3.97 (0.16)
	E27	4 (0)	3.89 (0.31)
	E39	3.52 (0.81)	3.13 (0 .78)
4	E16	4 (0)	3.86 (0.44)
	E04	3.89 (0.4)	3.82 (0.46)
	E40	4 (0)	3.89 (0.3)
	E28	3.93 (0.25)	3.82 (0.6)
5	E17	4 (0)	4 (0)
	E05	2.7 (0.6)	2.54 (0.8)
	E41	3.48 (0.75)	3.68 (0.47)
6	E18	3.55 (0.82)	3.5 (0.84)
	E42	3.9 (0.44)	3.75 (0.61)
	E30	3.95 (0.22)	3.93 (0.25)
7	E31	3.85 (0.6)	4 (0)
	E49	3.25 (0.98) (*)	3.78 (0.61) (*)
	E19	3.29 (0.77) (*)	3.95 (0.21) (*)
8	E08	3.86 (0.35)	3.89 (0.3)

	E20	3.93 (0.25)	4 (0)
	E32	3.55 (0.73)	3.34 (0.76)
	E44	3.82 (0.46)	3.93 (0.37)
9	E21	3.93 (0.25)	3.96 (0.19)
	E09	4 (0)	4 (0)
	E33	3.83 (0.46)	3.85 (0.36)
	E45	3.86 (0.34)	3.77 (0.42)
10	E10	3.63 (0.72)	3.77 (0.61)
	E34	3.77 (0.52)	3.83 (0.37)
	E46	3.86 (0.35)	3.96 (0.17)
11	E23	3.82 (0.38)	3.76 (0.6)
	E50	3.73 (0.54)	3.8 (0.54)
	E52	2.47 (0.59) (*)	3.5 (0.69) (*)
12	E12	3.34 (0.7)	3.5 (0.65)
	E24	3.84 (0.36)	3.75 (0.44)
	E48	3.46 (0.62)	3.7 (0.46)
GRAND Means (SD)		**3.72 (0.33)**	**3.78 (0.27)**

Table 3: Descriptive and inferential statistics on adequacy scores

Groups	**Evaluators**	**Non-automated PLS**	**Semi-automated PLS**
		Mean (SD) fluency score	**Mean (SD) fluency score**
1	E13	3.56 (0.58)	3.47 (0.77)
	E25	3.17 (1.07)	3.21 (1.13)
	E01	2.26 (1.17)	2.73 (1.04)
	E37	3.65 (0.57)	3.63 (0.68)
2	E14	3.11 (0.9) (*)	3.64 (0.56) (*)
	E02	3.38 (0.69)	3.36 (0.63)
	E26	3.33 (0.76)	3.12 (1.01)
3	E15	3.71 (0.56)	3.62 (0.59)
	E03	3.71 (0.46)	3.59 (0.64)
	E27	3.04 (1.07)	3.13 (0.91)
	E39	2.8 (0.81)	2.67 (0.7)
4	E16	3.79 (0.49)	3.65 (0.61)
	E04	3.79 (0.55)	3.65 (0.72)
	E40	3.24 (0.57)	3.44 (0.63)
	E28	3.68 (0.66)	3.75 (0.73)
5	E17	2.85 (0.76)	3.13 (0.63)
	E05	2.07 (0.72)	1.95 (1.04)
	E41	2.55 (1.01)	2.77 (0.86)
6	E18	3.9 (0.3)	3.72 (0.54)
	E42	3.45 (0.75)	3.54 (0.72)
	E30	2.55 (1.19)	2.7 (1)
7	E31	3.59 (0.79) (*)	3.92 (0.34) (*)

	E49	3.22 (0.93)	3.53 (0.67)
	E19	3 (0.96)	3.36 (0.58)
8	E08	3.75 (0.51)	3.86 (0.35)
	E20	3.89 (0.3)	3.96 (0.18)
	E32	3.13 (1.02)	2.62 (1.04)
	E44	3.68 (0.54)	3.62 (0.67)
9	E21	3.8 (0.48)	3.51 (0.7)
	E09	3.86 (0.34)	3.85 (0.45)
	E33	3.46 (0.62)	3.59 (0.57)
	E45	3.36 (0.88)	3.37 (0.92)
10	E10	3.09 (1.1)	3.61 (0.84)
	E34	2.22 (0.97)	2.09 (0.9)
	E46	3.04 (0.84)	3.35 (0.7)
11	E23	3.69 (0.47)	3.84 (0.36)
	E50	3.78 (0.51)	3.63 (0.71)
	E52	1.04 (0.2) (*)	1.6 (1.1) (*)
12	E12	3.78 (0.42) (*)	4 (0) (*)
	E24	3.9 (0.29)	3.87 (0.33)
	E48	3.31 (0.78)	2.91 (0.77)
GRAND Means (SD)		**3.27 (0.6)**	**3.33 (0.55)**

Table 4: Descriptive and inferential statistics on fluency scores

From Tables 3 and 4 it emerges that: (i) in terms of adequacy, the number of evaluators who rated semi-automated PLS higher (n=19) was just one greater than the number of evaluators who rated non-automated PLS higher (n=18); (ii) the number of evaluators who assigned a higher average fluency score to semi-automated PLS (n=22) was very slightly higher than the number of evaluators who assigned a higher average fluency score to non-automated PLS (n=19); (iii) most differences in average scores were not statistically significant. The only statistically significant increases in adequacy and fluency scores (for eight participants in total) were observed for semi-automated PLS. Moreover, after excluding evaluators with A1-A2 level of English proficiency (Section 4) and recalculating the grand mean of adequacy scores, we observed that the difference between the grand mean score assigned to non-automated PLS (M=3.76, SD=0.28) and the grand mean score assigned to semi-automated PLS (M=3.82, SD=0.25) remained slight. Overall, there was little difference in evaluators'

ratings of fluency and adequacy between the two corpora of PLS, thus indicating that the MT system used (i.e. Google Translate) did not consistently produce better raw MT quality when Acrolinx was integrated into the simplification process.

Overall fluency and adequacy scores were relatively high, suggesting that the MT system produced reasonably good raw MT quality. We would need to examine the Spanish quality and verify that the evaluations are fair. At the same time, evaluators are domain experts with some source text knowledge and we assume that they are reasonably capable of making a credible judgement of the content. A random sample check of scores suggests that the evaluation task was executed credibly, and the scores are fair judgements of quality. Moreover, evaluators who consistently assigned four on adequacy, gave varied scores on fluency (and vice versa), thus providing further indication that they paid attention to the different characteristics of the sentences. The data reported in Tables 3 and 4 also show that the vast majority of evaluators (n=38) assigned higher mean scores for adequacy than for fluency for both corpora of PLS—this is also reflected in the grand means, which are higher for the adequacy measure.

8 Conclusions and Future Work

This study focused on volunteer health practitioners facilitating the intralingual and interlingual translation of medical content at Cochrane. More precisely, we examined whether introducing the Acrolinx CL checker into Cochrane's non-automated simplification approach for the production of PLS would be beneficial in terms of volunteer authors' satisfaction and machine translatability (into Spanish) of their PLS. In this section, we will summarize and discuss the main findings, and we will present areas for future work.

We observed that most authors were satisfied with the CL checker, and would welcome its introduction to supplement Cochrane PLS guidelines during the simplification/intralingual translation tasks. This tool would reduce the need to remember,

check and manually implement plain language guidelines, which might represent a burden, particularly for individuals with no training in linguistics. A possible way to maximize the benefits of integrating a CL checker and guidelines would be to use the former to check for compliance with simplification rules that can be formalized, and the latter at the summarization stage, i.e. when authors need guidance on which content of the Systematic Reviews should be included in the PLS. However, CL checkers should be tailored to the characteristics of Cochrane content, and authors would need some time to familiarise themselves with this semi-automated approach.

When volunteering involves content editing and production (as in the case of Cochrane or Wikipedia), Nov and Rao (2008, 85) shed light on the threat of *asymmetry*, defined as "a lack of contributed resources for maintaining and improving the common pool of resources". In other words, there is some risk that volunteers' contributions will be scarce and not sufficient to expand and update content. Ensuring the commitment of volunteers to the authoring task (by boosting their satisfaction) might reduce the threat of asymmetry, and result in an increase in "volunteering rates" (Olohan 2014, 19), namely the amount of health content made available in plain English on different websites: as part of the Cochrane-Wikipedia partnership, evidence from Cochrane Systematic Reviews is being used in Wikipedia medical articles to ensure their accuracy (Cochrane 2016).

However, satisfaction with a task might not always result in motivation to volunteer. Motivation is a complex psychological construct. Previous works have shown that volunteers involved in online communities can be self-motivated to write in the first place (Joyce and Kraut 2006), or that motivation is determined by the perceived importance and uniqueness of the contributions (Ling et al. 2005). According to Clary et al. (1998), volunteering can serve six functions, which are: expressing altruistic values; engaging in favourably viewed activities; advancing one's career; reducing one's sense of guilt; enhancing positive mood;

and understanding, which is defined as the possibility for volunteers to be involved in new learning experiences, and to put their skills and abilities into practice. Future research, e.g. involving structured or semi-structured interviews, might shed light on the factors that motivate Cochrane authors to volunteer, and on whether the use of automated tools such as CL checkers affects their motivation.

Future work should also measure the duration of the authoring tasks, since time commitment is another factor likely to influence authors' motivation to engage in volunteer intralingual translation tasks. Moreover, these findings regarding the satisfaction and preference of Cochrane authors should be tested in studies with a larger and more homogeneous sample of participants (in terms of native language, plain language writing skills, and familiarity with Cochrane sets of guidelines) in order to be generalisable.

Regarding machine translatability, using the Acrolinx CL checker did not result in a significant increase in quality of the Spanish MT output of PLS. In line with Wu et al. (2011), we also observed that Google Translate produced output of relatively high quality. This result seems encouraging for Cochrane and other non-profit organizations that are increasingly relying on MT systems to streamline their translation workflow and to encourage contributions from volunteers, particularly non-professional translators (Section 1). Furthermore, for both PLS corpora, adequacy was rated higher than fluency. This finding is also promising since, compared with adequacy/content errors, fluency/style issues are easier to correct and less likely to have a detrimental effect on the well-being of readers (Koponen 2010; Stymne 2013).

It remains to be investigated if volunteer health professionals would be willing to post-edit MT outputs rather than translating PLS from scratch. Future work should also compare the human interlingual translation vs. post-editing scenario of simplified texts. Moreover, it might be interesting to analyse the impact of other CL checkers on machine

translatability (into other languages), as well as to conduct the evaluation of the MT output at the document (rather than the sentence) level (Läubli, Sennrich and Volk 2018). Even though volunteer health professionals might need some training on how to validate or post-edit MT outputs, combining CL checkers and MT might reduce their workload and increase the amount of health information that is made available in different languages.

Funding and acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement N. 734211 ("Interact"); and by the Irish Research Council (GOIPG/2017/1409). The authors would like to thank the Acrolinx team for their technical support, and Dr Silvia Rodríguez Vázquez for her assistance with the setup of the authoring study. The authors would also like to thank the anonymous participants, and Therese Docherty, Juliane Reid, Hayley Hassan, and Andrea Cervera from Cochrane for helping with recruitment.

Disclosure Statement

The authors have no financial interest or benefit arising from the direct applications of their research.

References

Abekawa, Takeshi, and Kyo Kageura. 2007. "A Translation Aid System with a Stratified Lookup Interface." In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Poster and Demo Session*, 5-8. https://bit.ly/2LYqqy5

Bangor, Aaron, Philip Kortum, and James Miller. 2008. "An Empirical Evaluation of the System Usability Scale." *International Journal of Human-Computer Interaction* 24(6): 574-594. doi:10.1080/10447310802205776.

Birch, Alexandra, Ondřej Bojar, Rudolf Rosa, Juliane Ried, Hayley Hassan, and Colin Davenport. 2017. *D5.5: Report on User Surveys, Impact Assessment and Automatic Semantic Metrics.* https://goo.gl/EGQJXK

Bojar, Ondřej, Barry Haddow, David Mareček, Roman Sudarikov, Aleš Tamchyna, and Dušan Variš. 2017. *D1.1: Report on Building Translation Systems for Public Health Domain.* https://goo.gl/MzXdKj

Brooke, John. 1996. "SUS – A Quick and Dirty Usability Scale." *Usability Evaluation in Industry* 89 (194): 4-7. https://goo.gl/kyFkmJ

Brooke, John. 2013. "SUS: A Retrospective." *Journal of Usability Studies* 8 (2): 29-40. https://goo.gl/ckpbm4

Byrne, Michael, Kristen Greene, and Sarah Everett. 2007. "Usability of Voting Systems: Baseline Data for Paper, Punch Cards, and Lever Machines." In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, edited by Bo Begole, Stephen Payne, Elizabeth Churchill, Rob Amant, David Gilmore, and Mary Beth Rosson, 171-180. New York: ACM.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Miceli Barone, and Maria Gialama. 2017. "A Comparative Quality Evaluation of PBSMT and NMT Using Professional Translators." In *Proceedings of MT Summit XVI, vol. 1*, edited by Sadao Kurohashi and Pascale Fung, 116-131. https://goo.gl/pHwfdB

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. "Approaches to Human and Machine Translation Quality Assessment." In *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 9-38. Basel, Switzerland: Springer International Publishing AG.

Chandler, Jackie, Julian Higgins, Jonathan Deeks, Clare Davenport, and Mike Clarke. 2017. "Chapter 1: Introduction." In *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.2.0,* edited by Julian Higgins, Rachel Churchill, Jackie Chandler, and Miranda Cumpston. https://goo.gl/BRfoYQ

Chaparro, Barbara, Mikki Phan, Christina Siu, and Jo Jardina. 2014. "User Performance and Satisfaction of Tablet Physical Keyboards." *Journal of Usability Studies* 9 (2): 70-80. https://goo.gl/qB1x9P

Clary, Gyl, Mark Snyder, Robert Ridge, John Copeland, Arthur Stukas, Julie Haugen, and Peter Miene. 1998. "Understanding and Assessing the Motivations of Volunteers: A Functional Approach." *Journal of Personality and Social Psychology* 74 (6): 1516-1530. doi:10.1037/0022-3514.74.6.1516

Cochrane. 2016. "The Cochrane-Wikipedia Partnership in 2016". https://goo.gl/Lh7b7z

Council of Europe 2011. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. https://bit.ly/2w4iMwg

Den Besten, Matthijs, and Jean-Michel Dalle. 2008. "Keep it Simple: A Companion for Simple Wikipedia?" *Industry and Innovation* 15 (2): 169-178. doi:10.1080/13662710801970126

Doane, David, and Lori Seward. 2011. "Measuring Skewness: A Forgotten Statistic?" *Journal of Statistics Education* 19 (2): 1-18. doi:10.1080/10691898.2011.11889611

Doherty, Stephen. 2017. "Issues in Human and Automatic Translation Quality Assessment." In *Human Issues in Translation Technology*, edited by Dorothy Kenny, 131-148. London: Routledge.

Dyson, Mary, and Jean Hannah. 1987. "Toward a Methodology for the Evaluation of Machine Assisted Translation Systems." *Computers and Translation* 2 (3): 163-176. https://goo.gl/fWC1ZX

Fiederer, Rebecca, and Sharon O'Brien. 2009. "Quality and Machine Translation: A Realistic Objective." *Journal of Specialised Translation* 11: 52-74. https://goo.gl/D5wf5z

Gigliotti, Giulia. 2017. "The Quality of Mercy: A Corpus-Based Analysis of the Quality of Volunteer Translations for Non-Profit Organisations (NPOs)." *New Voices in Translation Studies* 17: 52-81. https://goo.gl/L8eqjL

Heilman, James, Eckhard Kemmann, Michael Bonert, Anwesh Chatterjee, Brent Ragar, Graham Beards, David Iberri et al. 2011. "Wikipedia: A Key Tool for Global Public Health Promotion." *Journal of Medical Internet Research* 13(1): e14. doi:10.2196/jmir.1589

Higgins, Julian, and Sally Green, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0*. https://goo.gl/x1mgRS

ISO (International Standardization Organization) 2018. *ISO 9241-11:2018. Ergonomics of Human-System Interaction — Part 11: Usability: Definitions and Concepts.* Preview: https://goo.gl/57dkf8

Joyce, Elizabeth, and Robert Kraut. 2006. "Predicting Continued Participation in Newsgroups." *Journal of Computer-Mediated Communication* 11: 723-747. doi:10.1111/j.1083-6101.2006.00033.x

Kajzer-Wietrzny, Marta, Boguslawa Whyatt, and Katarzyna Stachowiak. 2016. "Simplification in Inter- and Intralingual Translation — Combining Corpus Linguistics, Key Logging and Eye-Tracking." *Poznan Studies in Contemporary Linguistics* 52 (2): 235-237. doi:10.1515/psicl-2016-0009

Koehn, Philipp, and Christof Monz. 2006. "Manual and Automatic Evaluation of Machine Translation between European Languages." In *Proceedings of the Workshop on Statistical Machine Translation (HLT-NAACL 06)*, edited by Philipp Koehn, and Christof Monz, 102-121. Stroudsburg: Association for Computational Linguistics.

Koponen, Maarit. 2010. "Assessing Machine Translation Quality with Error Analysis." In *Electronic Proceedings of the KäTu Symposium on Translation and Interpreting Studies.* https://bit.ly/2Ni6c7l

Laurent, Michaël, and Tim Vickers. 2009. "Seeking Health Information Online: Does Wikipedia Matter?" *Journal of the American Medical Informatics Association* 16 (4): 471-479. doi:10.1197/jamia.M3059

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. "Has Machine Translation Achieved Human Parity? A Case for Document-Level Evaluation." https://bit.ly/2M4cumf

Leroy, Gondy, David Kauchak, and Obay Mouradi. 2013. "A User-Study Measuring the Effects of Lexical Simplification and Coherence Enhancement on Perceived and Actual Text Difficulty." *International Journal of Medical Informatics* 82(8): 717-730. doi:10.1016/j.ijmedinf.2013.03.001

Lewis, James, and Jeff Sauro. 2009. "The Factor Structure of the System Usability Scale." In *Proceedings of the 13th International Conference on Human Computer Interaction (HCII 2009)*, edited by Masaaki Kurosu, 94-103. New York: Springer.

Ling, Kimberly, Gerard Beenen, Pamela Ludford, Xiaoqing Wang, Klarissa Chang, Xin Li, Dan Cosley et al. 2005. "Using Social Psychology to Motivate Contributions to Online Communities." *Journal of Computer-Mediated Communication* 10(4). doi:10.1111/j.1083-6101.2005.tb00273.x

Linguistic Data Consortium 2002. *Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations*. https://goo.gl/FMp5wj

MacKenzie, Scott. 2013. *Human-Computer Interaction: An Empirical Research Perspective*. Burlington, Massachusetts: Morgan Kaufmann.

Maguire, Lisa, and Mike Clarke. 2014. "How Much do You Need: A Randomised Experiment of whether Readers Can Understand the Key Messages from Summaries of Cochrane Reviews without Reading the Full Review." *Journal of the Royal Society of Medicine* 107 (22): 444-449. doi:10.1177/0141076814546710

McDonough Dolmaya, Julie. 2012. “Analyzing the Crowdsourcing Model and Its Impact on Public Perceptions of Translation.” *The Translator* 18 (2): 167-191. doi:10.1080/13556509.2012.10799507

Miyata, Rei, Anthony Hartley, Kyo Kageura, and Cécile Paris. 2017. “Evaluating the Usability of a Controlled Language Authoring Assistant.” *The Prague Bulletin of Mathematical Linguistics* 108: 147-158. doi:10.1515/pralin-2017-0016

Muñoz-Miquel, Ana. 2012. “From the Original Article to the Summary for Patients: Reformulation Procedures in Intralingual Translation.” *Linguistica Antverpiensia, New Series–Themes in Translation Studies* 11: 187-206. https://bit.ly/2w8FOCR

Nov, Oded, and Bharat Rao. 2008. “Technology-Facilitated ‘Give According to Your Abilities, Receive According to Your Needs’”. *Communications of the ACM* 51 (5): 83-87. doi:10.1145/1342327.1342342

O’Brien, Sharon. 2010. “Controlled Language and Readability.” In *Translation and Cognition - ATA Scholarly Monograph XV Series*, edited by Gregory Shreve, and Erik Angelone, 143-168. Amsterdam: John Benjamins.

Olohan, Maeve. 2014. “Why do You Translate? Motivation to Volunteer and TED Translation.” *Translation Studies* 7 (1): 17-33. doi:10.1080/14781700.2013.781952

Parra Escartín, Carla, Sharon O’Brien, Marie-Josée Goulet, and Michel Simard. 2017. “Machine Translation as an Academic Writing Aid for Medical Practitioners.” In *Proceedings of MT Summit XVI, vol.* 1, edited by Sadao Kurohashi, and Pascale Fung, 254-267. https://goo.gl/pHwfdB

Pérez-González, Luis, and Şebnem Susam-Saraeva. 2012. "Nonprofessionals Translating and Interpreting." *The Translator* 18 (2): 149-165. doi:10.1080/13556509.2012.10799506

Pym, Anthony. 2011. "What Technology Does to Translating." *Translation & Interpreting* 3(1): 1-9. https://bit.ly/2PEhfoG

Rada, Gabriel, Daniel Pérez, and Daniel Capurro. 2013. "Epistemonikos: A Free, Relational, Collaborative, Multilingual Database of Health Evidence." *Studies in Health Technology and Informatics* 192: 486-490. doi:10.3233/978-1-61499-289-9-486

Rodríguez Vázquez, Silvia. 2016. "Assuring Accessibility during Web Localisation: An Empirical Investigation on the Achievement of Appropriate Text Alternatives for Images." PhD dissertation, University of Geneva.

Sauro, Jeff, and James Lewis. 2011. "When Designing Usability Questionnaires, does It Hurt to Be Positive?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*, edited by Desney Tan, Geraldine Fitzpatrick, Carl Gutwin, Bo Begole, and Wendy Kellogg, 2215-2224. New York: Association for Computing Machinery.

Schriver, Karen. 2017. "Plain Language in the US Gains Momentum: 1940-2015." *IEEE Transactions on Professional Communication* 60(4): 343-366. doi:10.1109/TPC.2017.2765118

Stymne, Sara. 2013. "Using a Grammar Checker and Its Error Typology for Annotation of Statistical Machine Translation Errors." In *Proceedings of the 24th Scandinavian Conference of Linguistics (24SCL)*. https://bit.ly/2x5HRbr

Temnikova, Irina. 2012. “Text Complexity and Text Simplification in the Crisis Management Domain.” PhD dissertation, University of Wolverhampton.

The Cochrane Collaboration 2013. *Standards for the Reporting of Plain Language Summaries in new Cochrane Intervention Reviews (PLEACS)*. https://goo.gl/w8FpFv

TwB (Translators without Borders). 2016. “Translators without Borders Develops the World’s First Crisis-Specific Machine Translation System for Kurdish Languages.” https://goo.gl/tuvv2W

TwB (Translators without Borders). 2018a. “Development & Preparedness.” https://goo.gl/RP2BDb

TwB (Translators without Borders). 2018b. “Translators without Borders Code of Conduct for Translators.” https://bit.ly/2NONC3L

TwB (Translators without Borders). 2019. “Kató – TWB’s Translation Platform.” https://goo.gl/si6aHp

Von Elm, Erik, Philippe Ravaud, Harriet MacLehose, Lawrence Mbuagbaw, Paul Garner, Juliane Ried, and Xavier Bonfill. 2013. “Translating Cochrane Reviews to Ensure that Healthcare Decision-Making Is Informed by High-Quality Research Evidence.” *PLOS Medicine* 10 (9): e1001516. doi:10.1371/journal.pmed.1001516

Wu, Cuijun, Fei Xia, Louise Deleger, and Imre Solti. 2011. “Statistical Machine Translation for Biomedical Text: Are We There Yet?” In *Proceedings of the AMIA Annual Symposium*, 1290-1299. Bethesda, Maryland: American Medical Informatics Association.

Zethsen, Karen Korning. 2009. “Intralingual Translation: An Attempt at Description.” *Meta* 544: 795–812. doi:10.7202/038904ar

Appendix A.

	Strongly disagree				*Strongly agree*
1. I think that I would like to use this system frequently	○	○	○	○	○
2. I found the system unnecessarily complex	○	○	○	○	○
3. I thought the system was easy to use	○	○	○	○	○
4. I think that I would need the support of a technical person to be able to use this system	○	○	○	○	○
5. I found the various functions in this system were well integrated	○	○	○	○	○
6. I thought there was too much inconsistency in this system	○	○	○	○	○
7. I would imagine that most people would learn to use this system very quickly	○	○	○	○	○
8. I found the system very cumbersome to use	○	○	○	○	○
9. I felt very confident using the system	○	○	○	○	○
10. I needed to learn a lot of things before I could get going with this system	○	○	○	○	○

[i] The Cochrane Collaboration website is available at: https://www.cochrane.org/ [Accessed 7 January 2019]
[ii] An example of Cochrane PLS is available at: https://goo.gl/WKGpBS [Accessed 7 January 2019]
[iii] The Acrolinx website is available at: https://www.acrolinx.com/ [Accessed 7 January 2019]
[iv] The Cambridge English Test is available at: https://goo.gl/BFt6zF [Accessed 7 January 2019]
[v] A description of TeamViewer is available at: https://www.teamviewer.com/en/ [Accessed 7 January 2019]
[vi] We could not recruit authors who would produce PLS from scratch for our experiment. Producing a PLS from an entire Systematic Review is an onerous and time-consuming task that volunteer authors conduct in different sessions during their free time. Our only option was to recruit authors who had already produced a PLS in the past, and ask them to check and edit their PLS using Acrolinx. While this choice meant that we were not able to measure task duration, nor to control the conditions under which the production of PLS with Cochrane guidelines took place, it did enhance the ecological validity of the study since a laboratory/controlled setting would have represented an unrealistic scenario for Cochrane authors.
[vii] It is worth noting that authoring a PLS from scratch while consulting sets of guidelines and at the same time receiving automatic notifications on violations of readability and translatability might be overwhelming and distracting for authors. Therefore, we suggest that if Acrolinx (or similar) were integrated into the overall workflow of PLS production, it be used as a way to check the readability and translatability of an already existing draft. Moreover, text simplification is often an iterative process, in which one draft is successively edited and refined (Schriver 2017).