Using the Microsoft Translator Hub at The Church of Jesus Christ of Latter-day Saints

Stephen D. Richardson
Publishing Services Department
50 North Temple St.
Salt Lake City, UT 84150
stephen.richardson@ldschurch.org

Abstract

The Church of Jesus Christ of Latter-day Saints undertook an extensive effort at the beginning of this year to deploy machine translation (MT) in the translation workflow for the content on its principal website, www.lds.org. The objective of this effort is to reduce by at least 50% the time required by human translators to translate English content into nine other languages and publish it on this site. This paper documents the experience to date, including selection of the MT system, preparation and use of data to customize the system, initial deployment of the system in the Church's translation workflow, post-editing training for translators, the resulting productivity improvements, and plans for future deployments.

1 Introduction

The Church of Jesus Christ of Latter-day Saints (hereafter "the Church") has an extensive in-house translation effort to support communication among its more than 14 million members in 168 countries worldwide. This effort includes the translation of Church materials from English into over 100 languages. Besides scriptural texts in these languages, the Church publishes a magazine to its members in over 20 languages on a monthly basis and less frequently in an additional 30 languages. Several websites are hosted by the Church, the principal of which is www.lds.org, with content updated weekly in nine languages, and additional more static content in dozens more languages. The Church's semi-annual general conference of five 2-hour sessions is broadcast with live interpretation in 94 languages via TV, satellite, and internet, and hundreds of other events are broadcast throughout the year in dozens of languages. The translated text of many of these events is later published in the Church's magazine and on the internet. Materials are translated in diverse subject domains, including scripture, music, education, family history, humanitarian aid, welfare, legal, medical, finance, agriculture, real estate, technical, construction, manufacturing, facilities management, emergency response, and human relations. In 2011, the Church's translation department translated over 85 million words of (principally English) source text, generating many times that amount in the various target languages.

In its desire to increase significantly the availability of translated materials, especially those published on the internet, the Church launched an effort to deploy machine translation earlier this year. The initial objective of this effort is to reduce by at least 50% the time required by human translators to translate English content published on www.lds.org into nine other languages. If successful, the effort will then be expanded to several more languages and several additional sources and types of content. Machine translation into English and between other languages is also anticipated.

The following sections will discuss the selection of the MT system used in this effort, the preparation and use of training data to customize the system, the initial deployment of the system in the Church's translation workflow, certain aspects

of the post-editing training provided to translators, the productivity improvements observed over a period of months, and plans for future deployments.

2 Selection of the Microsoft Translator Hub

Several machine translation systems were evaluated initially as candidates for integration into the Church's workflow, taking into account such factors as output quality, customizability, performance, and overall cost, including both for initial deployment and for ongoing maintenance.

It is interesting to note that over the past year or more, several versions of statistical machine translation (SMT) systems or statistical/rule-based hybrid systems that provide for rapid customization using clients' bilingual data (such as that found in TMs) have become commercially available. These include (but are not limited to) Systran, SDL/Be Global, ProMT, Applied Language Solutions, and Microsoft, as well as the free, open-source Moses system. Each of these offers higher quality through rapid customization that was only dreamt of a decade ago.

Although the generic version of the Microsoft Translator has been available on the internet for five years (Richardson 2007), the beta version of the new Microsoft Translator Hub was just launched in February of this year, barely in time to be included in our evaluation. Even though certain problematic issues arose, as happens with any beta system, prompt attention from the Microsoft team resolved them sufficiently, and the Hub was selected.

In evaluating the output quality of certain systems, one or two languages were selected, system training was performed using a subset of our TM data, and a relatively small sample of test sentences was translated and evaluated by humans. While this was not a comprehensive quality evaluation by any means, we consistently observed that the Hub quality was comparable to, or in some cases slightly better than, some of the other systems we evaluated. In general, we observed that among the customizable MT systems, quality could be somewhat better in one or two specific languages, but across the board, it was hard to justify the selection of a system based on this criterion alone. Where quality was comparable, other pragmatic factors played a more decisive role.

One of the most distinguishing factors of the Hub was its cost. Microsoft provides subscriptions to its Translator service, including any trained systems in the Hub, at the rate of US$10 per million characters translated per month, with discounts at higher volumes, e.g., US$9 per million characters for 64 million characters per month and less than US$8 per million characters for one billion characters per month; (see: datamarket.azure.com). By comparison, Google Translate (which does not currently offer a service to customize its system with client data) costs US$20 per million characters, but it is billed only for characters actually translated, and is not a monthly subscription. Thus, if one translates less than half the amount subscribed to in the Microsoft monthly model, Google would be more cost effective, but again, it is unfortunately not customizable. Other systems evaluated were significantly more expensive than either Microsoft or Google to train, deploy, use, and maintain.

Included in the Hub cost is the capacity to train and deploy a (theoretically) unlimited number of language pairs and domains under the single price-per-character-translated umbrella, an easy-to-use customization capability using TM or other data (including data in various formats that is parallel but not yet aligned) an API that combines the trained MT systems with access to a TM in the cloud that contains our data, and an integrated set of tools for community review and correction of MT output. Microsoft also scored well on other factors generally considered with cloud services: system performance, availability, reliability, etc.

Having justified our selection of the Hub, it must be stated that we will continue to evaluate other systems – their quality, cost, and other factors – especially as we expand to other languages. In this regard, we are refining and better organizing our data for evaluations, and we have recently implemented a web-based community evaluation tool at the Church's volunteer crowd-sourcing site: vineyard.lds.org (see Figure 1).

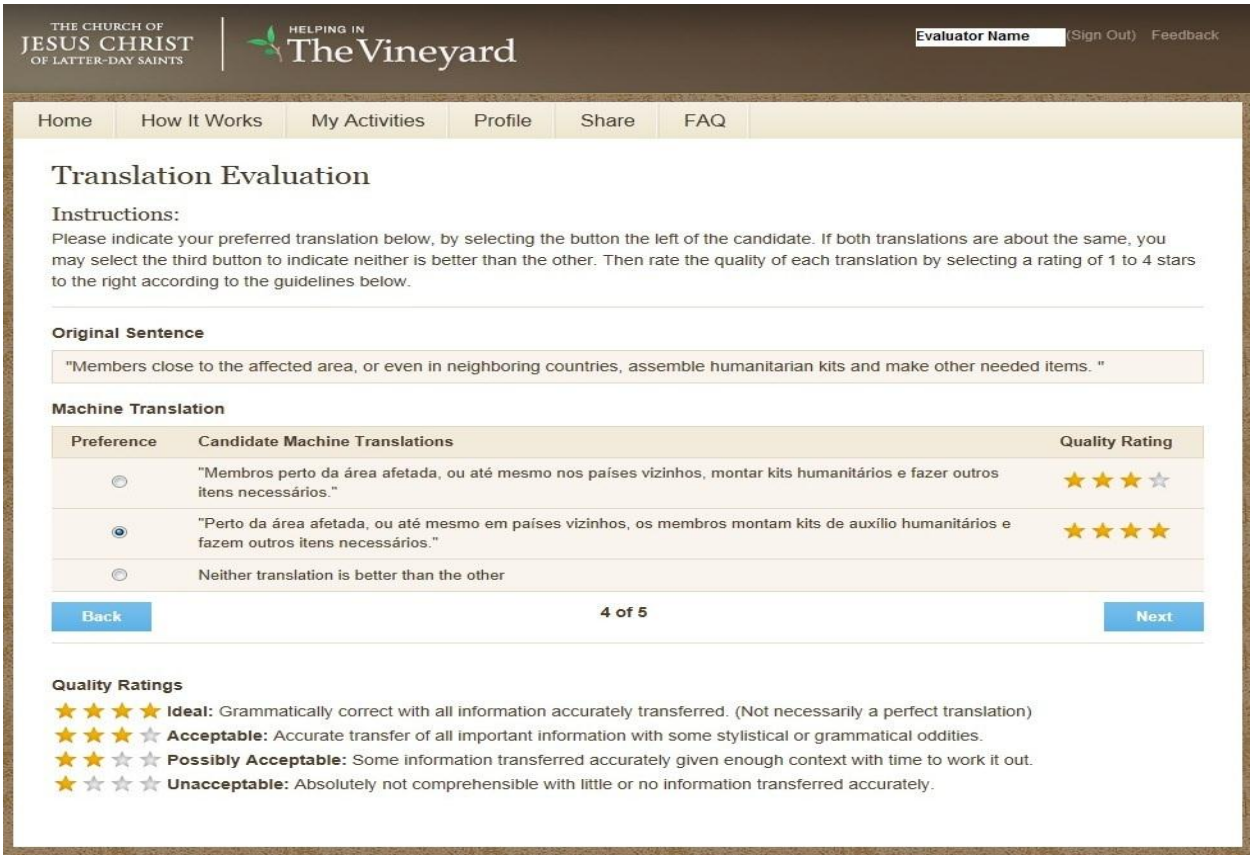

Figure 1: Community MT evaluation site

This tool employs the same evaluation methodology that has been used at Microsoft for human evaluations over more than a decade.

For a set of about two hundred source sentences, evaluators are asked to indicate a preference between two machine translations of each of those source sentences, and are also asked to assign a quality rating of from 1 (unacceptable) to 4 (Ideal) stars. Coughlin (2003) has pointed out that BLEU scores correlate highly with the preference task in this type of evaluation, as long as important factors are held constant. However, when dealing with different system types (pure SMT vs. Hybrid) trained on possibly different data, and to assess when a system reaches a quality level that will be acceptable for post-editing, this type of human evaluation has proven extremely useful.

3 Preparation and Use of Training Data

Although the Church has been translating materials in dozens of languages for decades, it has only been vigilant about storing and maintaining aligned data in TMs since around 2008. With the current strong emphasis in the translation industry on reuse, and now with the additional motivation of training MT systems, the Church has undertaken an effort to gather, clean, align, and store data from previous years.

The target languages for the MT effort this year are: Spanish, Portuguese, French, German, Italian, Russian, Chinese (Traditional), Japanese, and Korean. The raw data in the Church's TMs for these languages range from just over 600K translation units (TUs – aligned pairs of source/target segments) to around 1.2M TUs, with Spanish and Portuguese having the greatest number.

After cleaning and organizing this data, initial versions of trained systems have been created on the Microsoft Translator Hub for these languages with around 200K to 300K TUs each, depending on the language.

It is not the purpose of this paper to outline in detail the process through which system training takes place. That will be left to the Microsoft MT

group to describe. Briefly, however, after the data is cleaned, prepared, and exported to TMX files, it is uploaded to the Hub, a few options are chosen, such as whether to use the Microsoft models in addition to those created from our data and whether to generate random tuning and test sets from the data or to use sets we designate, and the system training is launched. A few hours later, notification is given by the Hub that the trained system has been created and the test set has been run against it. One can then examine the test set translations on the Hub or download them, review the BLEU score as compared to the score for Microsoft's generic Translator on the same test set, and invite members of a community to review and edit translations in the test set. The option exists to then perform another training with the same or different data and settings, or to deploy the trained system for production use through the Microsoft Translator API. Our experience is that the Hub dashboard that controls these features is well designed and easy to use (See Figure 2).

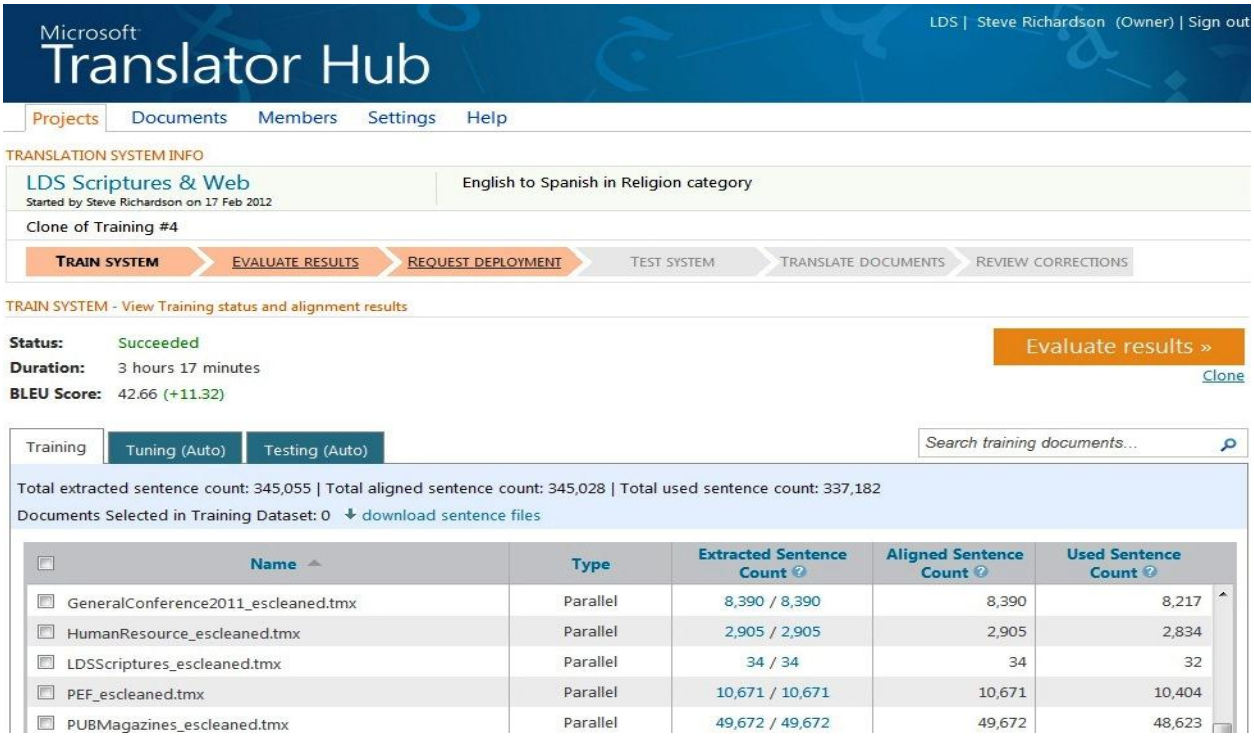

Figure 2: Microsoft Translator Hub dashboard for Spanish system

A few words should be added here about data selection and cleaning. Regarding organization and use of data by subdomain, we have chosen at this time to break out only the data in the family history (genealogy) subdomain, which is rather extensive. The remaining data is used as a single large "general Church" domain, since the topics covered on www.lds.org are fairly broad. Our strategy is to break up this general domain into subdomains only as experimental results suggest significant benefit in so doing.

More relevant, perhaps, is the following description of the main steps in our data-cleaning pipeline (not in any particular order), which is oriented toward the cleaning of TUs in a TMX file (another excellent source for data-cleaning guidelines may be found in TAUS 2009):

1. Verify consistency of character encoding throughout the data in the TUs.
2. Remove TUs where Src = Tgt, Src or Tgt = blank, Src or Tgt contain a high % of non-alphabetics.
3. Check for a consistent ratio of Src/Tgt lengths and eliminate outliers.
4. Segment and align sentences in TUs containing paragraphs if possible, and eliminate excessively long TUs that can't

be segmented into shorter aligned sentences.

5. Replace all entities with the strings they represent if possible, or with other generic strings.
6. Remove all tags/placeholders.
7. Ensure consistent use of terminology, capitalization, etc. (harder when training data spans many years and may require some manual review).
8. Eliminate duplicate TUs (where both Src and Tgt are duplicate). Note: in some systems, leaving or even introducing duplicates may increase the probability of correct terminology. Duplicates are not eliminated during the Hub training process.
9. Normalize quotes (e.g., substitute straight quotes for smart quotes or guillemets) and other punctuation. If this is done, post-processing of MT output may be necessary for specific languages to ensure correct punctuation usage.

The above steps have provided reasonably clean data for the training of our systems on the Hub. We have also observed that it is not necessary to remove every piece of "dirty" data, especially if it is fairly uncommon in the training set, since statistical processing will naturally filter out infrequent phenomena.

Table 1 below provides the results of cleaning and using the data in our TMs to create trained systems for English to Spanish and Portuguese on the Hub. The 10 to 11 point difference between the BLEU score from our trained systems and the BLEU score obtained by the generic Microsoft Translator on our test set is an indicator of the strong positive effect of our training data on output quality.

Target Language	Segments used to train	BLEU (Trained systems)	BLEU (Generic MS translator)
Spanish	337K	42.66	31.34
Portuguese	287K	42.46	32.80

Table 2: Training data and BLEU scores for Spanish and Portuguese Systems

4 Deployment of Trained Systems in the Church's Translation Workflow

To date, we have deployed trained MT systems in six languages: Spanish and Portuguese are already in production use, and French, Italian, German, and Russian have just come online. Three more systems - Japanese, Chinese, and Korean - will be deployed by the end of the third quarter. Additionally, a request has just been made to add four more languages that were not part of the original plan by the end of 2012: Samoan, Tongan, Cebuano, and Tagalog. Data is now being gathered and prepared to train these systems.

Feedback thus far from Spanish and Portuguese translators after 3 months of use indicates a very high level of satisfaction. Initial French and Italian impressions are also positive. German, Russian, and the three Asian languages will undoubtedly be more challenging, but we will continue to address issues, retrain systems, and re-evaluate alternatives as necessary.

The Church uses SDL/WorldServer to store and manage its TMs and to produce packages of documents that are sent to area offices around the world for translation. These packages include a bilingual representation of the documents, which is pre-populated with exact and fuzzy (>=80%) matches from the TMs stored in WorldServer as well as with output from our trained MT systems for any segments that are not matched in the TMs. An MT connector has been implemented in WorldServer that calls the Microsoft Translator API, specifying the identifier required for our trained systems.

On the receiving end of these packages, translators use workbenches such as SDL/Studio or WordFast to review and edit the TM matches and to post-edit the MT output. Lingotek has been used for volunteer community-sourced projects.

As translators have begun to post-edit the MT output and we have received their feedback, we have passed it on to the Microsoft team, but we have also implemented the following pre- and post-processing steps as a stop-gap measure to correct errors in the output that are superficial but nevertheless important to achieve higher productivity during post-editing:

1. Normalize smart quotes and guillemets to straight quotes before translation and restore them afterwards.

2. Adjust the relative placement of quotes, commas, and terminal punctuation according to the rules of the target language.
3. Remove extras spaces generated by the Microsoft Translator around tags and punctuation.
4. Add or delete tags and punctuation at the beginning and end of segments to correspond to the source sentence
5. Capitalize certain terms that the trained MT systems do not produce correctly

The above processing really does make a significant difference to our translators as they post-edit because of the many quotations, tags, and capitalized terms that occur in our materials. Although not that common in technical texts, and generally difficult for MT systems to process correctly, we anticipate that the Microsoft team will eventually incorporate these items into the generic Translator or cause them to be learned during the training process on the Hub so that all kinds of texts may be handled correctly.

5 Post-editing Training for Translators

It has been widely observed that experienced translators who have never post-edited MT output before are often quite negative about it and generally have a difficult time adapting to it. The Church's translators, as good as they are, are no exception.

Fortunately, however, the Church's leadership, and specifically the leadership of the Translation department, understanding that the goals of substantially increased volumes would never be achieved by traditional means, began working to change attitudes regarding the transition to MT post-editing well over a year ago.

As an essential part of the effort to deploy MT, a special training module was developed for the Church's translators who would participate in post-editing. This training laid out clearly the guidelines for the "moderate" level of post-editing to be applied to the content on www.lds.org, but rather than presenting the guidelines and heavy-handedly demanding adherence to them, a simple game-like activity called "Which is which?" was implemented.

In this activity, 5 or 10 English sentences are presented to the translators together with two translations in a random order: one is a human translation from scratch (created by a translator who didn't know that his translation would be used in this way) and the other is a post-edited machine translation (See Figure 3). The translators are then asked to guess which one is which and note it down silently on a piece of paper. No consultation with others is allowed.

After all the sentences and their translations have been presented, they are presented again, and the translators in the room are asked by a raise of hands which one is which. Typically, and especially if there are several translators in the room, some will pick one and some will pick the other. The true answer is then given (see Figure 4), and often, there are few chuckles and some discussion about why someone thought it was one or the other. The raw MT output is then displayed next to the post-edited translation, with the changes that were made in order to correct it (see Figure 5). It's important to pick a range of examples from those requiring no edits, to those with some edits, to those with many edits, so that expectations are properly set.

In all the cases that this activity has been used (with translators for 6 different languages so far), the translators generally came away with the attitude "that's not so bad – I can do that." They were then ready to see the post-editing guidelines, understand them, and accept them.

- This is the quiet, encouraging voice which sustains without pause those who walk in faith down to the last days of their lives.
- É a voz suave e motivadora que ampara constantemente aqueles que caminham pela fé, durante todos os dias de sua vida.
- Esta é a voz tranquila e incentivadora que apoia sem pausa aqueles que caminham pela fé até os últimos dias de suas vidas.

Figure 3: English sentence, human translation, and post-edited machine translation

- This is the quiet, encouraging voice which sustains without pause those who walk in faith down to the last days of their lives.
- É a voz suave e motivadora que ampara constantemente aqueles que caminham pela fé, durante todos os dias de sua vida.
- PE MT Esta é a voz tranquila e incentivadora que apoia sem pausa aqueles que caminham pela fé até os últimos dias de suas vidas.

Figure 4: Post-edited machine translation identified

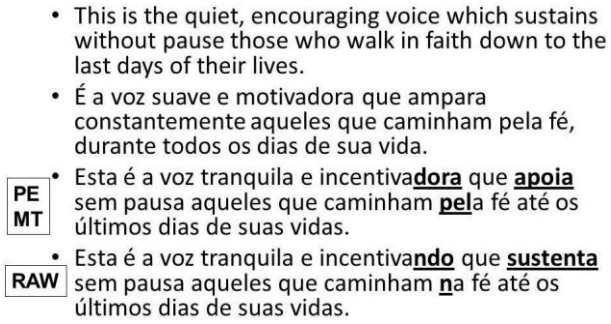

Figure 5: Raw MT displayed together with changes made during post-editing

We do not employ MT with post-editing for scriptural or doctrinal content. Rather, we use it for much of the more commonplace communications, news, stories, articles, and instructional material found on www.lds.org.

Hence, the guidelines for these materials target a "moderate" level of post-editing as defined by the following points:

1. Change only what is essential to ensure clear understanding and grammatical correctness.
2. Correct spelling, capitalization, and punctuation.
3. Ensure that any tags are present and in the correct positions
4. Do not use synonyms to make the translation more original, interesting, or stylistically pleasing.
5. Style does not matter as much as accuracy and adherence to the original text.
6. If an improvement is not immediately obvious, move to the next segment (avoid the temptation to make the translation sound the way you might say it instead of how someone else might legitimately, but less stylistically, say it).
7. Throughput expectations: high
8. Quality expectations: medium

6 Productivity Improvements

By following the post-editing guidelines given in the preceding section, and with ongoing practice, the Church's Spanish and Portuguese translators have achieved significant speed-ups in their work.

During 2011, the average time spent to translate one page of text (286 words) was approximately one hour. Following the actual translation, various reviews were performed: a content review (for accuracy), a language review (for fluency), and a proofing review (for formatting/publishing), adding up to another .9 hours of work. The goal for the MT effort was to reduce the translation time by 50%. Together with a targeted reduction in the review and publishing process, the overall goal was to translate and publish one page per hour.

Over the past three months, translators have been recording the time it has taken to translate the projects assigned to them as well as the time to accomplish other review, publishing, and administrative tasks. Table 2 below shows the translation (post-editing) times for a selection of the recent projects translated by two Spanish translators and one Portuguese translator. The projects were selected based on having a minimal number of TM matches in the material translated – projects with high TM matches were eliminated so that the numbers would not be skewed towards faster times.

According to these numbers, the goal of a 50% reduction in translation time, or translating two pages per hour, has clearly been reached and even surpassed. Depending on the difficulty of the texts, the translators are sometimes able to achieve rates as high as 3-4 pages an hour. These are offset, of course, by other more difficult material.

Translator	# of projects	# of pages	hours spent	hours per page
1	7	48.70	18.88	0.39
2	6	94.70	45.60	0.48
3	12	97.90	43.73	0.45

Table 2: Post-editing productivity of translators

An interesting comment heard from a number of translators, which they said they did not previously expect to make, is that it is easier to make a few changes when the (reasonably correct) words are already present than it is to have to think from scratch about how to translate something.

Another indicator of a significant shift in attitude is that all the Spanish and Portuguese translators who have been involved in post-edited have requested that MT output be included in

many of the projects that they are assigned to translate other than those for www.lds.org.

While Spanish and Portuguese adoption appears to be highly successful, there will undoubtedly be more substantial tests of our MT effort as we deploy other more difficult languages into production.

7 Future Plans

Beyond the nine languages, and the four more that have just been added for 2012, the plan for 2013 is to deploy these 15 languages: Polish, Dutch, Norwegian, Hungarian, Armenian, Finnish, Malagasy, Danish, Thai, Croatian, Czech, Fijian, Ukrainian, Swedish, and Swahili. Armenian, Malagasy, Croatian, Fijian, and Swahili are currently not available on the generic Microsoft Translator site, but with sufficient data, which we continue to generate, we should still be able to create systems in these languages. Iterative re-training and careful evaluation will determine whether these languages will actually be deployed into production.

We anticipate that the Church's initiative to gather, clean, and organize bilingual data from past years will continue to result in noticeable improvements in MT quality.

Acknowledgments

I gratefully acknowledge the assistance of, and contributions to this effort by Dan Higinbotham and Jerry McGhee of the LDS Church's machine translation team, as well as by Rahul Sharma and Chris Wendt of the Machine Translation group at Microsoft Research.

References

Deborah Coughlin. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In the *Proceedings of MT Summit IX*, September, 2003, New Orleans, Louisiana.

Stephen D. Richardson. 2007. Microsoft machine translation: from research to real user. Invited talk at *MT Summit XI*, September 2007, Copenhagen, Denmark.

TAUS (Translation Automation User Society). 2009. Technical guide to SMT training data. TAUS Report located at www.translationautomation.com.