Integrated Post-Editing and Translation Management for Lay User Communities

Adrian Laurenzi
Department of Computer Science and Engineering
University of Washington
Seattle, WA, USA
alaurenz@uw.edu

Megumu Brownstein
Northwest Center of Public Health Practice
University of Washington
Seattle, WA, USA
megumu@uw.edu

Anne M. Turner
Department of Health Services
Department of Biomedial Informatics and Medical Education
University of Washington
Seattle, WA, USA
amturner@uw.edu

Katrin Kirchhoff
Department of Electrical Engineering
University of Washington
Seattle, WA, USA
kk2@uw.edu

Abstract

Over the past decade machine translation has reached a high level of maturity and is now routinely utilized by a wide variety of organizations, including multinational corporations, language service providers, and governmental/non-profit organizations. However, there are many communities that could benefit greatly from machine translation but do not actively use it, either because of a lack of awareness of its current capabilities, or because of real or perceived barriers to adopting machine translation technology. In this paper we present a case study of introducing machine translation in combination with post-editing to one such community, namely employees at local and regional public health departments in the U.S. We describe a methodology for determining their translation needs, and describe the development of an integrated post-editing and translation management system specifically targeted to their typical workflow. We report results from user testing and participatory design studies and conclude with a set of recommendations and best practices for introducing machine translation plus post-editing to lay user communities.

1 Introduction

Over the past decade the development of machine translation (MT) technology has made rapid progress and has reached a high level of maturity. MT is now routinely being used for a variety of tasks. Machine translation plus post-editing (MT+PE) has been shown to significantly increase translator productivity (see e.g. (Guerberof, 2009; Plitt and Masselot, 2010; Green et al., 2013)) and has become a common procedure for many language service providers, corporations, and government organizations.

However, there are still many communities that could potentially benefit from MT but that do not currently use it. These are often non-profit, educational, faith-based, or research organizations whose members may be experts in a particular domain but who are "lay" users from an MT perspective, i.e. they are not trained translators or post-editors. Lay communities may be prevented from using MT by a lack of awareness of the current capabilities of MT systems, lack of technological know-how, costs, or by an inherent bias against MT. The adoption of MT technology would likely make the work of these communities more widely accessible, which would ultimately result in a benefit to the public. In addition it might provide unique insights or interesting data for future MT research, e.g. data from non-mainstream language pairs or specialized domains. Finally, lay communities provide a more general testbed for user modeling, adaptation, and human-computer interface design, which all need to be addressed if MT is to become a ubiquitous technology.

In order to introduce MT to lay communities it is necessary to study their current translation practices, actual translation needs, and, most importantly, to develop a model for integrating MT+PE into their typical workflows. In this paper we describe our experiences with introducing MT+PE to a community of public health professionals in lo-

Sharon O'Brien, Michel Simard and Lucia Specia (eds.)
Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, September 2, 2013, p. 27–34.

cal and regional health departments in the U.S. Although our focus is on one particular community we believe that our case study can serve as a general model for lay user communities.

The paper is structured as follows: Section 2 provides the background for this study. Section 3 describes an initial data gathering phase and a pilot PE study to shed light on translation practices, translation needs, typical workflows, and the feasibility of having lay users post-edit translations of health and safety materials. Based on these initial observation we have developed an integrated post-editing and translation management system for lay user communities, described in Section 4. Section 4.2 details initial user studies and an iterative participatory design process used to refine the system. Section 5 compares this system to related work. We conclude with a summary of insights and best practices for introducing MT+PE to lay communities (Section 6).

2 Background: Machine Translation and Public Health Practice in the U.S.

The U.S. population is characterized by a fair amount of linguistic diversity. According to the 2011 American Community Survey estimates (ACS, 2011) 20.8% of the population over 5 years of age speak a language other than English at home; of these, 41.8% report speaking English "less than very well". This percentage is even higher for certain demographic groups; e.g., it reaches 63.5% for Spanish speakers of 65 years of age or older. 24.7% to 27.7% of all households speaking Spanish or an Asian/Pacific-Island language are classified as linguistically isolated (all household members 14 years or older speak English less than "very well"). Such limited English proficiency (LEP) is correlated with adverse health outcomes. Previous studies have shown that LEP populations have more difficulty in gaining access to health care, fewer preventative health screenings, and poorer health status than English-speaking minority groups (Goel et al., 2003; Jacobs et al., 2004; Ponce et al., 2006).

This situation persists despite federal mandates requiring special provisions for LEP populations. For example, guidelines issued by the U.S. Department of Health and Human Services (DHHS) require that agencies receiving financial assistance from DHHS must take "reasonable steps" to make their services accessible to LEP populations, which includes linguistic accessibility. In the overall healthcare context, linguistic accessibility needs to be addressed at different levels and in different forms, including providing interpreting services during patient-provider interactions, hospital discharge instructions and consent forms in different languages, or translations of newsletters or flyers on disease prevention and available health services. Here we focus on the translation and dissemination of consumer-oriented health information documents created by public health departments.

The DHHS mandate does include providing translation of vital documents (DHHS, 2003). However, in practice there is a lack of high-quality, up-to-date health information materials in languages other than English, especially at the state and local community level. The primary reasons for this situation are the lack of funds and staff time to create multilingual documents. Currently, translation practices in regional public health departments are non-standardized and vary widely. In a survey of translation practices in regional health departments in the U.S. we found that they exclusively use traditional human translation processes. Departments typically contract with a small number of language service providers. When a translation is needed, the first step is to obtain quotes from providers. Documents are then sent out to the winning bidder. The average turnaround time for translations to be completed is 15 days, with a minimum of 2 days even for rush orders. Translations then go through another internal review and quality control step before they are published. This is a time-consuming process, especially in situations where a rapid response to an emerging health crisis is required. Additionally, health departments have very scarce financial and staff resources, e.g. one medium-sized health department in Washington state reported having a monthly budget of only $50 for translation work. Per-word translation costs reported to us by health departments participating in our study range between $0.20 and $1.73 (for rush orders); thus, even when using the lowest-cost service this budget allows for the translation of only 250 words per month.

MT could significantly accelerate and streamline the process of producing multilingual health information materials by eliminating the time-

consuming and costly step of outsourcing translation to external vendors. Under this model documents would first be translated automatically before being post-edited by a bilingual in-house staff member. We previously conducted a pilot study demonstrating that MT+PE leads to faster turnaround times and lower cost while the quality of the output is equivalent to human translations (**?**). However, our initial studies also indicated that, in order to be adopted as a standard tool, MT needs to be properly integrated into the typical workflow of public health departments and needs to be adapted to employees' needs as far as possible. Another barrier consists of attitudes and beliefs about MT. We found that employees typically are not aware of the quality of state-of-the-art MT engines, the concept of PE, or of standard support tools available.

3 Initial Feasibility Study

We conducted an initial study of human factors involved in integrating MT+PE into the standard workflow of public health departments. Our focus was on health and safety information documents (regarding e.g., vaccines, preventative screenings, infectious diseases, and maternal and child health) that are disseminated as websites, flyers, or mass mailings. We conducted 41 semi-structured interviews and 4 focus groups with health department staff involved in translation processes. The health departments included a state health department, a large municipal health department and several rural health departments that serve populations with a high percentage of LEP speakers. The participants were asked to provide a description of their current translation processes, obstacles and incentives to creating multilingual information materials, and attitudes towards MT. Transcripts of these interactions were coded using the Cognitive Workflow Analysis framework (Vicente, 1999) and the method of constant comparisons (Glaser, 1965).

The most important insights from this study can be summarized as follows:
1. Health departments typically do not have dedicated budgets or support staff for acquiring, installing and maintaining translation software, or for training employees in its use. Therefore, any MT based system targeted at health departments must be as low-cost as possible, intuitive, and easy to use, requiring as little prior technological knowledge or initial training time as possible.
2. There is a moderate volume of translation work overall, but it may spike in response to emergencies (e.g., disease outbreaks or natural disasters) and require fast turnaround times.
3. Employees do not work on translation continuously but intermittently, in addition to other work tasks. There is a need for tracking post-editing progress, saving intermediate and partially completed work, and accessing different evolving versions of the same document (version control).
4. Different health departments could benefit greatly from sharing already-translated documents and language expertise. Currently there is no system in place to archive and share translations. Health departments often have one or two bilingual staff members but they only represent the largest language groups in the areas they serve. Thus, ideally the system should also support the online collaboration of geographically distributed workers with complementary language expertise, and the sharing of translated documents.
5. Health department employees are very concerned about the accuracy of their translated documents since they convey health and safety information. In addition translations need to be culturally appropriate and are often targeted to a specific demographic group. An MT-based system needs to have provisions for multiple layers of quality control and needs to be able to accommodate information about requirements for specific documents (e.g. desired reading level, target group, etc.).

We next conducted a pilot PE study with public-health professionals to assess whether staff members who are domain experts but not trained translators can post-edit translated documents to an acceptable standard. To this end we implemented a Java-based in-house tool that provides a simple PE interface and includes timing and keystroke logging. Source documents and their translations are displayed side-by-side in two aligned text windows, with one sentence per line. While the user is editing the translated sentences, the corresponding source sentences are highlighted. Optionally, the original non-edited machine translation can be redisplayed in a third window. Timing begins when the user first clicks inside the editable text area; it can be paused by clicking a button, which prevents the user from editing until they resume the session. All user keystrokes are logged in a single

text file per document, along with the time point at which each key was pressed. After the user has finished editing, the session is completed and the post-edited machine translation is saved as a text file along with meta-information (filename, start time, end time, pause times).

A total of 25 English health documents with an average word count of 923 (standard deviation: 452) were used for the study. Spanish translations of the documents were created using Google Translate. Eight bilingual health department staff members fluent in English and Spanish were recruited to post-edit the translations in two different sessions. They had worked in their current organization between 1.5 and 20 years and performed functions such as "Immunization Coordinator", "Health Services Consultant", or "Research coordinator". Two of them had previously worked as medical interpreters; all had some experience reviewing manual translations. However, none of them were trained translators or post-editors, and they did not have experience working with machine translation or with professional translation tools. They were instructed to perform all necessary edits to create grammatically correct and accurate translations in a timely manner. Four independent quality reviewers (trained translators or interpreters recruited from health departments) were then asked to perform a blind comparison of the post-edited machine translations and human translations of the same source documents. They classified translations according to whether they were equivalent or different, and if so, which one was preferred. Post-editors were also asked about their impressions of the machine translations, which errors they found most difficult to edit, and their impressions of the PE tool.

Post-editing took on average 24.5 minutes (standard deviation: 14.9) per document. We computed the overall PE time and the average duration of pauses for each document and post-editor but did not find any consistent patterns – there was neither a strong correlation between document length and PE time nor a consistent correlation between PE time and the number of documents already completed. Some post-editors became faster from Session 1 to Session 2, indicating a learning effect, but others did not. Timing seems to be largely dependent on the individual, with some post-editors taking more time to double-check and ensure correct translations after the initial post-editing pass, whereas others do a single integrated pass over the text. This is in line with similar observations reported in the literature (O'Brien, 2006; Koponen et al., 2012).

MT+PE preferred	HT preferred	Equivalent
18	16	16

Table 1: Number of votes assigned to categories in qualitative comparison of human-only translation (HT) and MT+PE output. Each of the 25 document was rated twice by two independent reviewers.

The quality rating results (Table 1) showed that overall the quality of post-edited documents did not differ from their human-translated counterparts – preference ratings for documents that were not judged equivalent were distributed approximately evenly across the three different categories. With regard to translation errors post-editors found word order errors the most difficult to process and to correct, followed by word sense errors. They did notice a fair amount of morphological errors but these were considered less distracting. Again, this is similar to results reported in other studies (Koponen et al., 2012). They uniformly found the PE interface intuitive and easy to use and did not voice any needs for more advanced functionality.

4 Post-Editing and Translation Management System Design

In this section we describe the development of an online system designed to be integrated in the day-to-day workflow in public health departments. An initial prototype was presented in (**?**). Since then, we have conducted user testing with actual public health professionals and have utilized their feedback to modify the system.

4.1 System Implementation

A prototype PE and translation management system has been implemented in the form of an web-based application using the Kohana PHP framework and a MySQL database. The front-end interface was built using JQuery, Twitter Bootstrap, HTML and CSS. There are four main modules that support the main tasks of (1) uploading a document, (2) applying MT, (3) post-editing MT output, and (4) sharing and downloading the finalized translation of a document. Users

can connect to the system via any standard web browser. After registering with the system they can upload a document in the source language (English). Registration includes establishing a user profile, which include information such as the user's agency or affiliation, language expertise, and experience (such as certifications from specific professional organizations). Upon being uploaded documents are automatically translated. Our current prototype system uses the Microsoft Translator API[1], which is free for low-volume use (up to 2M characters per month) and supports up to 39 languages. However, it is in principle possible to use the Google Translate API or any other API-based translation service instead. The translated document is then added to the pool of documents in the system. Users can start post-editing a translation by "claiming" the document, which locks a document and prevents other users from post-editing it simultaneously. Once a user "unclaims" a document, other users can access it to double-check or finish the post-editing. Users can save a claimed document even if post-editing has not yet been completed and can return to it at a later time. The system tracks the progress of each document through the translation pipeline and marks each of the four stages (1) uploaded, (2) claimed, (3) post-editing in progress, and (4) completed.

For post-editing the translated documents are automatically divided into smaller chunks based on delimiters (line breaks) in the source document. For each chunk the source text is shown above the editable machine translation. When the post-editor has finished editing a block it will be saved and marked in green. Compared to an interface where the post-editing is done in a single text area this design makes it easier for post-editors to resume post-editing because all of the previously completed blocks are clearly marked. Furthermore, this design makes it easier for multiple users to contribute to the post-editing of a single document because they can clearly see which sentences still need post-editing and which ones have been completed. A screenshot of the post-editing interface for a document being translated into Spanish is shown in Figure 1. The status bar at the top of the screen is used to track a document's progress through the pipeline. Finally the completed, post-edited documents can be downloaded.

Users can utilize the system to upload and post-edit documents created within their own health departments; they can also volunteer to post-edit documents that have been uploaded by other departments. This allows both documents and linguistic expertise to be shared across different (possibly geographically remote) health departments, which will eventually result in a more efficient utilization of resources. The interaction among different users is facilitated by a virtual discussion board included in the post-editing interface (visible in Figure 1). Users can post comments to others, e.g., in order to discuss the translation of a particular technical term.

Users can opt in to allow the source documents, original translations and post-edits to be collected for research purposes or to train customized MT models. Every document for which permission has been given is placed in an archive which can then be downloaded by the main system administrator. Thus, the system can simultaneously act as a data collection platform to collect new parallel source and target language corpora for training, updating or adapting MT models, or to create parallel corpora of machine translations and post-edits to train statistical post-editing models.

4.2 User Testing and Iterative Design

We conducted informal usability testing with six public health workers from four different programs at a state health department to obtain feedback from target users on our prototype system. Two of them were health educators, one was an environmental health investigator, one was a policy liason, one was a graphic designer, and one was an administrative staff member.

The facilitators presented the participants with an overview of the purpose of the system and quickly demonstrated how to upload and post-edit a provided example document. For user testing the participants were separated into two teams, each of which consisted of a project manager and a bilingual post-editor who spoke English and Spanish. Each team member was assigned a task that simulated how the system would be used in practice. The tasks assigned to each team member were con-

[1] http://www.microsofttranslator.com/dev

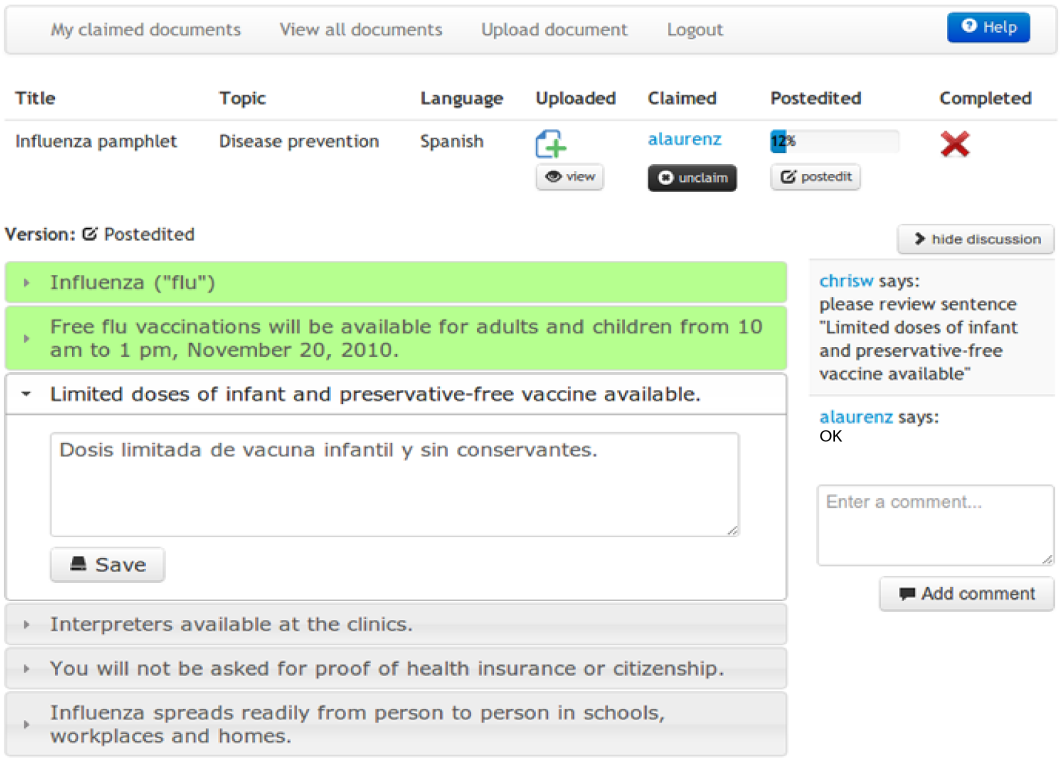

Figure 1: Screenshot of the post-editing page in our translation management system.

sistent with their roles in the translation processes used at their agency. The project manager on each team was assigned the task of uploading a provided example document. The post-editors were assigned the following tasks: (1) claim the document that was uploaded by the project manager, (2) post-edit the MT output, and (3) download the completed translation. All participants were instructed to verbalise their thought processes during the testing and were told they would not be provided with any help or guidance. During the testing the facilitators recorded general observations and critical incidences, such as difficulties encountered with the system.

After completing the assigned tasks participants were given a questionnaire to record their feedback after using the system. The questionnaire was broken up into four sections: (1) question specific to uploading a document, (2) questions specific to post-editing,(3) general questions, and (4) ranking and suggesting potential additional features. After the participants completed the questionnaires we held a focus group discussion with the participants to answer questions and record feedback that arose from the discussion. All participants successfully completed the assigned tasks without any help from the facilitators. Based on the questionnaire responses, focus group discussion, and facilitator observations all participants were interested in using the system in their translation work and found the system to be easy to understand and use. The responses to the question "What do you like most about the system?" included:
"Fast";
"Ease of use and visual appearance";
"It's intuitive - easy to use"; and
"Very efficient, clear - I like the post-editing page"
Only one participant gave feedback other than "None" to the prompt "Please describe anything that confused you while using the system." The reason was confusion between the "Available" icon signifying a document has not yet been claimed, and the actual "Claim" button.

Based on feedback from the participants several minor adjustments and additions were made to the look and feel of the interface. Other desired features included support for more document formats (currently only plain text files and MS Word formats are supported), which will be added in the near future. The most important changes, however, involved the addition of more features that support the typical interaction of different commu-

nity members with different roles. As one participant put it: "I miss the human factor." For example, post-editors indicated it was undesirable for a document to be automatically marked as completed after all sentences were saved. The reason is that they would like to let other staff members verify their translations and obtain feedback before finalizing a translation. In response we added a "Mark document as completed" button that appears after all sentences are saved and that needs to be explicitly clicked before advancing the document in the pipeline. In response to the project manager's desire to communicate guidelines or notes to post-editors (e.g., desired reading level of the upload page where document-specific information can be the document, target audience, etc.) we added a field on communicated and can later be edited. Participants also expressed interest in automatic email alerts that are sent out to post-editors whose language expertise matches the desired target language of the uploaded document. Finally, they strongly advocated being able to assign ratings to, or "Like", particular post-edits, as these would over time help to identify reliable and trusted post-editors. In sum, these are features that mirror not only the typical workflow but also the professional hierarchy or social network that exists within their community.

5 Related Work

A variety of translation management and post-editing systems have been developed in the past. Most of them (e.g. SDL Trados[2], Wordfast[3], etc.) are commercial products aimed at language professionals, such as translators and language service providers. Their price is often prohibitive and they frequently require software installation on the user side. Other systems, such as MemSource Cloud [4], SmartMATE (Penkale and Way, 2012), or Wordbee[5], work in the cloud but may still be too expensive for non-professional users.

Among free or open-source systems, Google's Translator Toolkit[6] comes close to our requirements in that it allows collaborative post-editing and document sharing. On the other hand it lacks essential features for our intended use, such as incorporating meta-information about documents and post-editors. A web-based translation management system intended for lay users was described in (Federmann and Eisele, 2010). However, the system solely accepts translation requests and distributes them to several back-end translation engines; there is no functionality for post-editing, version control, or user communication. Pootle[7] is an online translation management system primarily aimed at software localization rather than document translation. Although it supports document sharing and collaboration, it requires software installation on the client side. It does support human translation and editing but does not have an integrated MT component and is thus not suitable for our purposes.

6 Conclusions

We have presented an initial feasibility study and user testing with an integrated post-editing and translation management system for delivering MT+PE technology to communities of non-professional MT users (public health professionals). Despite focusing on a single user community we believe that the insights gained from these studies apply to other, similar communities of lay users.

First, lay communities often have severely limited financial and staff resources. Software tools that support machine translation, PE, and translation management should be easily available, low-cost (ideally free), and should not require software installation and management on the client side. The user interface should be intuitive, immediately usable and should not require extensive user training.

Second, users are likely to be domain experts rather than language professionals, and they tend to work on translation on an intermittent basis in addition to other tasks. Translation tasks may be shared among different users or among different geographically distributed groups. A PE system should enable users to save and archive partially completed work, hand partial work over to other users for completion or quality control, and it should support version control.

Third, the accuracy and appropriateness of the

[2]http://www.trados.com
[3]http://www.wordfast.net
[4]http://www.memsource.com/translation-cloud
[5]http://www.wordbee.com
[6]http://www.google.translate/toolkit
[7]http://www.pootle.org

translations for particular cultural or demographic groups are often more important in scenarios such as the one described here compared to other domains. Software tools need to support meta-annotation of documents with constraints on e.g, target user group, desired reading level, etc.

Fourth, users normally do not work in isolation but in teams who place much emphasis on personal communication in their work. They often rely on the judgment of trusted users or those with a high status in their community. These intra-community social networks should be replicated in a software application by facilitating user communication via email, virtual discussion boards, etc., and by allowing users to assign ratings to specific post-editors. At the same time users would like all communication, contact information to remain private (i.e. within the organization), and they do not wish to use their personal social networking sites for work-related purposes.

The software system described above was designed to fulfill those needs; it will shortly be released as an open-source package. We hope it will be of use to other communities, e.g. non-profit organizations providing legal assistance to LEP communities. Obvious additional elements that could be integrated are a management module for domain-specific terminology lists, or translation memories. These may be added in a future version of our system.

7 Acknowledgements

This study was funded by a University of Washington Mary Gates Research Award to the first author and grant #1R01LM010811-01 from the National Library of Medicine (NLM). Its content is the sole responsibility of the authors and does not necessarily represent the view of the NLM.

References

ACS. 2011. American community survey. http://factfinder2.census.gov/, downloaded June 4, 2011.

DHHS. 2003. Department of health and human services-guidance to federal financial assistance recipients regarding title VI prohibition against national origin discrimination affecting limited english proficient persons. *Federal Register*, 68(153):47311–47323.

Federmann, C. and A. Eisele. 2010. MT server land: An open source MT architecture. *Prague Bulletin of Mathematical Linguistics*, 94:57–66.

Glaser, B.G. 1965. The constant comparative method of qualitative analysis. *Social Problems*, 12(4):436–445.

Goel, M.S, C.C. Wee, E.P. McCarthy, R.B. Davis, Q. Ngo-Metzger, and R.S. Philips. 2003. Racial and ethnic disparities in cancer screening: the importance of foreign birth as a barrier to care. *J Gen Intern Med, 18(2)*, pages 1028–1035.

Green, S., J. Heer, and C. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of ACM Human Factors in Computing Systems (CHI)*, pages 439–448.

Guerberof, A. 2009. Productivity and quality in MT post-editing. In *Proceedings of MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators MT*, page 8pp, Ottawa, Ontario, Canada.

Jacobs, E.A., D.S. Shepard, J.A. Suaya, and E.L. Stone. 2004. Overcoming language barriers in health care: Costs and benefits of interpreter services. *Research and Practice*, 94(5):866–869.

Koponen, M., L. Ramos, W. Aziz, and L. Specia. 2012. Post-editing time as a measure of cognitive effort. In *Proceedings of the AMTA Workshop on Postediting Technology and Practice*, pages 11–20.

O'Brien, S. 2006. Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures*, 7(1):1–21.

Penkale, S. and A. Way. 2012. SmartMATE: An online end-to-end MT post-editing framework. In *Proceedings of the AMTA Workshop on Postediting Technology and Practice*, pages 51–59.

Plitt, M. and F. Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. 93:7–10.

Ponce, N.A., R.D. Hays, and W.E. Cunningham. 2006. Linguistic disparities in health care access and health status among older adults. *Journal of General Internal Medicine*, 21:786–791.

Vicente, K.J. 1999. *Cognitive work analysis: Towards safe, productive, and healthy computer-based work*. Lawrence Erlbaum Associates, Mahwah, NJ.