Translation Tools
Status, Practice, and Gaps

Jennifer DeCamp
MITRE Corporation
7515 Colshire Drive
McLean, VA 22102 USA
jdecamp@mitre.org

Abstract

This paper describes the state of the practice for Human Translation (HT), the established tools, the research, and the capability gaps. The paper is a summary of the tutorial at the Association for Machine Translation of the Americas 2009 conference.

1 Introduction

At the 2008 conference of the Association for Machine Translation of the Americas (AMTA), Mark Tapling, President and CEO of LanguageWeaver, commented in a keynote address (2008) that —Our market tends to promote science; as opposed to the solution value.” Developers and researchers throughout the audience demanded —Give us the user problems!” This paper is an effort from a user of MT and a consultant to users of MT on user-oriented problems.

2 Fundamental Research Questions

Fundamental research questions in translation include:

1. How can you get better accuracy in MT and HT, or as Tapling (2008) phrased it, better —communicative value”?
2. How can you make a C-level translator into a B-level translator? Kay was one of the first people to raise this issue in the early 1980s (Kay 1997).
3. How can you make human translators more productive?

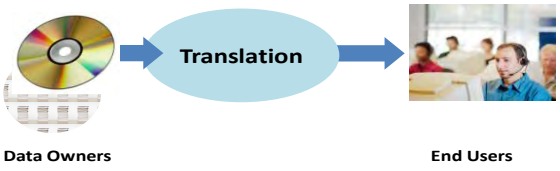
3 Types of Translation

There are many types of translation, including high-quality translation for literature, marketing materials, etc.; gists and summaries (Taylor and White, 1998; Egan 2008); obtaining answers to specific questions; and sorting (i.e., figuring out the language, subject and needed language proficiency level in order to route the material). There is translation of text, email, television news broadcasts (Egan 2008) and other media. There is MT embedded in chat and search tools. There is speech-to-speech translation, the focus of a Defense Advanced Research Program Agency (DARPA) project called Spoken Language Communication and Translation System for Tactical Use (TRANSTAC).

Some of the major distinctions in translation include publisher-centric vs. user-centric; human translation vs. machine translation vs. mixed translation; and caution vs. find-anything. These distinctions are described below.

The term —HT” is used to describe translation done by humans, since the more established terms —Machine Assisted Computer Translation” and —Computer Assisted Machine Translation” were set up to originally include word-processing, and just about every translator uses word processing (Hutchins 2001).

3.1 Publisher-Centric HT



In publisher-centric applications, materials such as foreign language editions of newspapers or product documentation, are prepared in one language and

then sent to a Language Service Provider (LSP) either inhouse or outsourced.

The LSP translates the material by HT, Machine Translation (MT), or some combination of the two.

3.2 Publisher-Centric HT

Translators use a wide range of tools, including electronic dictionaries, Translation Memory, Terminology Management Systems, and workflow management tools.

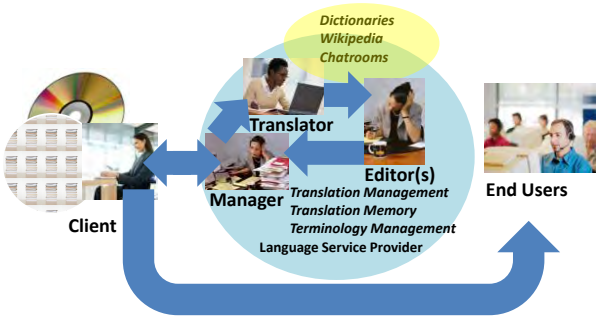


Figure 2: Publisher-Centric Human Translation.

Larger translation groups may have a manager who reviews the document to then route it to a human translator with the appropriate language proficiency and technical knowledge. Such groups may also have editors who review the translated documents for accuracy. The editors often work with the clients (e.g., with the software companies) to determine appropriate terminology, formats, styles, etc.

Commonly used tools include Translation Memory, such as Across, Bee-Text, Multicorpora, SDL Trados, WordFast, and LingoTek. Open Source tools include GlobalSight and Omega-T. Translation Memory (TM) is based on MT research (Hutchins 2001) for aligning source and translated texts. The system presents past translations to the translator, thus saving the translator the time of researching this information and further helping to standardize the translations. The translator can select to receive only exact matches, or can accept partial or fuzzy matches.

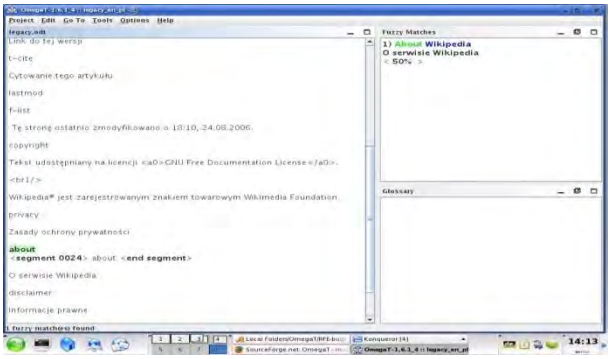


Figure 3: GlobalSight Open Source TM.

In some cases (e.g., TRADOS), companies can extract only the text where there are no precedent translations and send just that text to their human translators. A frequent complaint of some translators is that so little context is provided with this approach that accurate translation becomes very difficult.

There are frequently problems with alignment of text, where terms may become split and thus nonsensical. LanguageWeaver provides Translation Customizer, a tool designed for MT that enables experts to realign translation memories and then have the improved translation pairs take precedence over other statistical MT input.

Currently, most human translators work with only a small set of source-translation pairs. They also tend to catch most alignment issues when reviewing entries in Terminology Management Systems. However, the tool may have applications for HT in environments with large quantities of Translation Memory that need to be enhanced and prioritized.

Terminology Management Systems, such as MultiTerm and Terminotix, provide tools for handling terminology, where the terms are often drawn from Translation Memory. The terms may also be drawn from other research, including from surveying foreign sales offices. Terms may be provided across a range of languages (e.g., what the same machine part is to be called in English, French, German, Spanish, and Italian). The approach is prescriptive rather than descriptive: the intent is to provide standardization, consistency, and clarity.

The resulting terminologies document not only definitions but also the details when there is no clear correspondence in terms across languages. The terms traditionally are presented by subject

matter, but with standards such as the International Organization for Standardization Lexical Markup Framework, the terms can easily be viewed in a number of formats.

Senior translators—and more often, trained terminologists—communicate with the authors of the text being translated, system developers, Subject Matter Experts, end users, foreign marketing offices, and other language experts to determine the appropriate translations. The terminologists also work with multiple transliteration standards and with conventions for abbreviations and acronyms, often needing to go back to the original native-script full terms to be able to render the term correctly in a new standard or convention. They also work to disambiguate terms.

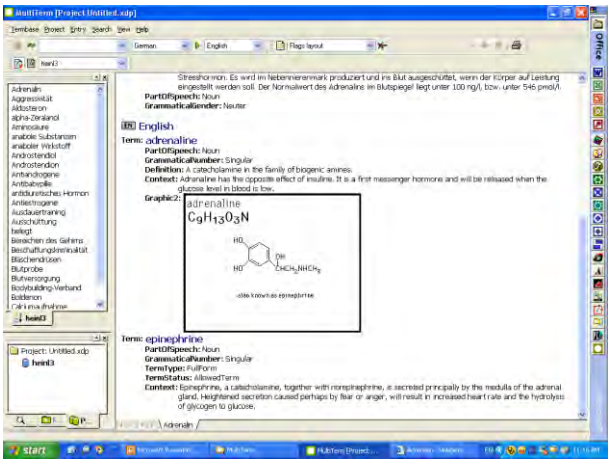


Figure 4: MultiTerm.

Translators consult other resources, such as internet dictionaries, chat rooms, and Wikipedia. Junior translators sometimes use online MT systems to look up terms—a practice that is quick and convenient but that does not provide a term within context to the MT system or provide context on the term back to the translator.

Translators and/or translation editors also practice Quality Assurance (QA). There are numerous tools that are used for QA in HT, as described by Makoushina (2008) in her evaluation of QA capabilities in Déjà Vu X, SDLX QA Check, Star Transit, Trados QA Checker, Wordfast, ErrorSpy, QA Distiller, and the Open Source XBench. Such Quality Assurance tools find untranslated segments, partial translations (where some source text was left), incomplete translations (significantly shorter than the source text), identical segments that are translated differently, differing segments that are translated the same, and segments with corrupt characters. They check for number values and formatting, untranslatables (terms that should not be translated), punctuation problems, and adherence to project glossaries. Such tools can also be used to alert translators to problems in MT output. In addition, they could be used to better inform the consumer of raw MT output.

3.3 Publisher-Centric MT-Assisted Human Translation

Translators are increasingly using MT, with pre-editing and/or post-editing. Pre-editing was pioneered by Xerox Corporation in the 1980s, which used software to check English documentation for types of text (e.g., long and/or convoluted sentences; words not in the MT system) that might cause problems with the MT output (Ryan 2003). Xerox claimed that these tools improved the readability and clarity of the English documents as well as of the machine translated material.

A newer take is work by Bernth and Gdaniec (2001) on Translatability Ratings, where they have identified problems in source material and communicate those issues to the authors. There is also various authoring software such as MaxIT, Acrolynx, and AuthorIT that might be used in this context.

Post-editing was also pioneered by Xerox Corporation, where uncertain translations were were highlighted for translator attention. This highlighting is particularly useful since it enables translators to focus on problem areas and not necessarily to read the full source and translated texts, comparing line by line. Systran developed similar post-editing capabilities. The Pan American Health Organization developed their own post-editing system which is still in use today (Aymerich 2006). There are also experiments in conducting automated pre- and post-editing (Doyen et al. 2008).

A modern version of post-editing support is provided with LanguageWeaver's confidence ratings, shown in Figure 5. These ratings are based on degree of uncertainty, with the degree of purple hue indicating the degree of lack of confidence (Muslea 2009).

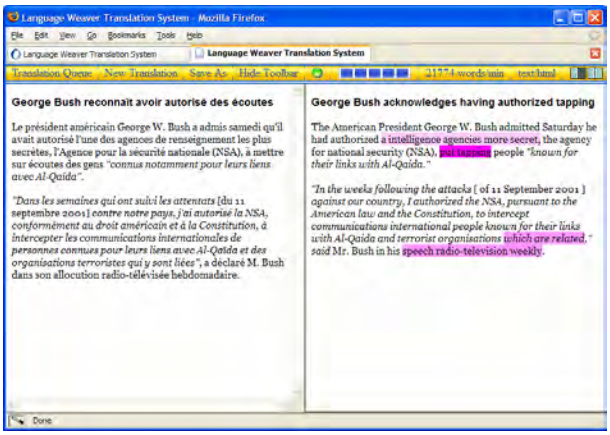


Figure 5: LanguageWeaver Confidence Ratings.

There are also various tools for supporting the translation process, including the European Commission (EC) Information Society Technologies (IST) effort with predictive MT known as TransType (Macklovitch 2004). TransType saves a translator keystrokes by predicting the completion of a word or phrase, similar to the function in an Excel spreadsheet. It also provides translators with alternatives drawn from Translation Memory, as shown in Figure 6.

Figure 6: TransType II.

There are beginning efforts to provide editing and spotchecking of human translation. For instance, TransCheck (also part of the IST effort) compares the source and target text to identify problems with omissions, numerical expressions, and source language interference (drawing on a negative dictionary). It also checks for the consistent use of terms.

Quality Assurance software is particularly useful in that there are not always the resources to provide thorough human editing. Providing feedback to the translator enables the translator to correct his/her own problems. Providing feedback to an editor alerts that person to possible problems with training needs, burnout, and other translator issues.

3.4 Publisher-Centric Globalization, Internationalization, and Localization

A large quantity of translation material is intended to provide documentation for products and services for markets using different languages. The general term for this substantial professional field is Globalization, Internationalization, and Localization (GIL). Translation Memory and Terminology Management Systems are commonly used. Accompanying or separate software is often used to extract text from and reinsert translated text to programming code or HTML for translation.

The Unicode Consortium and the Object Management Group (OMG) have been working on the Common Locale Data Repository. This internet set of libraries includes information by language and country of voltage requirements, plugs, character sets, font preferences, spelling conventions, and even color and image preferences. There may be many applications for this kind of resource in HT and MT.

3.5 Publisher-Centric MT

MT is sometimes used to produce predictable texts, such as photocopier or computer documentation. One of the advantages of MT is that it can provide translations very quickly once a source text is available, particularly if the system is primed with new terminology or parallel texts.

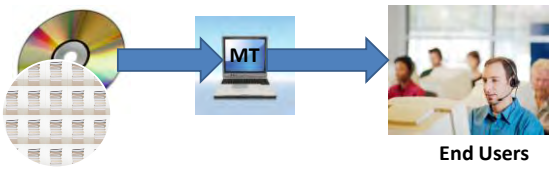


Figure 7: Publisher-Centric MT.

3.6 User-Centric MT and HT

In user-centric MT and HT, the user generally does not have the skills, time, or inclination to read the material in the original language but prefers to have the translation provided. An example of user-centric HT might be where a user selects certain research articles and sends them to an internal or external set of human translators to have them converted to English.

Figure 8: User-Centric MT.

User-centric MT is particularly appealing to companies providing products and information services, since they need only provide and maintain the products or services in a single language. The users then employ MT (such as Systran Translate) to obtain their own translations of the latest material. Since the users are selecting and employing a tool to provide the translations, the information provider presumably does not bear the legal responsibility for mistranslations.

Microsoft helped to pioneer user-centric MT for documentation by providing their MT system as a perk to VIP customers. The customers could then translate larger sections of the Microsoft website and/or help documentation. Microsoft is now making their MT capability more broadly available.

Given the cost and time of human translation, it is becoming increasingly popular among users to send electronic documents and other text to online-MT, particularly to one of the many free services now available on the internet (e.g., Bablefish, Altavista, Systransoft, Systranet, Google Translate, and Microsoft Translate). The problem is that users frequently have little or no understanding of the limitations of MT and little or no way to check the original text. As a result, the translations may deviate considerably from the original text, but the user might not realize this deviation.

A promising research and development area of MT is how to provide information and tools to users along with the MT so that the users can better understand the reliability of the MT output and can correct some of it themselves. LanguageWeaver's confidence ratings, and the output of several Post-Editing tools may be applicable. This area is discussed at length in a paper on "What is Missing in User-Centric MT" (DeCamp 2009).

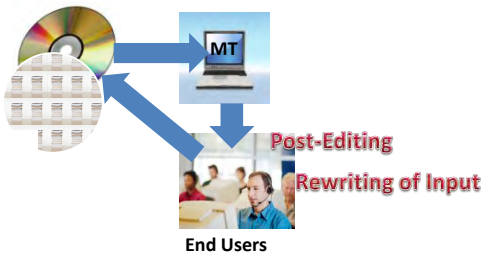


Figure 9: User-Centric MT with Review Tools.

4 Areas Where MT Can Help HT

There are many roles for MT in the HT environment, including to provide triage (routing), overviews, first cuts, guided or predictive typing, templates, additional information, and training.

4.1 Triage

Any translation bureau—commercial or government—needs to sort through incoming source material to determine the subject matter and level of technical difficulty and thus to assign the material to an appropriate translator. In government situations, where large quantities of material are received from a wide range of sources, it is particularly helpful to have tools such as MT (Bemish 2008). Adding entity tagging also helps in the assignment of documents (Day et al. 2006)

Determination of genre may also become important in providing templates to the translators, particularly intelligent translation templates (Kay 2001).

4.2 Overviews and Hypotheses

In a study by Day et al. (2007), translators—particularly those conducting gist translators—found the system helpful obtaining an overview of the material. Color-coded entity tagging was also helpful.

4.3 First Cut Translations

MT has long been used for providing first-cut translations that are then post-edited by human translators (Hutchins 2004). CACI's Language WorkBench provides translators with machine translation as one of many resources. An outstanding example of predictive typing is TransType II.

Another approach taken by Kay and Xerox PARC (2000) was to provide intelligent templates that would constrain the choices of the writer or translator.

4.4 First Cut Translations

MT can also be a reference resource for translators. In Day 2005, output from three MT systems was displayed on the translator's screen, with the assumption that the errors would probably be different across the systems and that the translator could then triangulate (Kay 2000) across the systems to get the meaning. The translator could get suggestions of vocabulary. He or she could get a hypothesis of what the text was about. He or she could also cut and paste selected text into the translation.

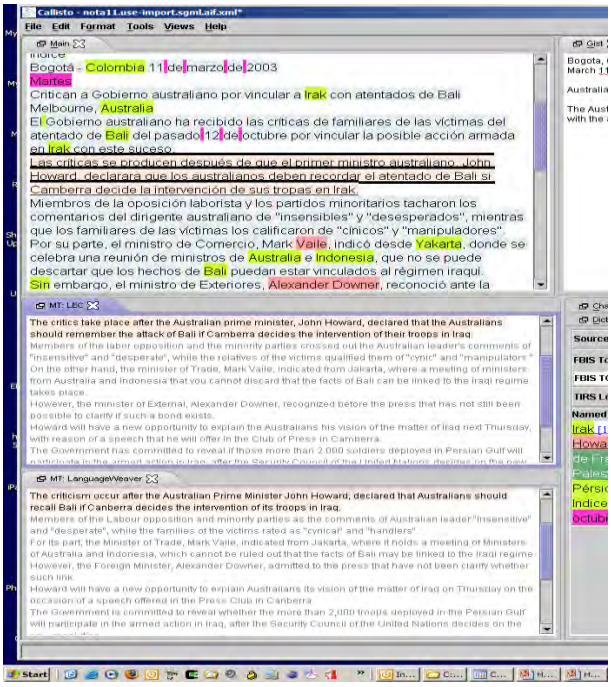


Figure 10: C-FLEX with Three MT Systems.

4.5 First Cut Translations

Providing editing and quality checks is very time intensive in an HT environment. Most post-editing tools can be used by editors. A few tools, such as TransCheck (Macklovitch 2008) automate checking for a few of the problems encountered by translators and editors, such as problems with omissions, numerical expressions, source language interference, and inconsistent use of terms.

4.6 Training Tools

MT can also be an effective training tool for translators. An example is provided by Egan (2008) using translation of news broadcasts.

4.7 Dictionaries and Term Harvesting

Traditional paper dictionaries were set up to consolidate information, such as by providing all forms of a word under one entry (e.g., "run", "runs", "running", "ran" are all provided under "to run"). Computer memory makes such constraints obsolete.

There is considerable testing and research to be done in how we can make the right word with the right conjugation appear in the user's translation with the least number of keystrokes. Translation Memory and tools such as TransSearch may be part of the solution. Wordnet—particularly bilingual or multilingual wordnets—may help translators to more easily distinguish appropriate meanings. Morphological analyzers may provide a bridge from traditional dictionaries to what the translators need in their documents.

One of the biggest problems in any translation is dealing with the large amount of new vocabulary. Using tools such as TransSearch (also from the IST project) saves translators time in finding bilingual concordances. Such concordances are useful in checking the context of the terms, researching alternative translations, and developing new terminologies and/or dictionaries. There is also the potential for greater use of entity extraction to build dictionaries.

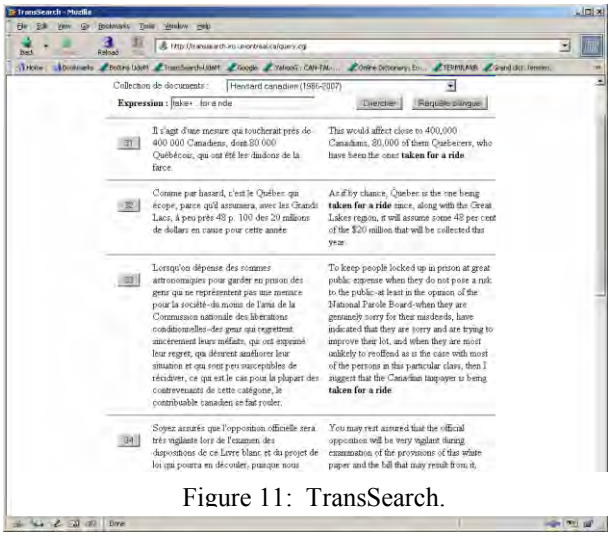


Figure 11: TransSearch.

4.8 Translation Memory

One of the key problems with TM systems is that terms become broken due to mistakes in the alignment of text. For instance, a term such as "White House" can be split into merely "white" and "house". Research by the MT community in this area can also be applied to translator tools.

5 Areas Where HT Can Help MT

MT has drawn on human translation and human translation technology in numerous ways, including with dictionaries, Translation Memory, and Transliteration.

5.1 Term Translations

Where terminologies and specialized vetted dictionaries exist, these materials can override statistical MT to provide a higher degree of accuracy, particularly for the translation of a specific customer. Statistical frequency may not always be the best method of determining the appropriate meaning or of standardizing terminology.

HT practices may also provide insights into how to deal with problems and ambiguities in translation. A common practice in human translation is to provide footnotes and inline references to further qualify the translation. Such notes are usually to provide further information about a term when there is not a clear equivalent in another language. These kinds of notes could be automatically inserted to signal areas of uncertainty and/or to provide further information about possible translation alternatives.

A major advantage of many HT tools is that they provide alternative translations for terms, with sufficient information for the user to make a selection. It may be interesting to experiment with MT tools that provide similar functionality, such as an "Again" button to try a dubious translation with terms having a less high statistical frequency. A user might be able to cycle through translations.

5.2 Translation Memory

Of course, one of the main ways that HT benefits MT is by providing high quality translations. Organizations are increasingly looking at ways to aggregate such translations. LingoTek, for instance, had a system where users of their Translation Memory system could elect to contribute their source and translated pairs of documents to a common Translation Memory database in return for the right to use the community database themselves.

5.3 Transliteration and Name Translation

Transliteration and name translation software was developed by Basis Technologies for use in HT. The software is now being successfully used with LanguageWeaver machine translation to provide higher accuracy of name transliteration and name translation. There is still much work to be done in developing transliteration systems and tools, particularly those that provide full backwards transliteration with no loss of data.

6 Coordination with HT

The main model used by MT has been for individual companies to maintain parallel corpora and/or lexicons separately, particularly since the quality of the parallel corpora and/or lexicons impact the accuracy and thus the marketability of the MT. However, from the perspective of the user—particularly from the perspective of a very large and diverse user of translation such as the United States Government—there is need for coordination among tools so that the same translation terms will be used regardless of the tool.

In the past ten years, there has been increasing sharing of corpora and lexicons. The Open Lexicon Interchange Format (OLIF) standard was developed to enable the exchange of data between different MT systems. A new standard, OLIF II is now available. Even so, the exchange of lexicons between companies is limited to a very few organizations, with no published studies as to the effectiveness of the standards. In addition, the exchange is only between MT systems and not between MT and HT systems.

The Lexical Markup Framework was completed by ISO Technical Committee 37 in 2009, and work continues to develop more specific guidelines for exchanging dictionaries among and between MT and terminology systems. This new standard has received considerable international support. Wittenberg and Romary at the Common Language Advanced Research Infrastructure (CLARINS) have developed an online tool for easily annotating data with LMF. The U.S. Government is a major

user of LMF and in fact, held a leadership position in its development.

There are many means for coordinating terminology. One means is to add the dictionaries to the MT systems, including to statistically-based systems. This kind of approach would be particularly productive in areas that practice extensive terminology management—i.e., where terms are researched, reviewed, and selected in order to have high-quality translation across a workgroup.

In addition, there is a need for addressing new terminology.

Another means is to share parallel corpora, so that SBMT and translator-based Translation Memory could use the same resources.

7 Adoption

A key issue with any tools is their adoption and use. Receptivity to and value of the tool may vary by demographics. For instance, Day et al. (2006) found differences in tool use in beginning vs. advanced translators, with beginning translators making greater use of the MT other features. Use may also vary by the tasks and objectives, and by the degree of training and exposure the translator or end user has to the tools. In addition, use may be affective by perceived helpfulness of tools. A tool with seemingly great potential can prove unhelpful or unacceptable for a simple reason such as that it is difficult to access from a translator or users' typing environment.

Obtaining funding to develop or acquire such tools is also sometimes hampered by difficulties in assessment. Human translation is affected by so many factors (e.g., time of day, number of translations already completed, etc.) that assessment data to date has not made a compelling case for funding (Day 2006).

8 Conclusions

The term ―translation" covers many tasks and requirements, as has been well established (White and Taylor 2006; Egan 2008; etc.), and different task and accuracy requirements may need different skills and tools. There are many promising areas where the MT/NLP community is improving and/or can improve HT, including through increasing translator and editor productivity. There are also many promising areas where the HT community can offer their tools, practices, and standards to increase the quality of translation, including through providing terminologies, footnoting, and annotation. Many of these tools for MT and for HT can also be provided to users of MT who currently tend to have little knowledge of the source texts or the machine translation.

Acknowledgements

I would like to thank Laurie Gerber and Nick Bemish for reviewing this paper and Carl Rubino for reviewing parts of it. I would like to thank the United States Defense Intelligence Agency Foreign Language Program Office for sponsoring my participation in the MT Summit.

References

Julia Aymerich and Hermes Camelo. 2006. Post-Editing of MT Output in a Production Setting. *Proceedings from the Association for Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications.* Cambridge, MA.

Nicholas Bemish. 2008. Can MT Really Help the Department of Defense? *Proceedings from the Association for Machine Translation in the Americas (AMTA 2008).* Cambridge, MA.

Arendse Bernth and Claudia Gdaniec. 2001. MTranslatability. *Machine Translation* Vol 16, 3, 175-218.

Kenneth Church and Eduard Hovy. 1991. Good Applications for Crummy Machine Translation. in J. Neal and S. Walter (eds.), *Natural Language Processing Systems Evaluation Workshop*. Rome Laboratory Report #RL-TR-91-362:147-157.

Jennifer DeCamp. 2009. ―What is Missing in User-Centric MT?" Proceedings of the MT Summit. Ottawa, Canada.

Jennifer Doyon, Kathryn B. Taylor, and John S. White. 1999. "Task-Based Evaluation for Machine Translation." *Proceedings of Machine Translation Summit VII '99*. Singapore.

Kathleen Egan. 2008. User-Centered Development and Implementation. *Proceedings from the Association for Machine Translation in the Americas (AMTA 2008).* Cambridge, MA.

Lauren Friedman and Stephanie Strassell. 2008. Identifying Common Challenges for Human and Machine Translation: A Case Study from the GALE Program. *Proceedings from the Association for Machine Translation in the Americas (AMTA 2008).* Cambridge, MA.

John Hutchins, 2001. Machine translation and human translation: in competition or in complementation? *International Journal of Translation*, 13(1-2), 5–20.

Martin Kay. 1997. The proper place of men and machines language translation. *Machine Translation*,12:3–23. First appeared as a Xerox PARC working paper in 1980.

Martin Kay. 2000. Triangulation in translation. Keynote at the MT 2000 Conference, University of Exeter.

Elliott Macklovitch. 2004. The Contribution of End-Users to the TransType2 Project *Proceedings of Association for Machine Translation of the Americas (AMTA 2004)*, Washington DC.

Elliott Macklovitch, Guy Lapalme & Fabrizio Gotti. 2008. TransSearch: What are translators looking for? *Proceedings from the Association for Machine Translation in the Americas 2008 Conference (AMTA 2008).* Honolulu, HI.

Elliott Macklovitch, Guy Lapalme and Nelida Chan. 2009. Term-spotting with TransCheck: A Capital Idea. *First International Workshop on Terminology and Lexical Semantics*, Montréal, Canada. 3-12.

Elliott Macklovitch and A. Valderrábanos. 2001. Rethinking Interaction: The solution for high-quality MT? MT Summit 2001, Santiago de Compostela, Spain.

Julia Makoushina. 2008. A Comparison of Eight Quality Assurance Tools. Multilingual, June 2008. Multilingual Computing.

Daniel Marcu. 2008. Connecting Consumers to User Generated Content with Machine Translation. *Proceedings of the Conference of the Machine Translation Association for the Americas (AMTA 2008),* Honolulu, HI.

Daniel Marcu. 2008. Human vs. Machine: Competition or Collaboration? *Proceedings of the Conference of the Machine Translation Association for the Americas (AMTA 2008),* Honolulu, HI.

Ion Maslea. 2009. Personal correspondence.

JoAnn Ryan. 1993. Machine Translation: Matching Reality to Expectations. *Progress in Machine Translation*, ed. Sergei Nirenburg. Amsterdam: IOS Press, 225-235.

Serena Shubert and Jan Spyridakis, The Translatability of Simplified English Documents. Matching Information to Audience. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00554897

Tapling, Mark. 2008. Science Meets Solution. *Proceedings of the Conference of the Machine Translation Association for the Americas (AMTA 2008)*, Honolulu, HI.

Kathryn Taylor and John White. 1998. Predicting What MT is Good for: User Judgments and Task Performance. *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*. Washington DC.

Vasconcellos, Muriel. Functional considerations in the postediting of machine-translated output. *Machine Translation* 1(1). 21-38.