Applied Language Learning

VOLUME 18 · NUMBERS 1 & 2

APPLIED LANGUAGE LEARNING

2008 VOLUME 18 NUMBERS 1 & 2

Applied Language Learning

Volume 18 Numbers 1 & 2

Articles

Review

General Information

From the Editor

Reviewers for *Applied Language Learning*

The individuals listed below served as reviewers of manuscripts submitted to *Academic Journals including Applied language Learning* in 2007 and 2008. We express our gratitude for expert service to:

Christine M. Campbell
Defense Language Institute
Foreign Language Center
Marianne Celce-Murcia
University of California
Los Angeles
Andrew Cohen
University of Minnesota
Tracey M. Derwing
University of Alberta
Edmonton
Dan Douglas
Iowa State University
Donald Fischer
Defense Language Institute
Foreign Language Center
Teresa Gryminska
Defense Language Institute
Foreign Language Center
Evelyn Hatch
University of California
Los Angeles
John S. Hedgcock
Monterey Institute of
International Studies
Eli Hinkel
Seattle University
Gordon Jackson
Defense Language Institute
Foreign Language Center
Renee Jourdenais
Monterey Institute of
International Studies
Betty Lou Leaver
Defense Language Institute
Foreign Language Center
James F. Lee
University of Indiana
John A. Lett, Jr
Defense Language Institute
Foreign Language Center
Ronald P. Leow
Georgetown University
Paul Nation
Victoria University of
Wellington
Maria Parker
Duke University
Thomas Parry
Defense Language Institute
Foreign Language Center
Victor Shaw
Defense Language Institute
Foreign Language Center
David J. Shook
Georgia Institute of
Technology
Richard Sparks
College of Mount Saint Joseph
Charles Stansfield
Educational Testing Service
Swathi Vanniarajan
San Jose State University
Heejong Yi
Defense Language Institute
Foreign Language Center

Applied Language Learning
2008, Vol. 18, Nos. 1 & 2, pp 17 - 25

Quantifying the Efficiency of a Translator

The Effect of Syntactical and Literal Written Translations on Language Comprehension using the Machine Translation System FALCon

(Foreign Area Language Converter)

Ian A. McCulloh, Jillian Morton, Jennifer K. Jantzi
Amy M. Rodriguez, and John Graham
United States Military Academy

Today, there are approximately 20,000 linguists with language training in either the Active Duty or Reserve components of the U.S. Army. Of those 20,000, half belong to the Military Intelligence branch (LaRocca 1). More than ever, the Army has increasing interest and need for accurate language translation especially with the Global War on Terror (GWOT). Coalition operations and U.S. presence in Iraq, Kuwait, and other areas in the Middle East require Arabic translation. Unfortunately, the Army has never been able to maintain the number of linguists it needs, particularly in the hard-to-fill, low-density languages (Dunn 1).

The U.S. Army operating on foreign soil in both peacekeeping and combat operations cannot afford to ignore the language barrier. In addition to communicating with the populace or gleaning intelligence from enemy documents, the Army is increasingly cooperating with coalition forces, which also introduces a variety of languages and thus other language barriers. To overcome this problem, the most obvious solution would seem hire more translators. However, there are disadvantages to this. For one, the quality of translations can be mixed; some translators may be better communicators than others. More importantly, hiring translators can be downright dangerous. Translators could be operating in conjunction with the enemy or be providing false information to that effect. To alleviate the burden of language translation, many are looking toward machine translation, as a means to augment linguists in theater.

Opponents of automated machine translation cite the multiple errors that occur and thus conclude that machine translation does not add significant benefits. Previous evaluations of machine translations usually rely on word error rate. Word error rate, designed to measure accuracy, is calculated by adding the number of insertions, deletions, or substitutions of words in one language to another language (LaRocca 1). Usually word error rate is determined using a computer program which calculates using the following equation, $\frac{n - (numberoferrors)}{n}$, where n is the number of characters in the groundtruth file, and every character inserted, substituted, or deleted counts as an error since the translation is based on optical character recognition (OCR). Essentially, one computer application rates the accuracy of another, the machine translation.

The problem with this evaluation method is that it does not take into account human cognition or context. In other words, a machine translation might have a high word error rate but the user can still understand the "gist." In short, past methods of evaluation do not consider user knowledge or experience. Machine translation systems should be

rated not in terms of their word error rate but in terms of human comprehension and usefulness, which is some function of word translation, syntax translation, and semantic interpretation. Where, semantics refers to the basic linguistic meaning of morphemes, words, phrases and sentences. In short, usefulness is a function of "gist" where "gist" is the human interpretation of the machine translated text.

The purpose of this study is to introduce a new method of evaluating human comprehension in the context of machine translation using a language translation program known as the FALCon (Forward Area Language Converter).

In the past, machine translation systems have been judged on their word error rate (number of substitutions, deletions, and insertions).

Our experiment will introduce another evaluation method for determining the success of machine translation by focusing on human cognition.

Figure 1

The FALCon works by converting documents into digital images via scanner, and then converting those images to electronic text by use of the Optical Character Recognition (OCR) (ARL, 2004). Foreign text is then converted to English using the Machine Translator (MT). In all, the FALCon can negotiate 61 languages, though some languages do not have OCR capacity and the quality of translation varies between languages. Semitic languages such as Arabic, tend to be the most challenging for machine language translators. An example translation of an Arabic passage by a human and a machine translator are provided:

Human Translation:

> Army Major General Richard Zahner told reporters at the Combined Press Information Center in Baghdad, Iraq, September 27 that, as part of the Coalition's strategy for success in Iraq, "We're having to block the Shiite extremists from linking with Iran."

Machine Translator (Cybertran) Translation:

> informed the float/general in the military Richard ZANR thee correspondents from the station the information your joint in the capital the Iraqi Baghdad day 27 september current groan? "loss forced prevention the extremist/extreme Shiites from the attachment with Iran."

The FALCon has two translation systems; Cybertran and Transphere. Cybertran negotiates text at the literal level while the Transphere attempts to incorporate syntactical meaning in its translations. Syntax refers to the rules of sentence formation; the component of the mental grammar that represents speakers' knowledge of the structure of phrases and sentences. For example, in Spanish, syntax includes a noun followed by an adjective: i.e. "*Tengo la camisa negra*" which taken literally in English means: "I have the shirt black" Cybertran would translate in this manner. However, Transphere would syntactically adjust that same sentence to: "I have the black shirt."

The National Institute of Standards for Technology (NIST) standard for testing competing machine translators may not be appropriate for measuring the ease and quality of reader comprehension. Cybertran leaves the reconstruction of a sentence and the context to the analyst, relying on the analysts' learned understanding of Arabic sentence form. Transphere attempts to incorporate syntactical rules into the translation. In this way, Transphere attempts to make the sentence structure more similar to the English language; however, it introduces random words into the process that decreases performance based on word error rate.

In addition to structure of language, readers also rely on schema to increase the understanding of text. Schemas help linguists understand the story structure (Braintree). Though literal translation is a priority for the reader, the coherent meaning constructed by the reader will often reflect a reader's prior experience. Recall protocols of foreign language students demonstrate that though students can often recognize words, they seriously misread or misconstrue their meaning within different contexts (Swaffar 123). The more familiar a linguist is with the structure of a language, the better they will be at grasping the "gist."

Others argue that machine translators have a poor reputation because people have the wrong expectations of what machine translators are capable of. They should be seen as a tool that can be used to assist in translating because "even if machine translation systems can never duplicate human translations, can't they at least generate output that is understandable and useful (Myers 2)?"

It is hypothesized that semantic machine translations (Transphere) will result in better reading comprehension as the reader begins to develop an implicit understanding of the sentence structure. A second hypothesis proposes that over time and with practice, the literal machine translation system (Cybertran) will produce a reading comprehension curve that increases over time while the semantic translation system will initially be higher due to its resemblance of the English language, but then over time, will level off because of the noise it introduces. This concept is illustrated in Figure 2.

Methodology

An experiment was conducted to compare the two proposed hypotheses and suggest an improved metric for evaluating a machine translation. The participants for this experiment included 48 freshmen from the United States Military Academy enrolled in the General Psychology course, PL100. Seven Arabic news documents were translated using the FALCon software, specifically the CyberTran and Transphere programs. The articles ranged in topics from reports on global terrorism to the weather. Participants were asked to read the machine translations of the seven Arabic documents and answer a series of corresponding comprehension questions.

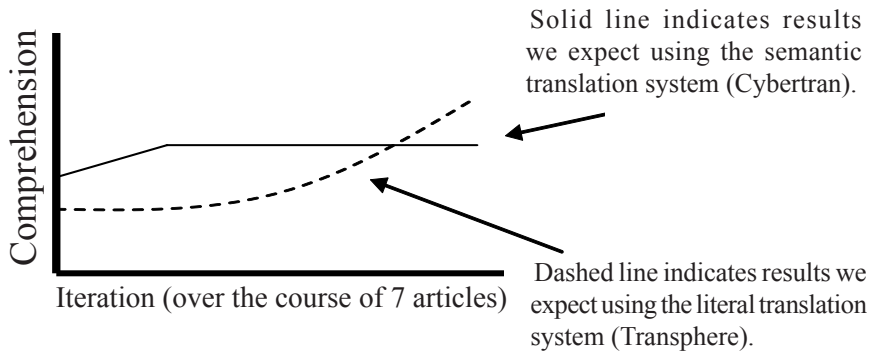

Figure 2. Illustration of Hypothesized Learning Curves for Machine Translators

This experiment was a between subjects design. The participants were equally divided into two groups. It is important to note that when using the FALCon program, this experiment used electronic mediums, thus eliminating the step involving the OCR (See Figure 3). The Arabic documents were the same for each group except for the type of translation used to convert them to English. Each group was exposed to a different condition; one group received the seven Arabic documents translated into English using the Transphere program. The second group received the documents translated into English using the Cybertran program. Each participant in both of these groups received the articles in a random order.

For each of the seven articles, the participants received a set of corresponding comprehension questions. The questions were the same for each participant, despite the condition. The participants were instructed to read and answer the comprehension questions to the best of their ability. Once they finished answering the questions, the participants were given the master English copy of the article so that they could compare this document to the translation produced by either the Transpere or Cybertran to see if they could better understand syntax, vocabulary, etc.

Each test for the seven articles was designed in the same format consisting of two multiple choice questions, one fill-in the blank question, one true/false question and a two-part question wherein subjects must re-structure a translated sentence from both the Transphere and Cybertran translators. The goal of putting the questions in a particular order was to gear the reader toward intelligence gathering and to see if he could grasp main concepts and details, and overtime (though not yet evaluated in this study), have him answer these kinds of questions without being prompted. The first question is always a main idea question, to gauge the reader's overall understanding and force him to think about the main concepts of the subject before he answers smaller questions. The second question was a detail within the article that was important to the overall article. This detail either asked for a key person or place within the article. The fill-in the blank question was another detail, but not as specific as the multiple choice question; it would ask for how often an occurrence happened or who a significant person was (based on position more than specific name). The true/false question was geared to be a little tricky to readers to see if they truly understood a broad concept of the article. The question would entail a detail that encompassed the overall significance of the article. For instance, one question read, "True or False: Each of these Iraqis thinks that the establishment of a government

will solve the problems in the country." Listening to current press reports in general, most people would choose false, but those who read and understood this particular article would correctly answer true to this question, thus the question aids (but does not define) the assessment of one's understanding of the article. The last two questions are meant to gauge which article would be more conducive to translation from "translator garb" to understandable English. This is done by asking the subjects to re-construct two sentences, one from each of the translators, into a coherent sentence. This last question really helps evaluate whether human understanding and interpretation can fill in the gaps of a poorly-written document by having the reader re-configure the sentence in their own words, while retaining the original meaning of the article. Each test received a score based on a 24 point scale; much like a teacher would grade a test for students. Each multiple choice and true/false question was worth 3 points, while each fill-in the blank and short answer question was worth 5 points. Partial credit was awarded to those answers that showed some valid comprehension of the material.

Analysis and Results

The initial and most important analysis sought to assess whether there was actual "learning" among subjects. If learning were present, the results would show an increase of correct responses over time. There was insufficient power to show statistically significant learning for overall test scores, however, it appears that for certain questions, correct responses may increase over time. The questions that showed an increase in correct responses over time were the ones which required specific answers, such as the multiple choice for detail, fill-in the blank, and true/false. Figure 3 shows the average scores for participants over the span of the test from the first article they were given to the last article for the specific-response type questions.

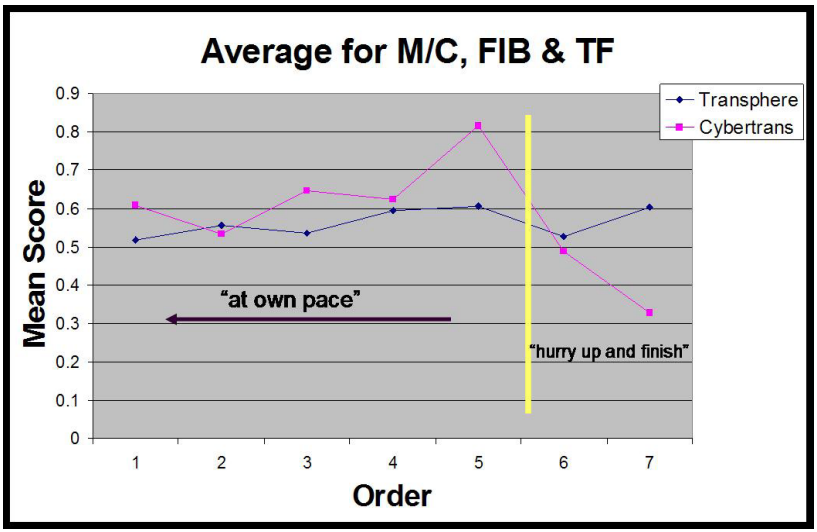

Figure 3. Average Test Score for Multiple Choice, Fill-in-the-Blank, and True/False Questions

The vertical line in Figure 3 separates the final two tests, wherein subjects had to rush to finish the test within the given time period. Until that time, the number of correct Cybertran responses was improving consistently, while the Transphere scores were rather steady. Cybertran's word-for-word translations seemed to make picking out details within the article a simpler task for subjects than the Transphere's translations. Showing even more evidence of the difference between the two types of translation is the graph of the multiple choice detail questions over time, as seen in Figure 4.

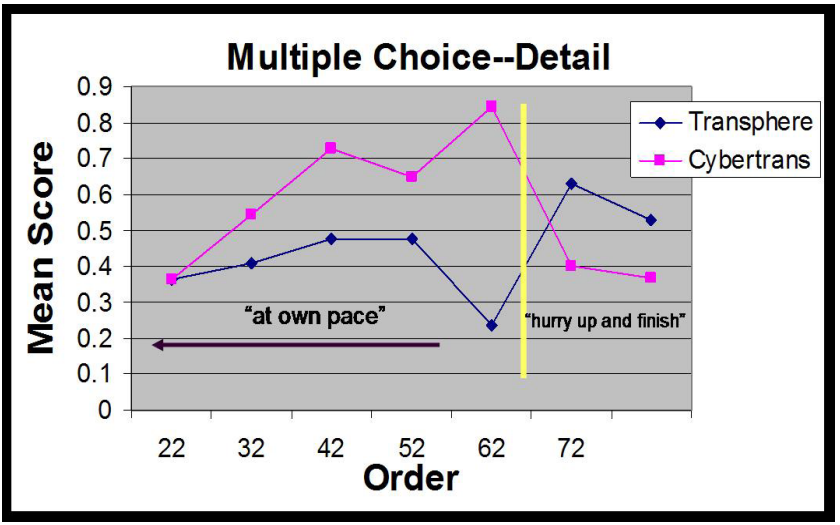

Figure 4. Average Test Score for Multiple Choice Detail Questions

While the two types of translations started at the same place for the first question, each question after that showed a steady rise in percent answered correctly for subjects reading the Cybertran translations. However, it is of note to mention that when answering in haste, the subjects reading Cybertran translations struggled, probably because they did not have time to look back in the article for key words. The Transphere most likely did better because it gives readers a better idea of the broad sense of the article, so they could still venture a good guess even when in a hurry.

The difference in overall scores for articles and for individual questions was also studied. There is more evidence supporting the strength each translator has in either the detail aspect or the broad idea aspect. For instance, Figure 5 depicts the difference in correct responses between each translator by questions asked.

Those questions for which Transphere translations prompted more correct responses were broad idea questions, which means that instead of asking for a particular person or fact, they ask for an idea or underlying concept. The two bars on the right of Figure 5 are the sentence re-structuring questions, and Transphere has done better on that as well. Subjects who read either type of machine translation found that restructuring the Transphere sentences was easier, although subjects who read only Transphere documents responded better to the Transphere re-structuring by 7.06 percent. Those subjects who only read Cybertran sentences only did better restructuring the Cybertran sentences by 0.93 percent than the Transphere readers.

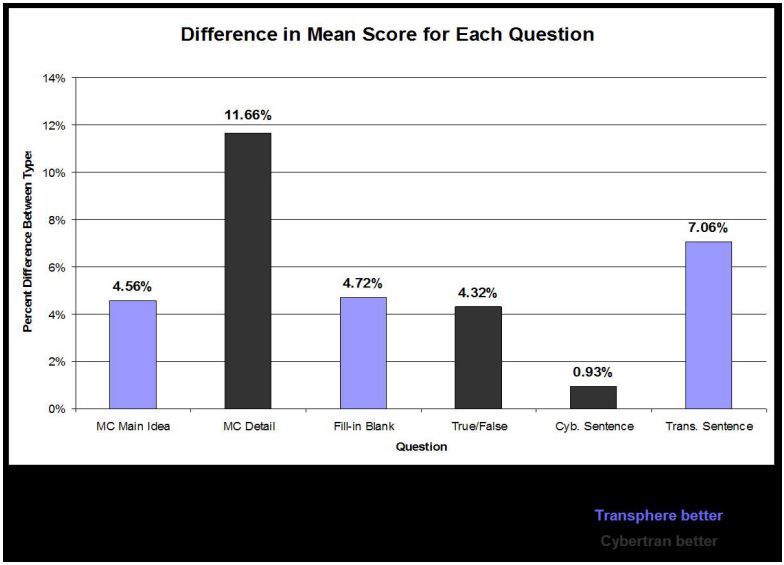

Figure 5. Difference in Mean Score for Each Question

Analysis of each article provides insight into types of articles and the capacity of understanding that each type affords its reader. Figure 6 displays the difference in average scores for the two translation types for each of the seven articles.

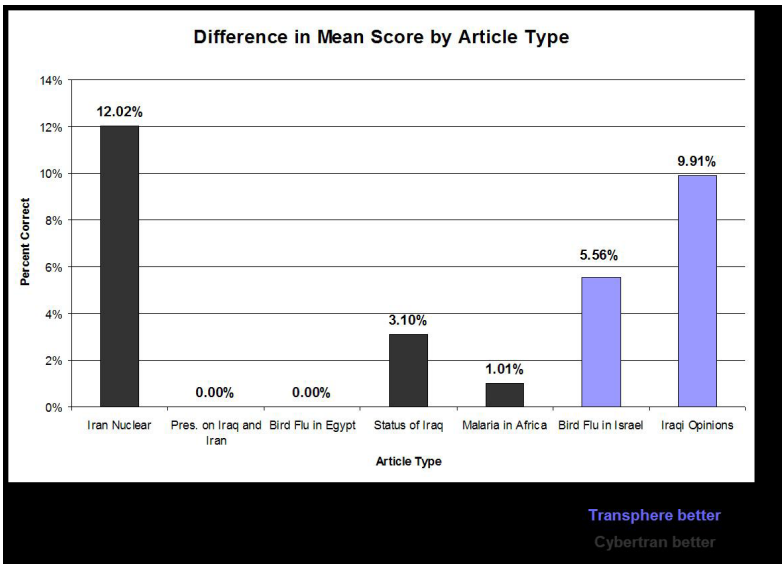

Figure 6. Difference in Mean Score for Each Article

It can be seen that those subjects who read the Cybertran translation of the article concerning Iran's nuclear intentions had a better average score by 12 percent. Meanwhile the subjects who read the Transphere version of an article on Iraqi opinions had a better average score by about 10 percent. Again, this difference is likely attributed to the content and technicality of the article. The Iran article contained more details, which are easier to pinpoint using the word-for-word Cybertran translation. The understanding of the Iran article is extremely dependent upon key actors, which are generally proper nouns, an aspect of translation wherein Cybertran outperforms Transphere. To understand the meaning of the "Iraqi Opinions" article, the reader would need to understand full concepts rather than small details, an aspect of translation that syntactical correction aids humans in doing. Since the Transphere translator uses these syntactical translations, more people understood the meaning of the "Iraqi Opinions" article using this type of translation.

Conclusion and Future Study

The results of this study have brought a few key points for consideration in Arabic machine translation. First, human understanding is not a factor to be ignored in gauging the usefulness of such translators. Secondly, the type of translation used can depend on the type of information needed, whether it is key people and places or the general plans or opinions. An even better method would combine the two types (probably through human interpretation) to have one complete translation with both key details and the right concepts. Combining the strengths of the two types is especially important in developing a training strategy to employ translators like the FALCON for intelligence gathering.

The benefits of the research for machine translators are expansive. Machine translators could be great tools for Army intelligence, that is, if humans could readily understand the texts. One of the biggest problems with Arabic machine translators is that the translated text still has to be sent to a linguistic expert for interpretation. If a soldier can be trained to interpret within a relatively short period of time, then the lengthy process of finding a linguist and sending and receiving a document can be eliminated, and articles can be processed and interpreted in a very short time by a member of the unit.

For further study, the learning factor requires further evaluation. Can people be trained to understand the machine translation better? If so, is one of the two types of translators easier to learn? To evaluate these questions, subjects could perform numerous test sessions over the period of a week or two instead of working for only an hour. During this time, subjects may slowly adapt to a different kind of test that would change from some multiple choice questions at the beginning to short answer and eventually to straight essay at the end, wherein they would attempt to touch on all the same key concepts from the first tests. If a person could obtain all the important information without being guided by questions, then that would truly test his understanding of the article and prove the translator's value to the intelligence community. To further assess the learning involved, the subject could be made aware of the exact rules that go into each translation type, and then be given the tests, instead of attempting the test without knowing anything about the type of translation they are reading.

References

Army Research Laboratory. (2004). "Training: Forward Area Language Converter (FALCon)."

Army Research Laboratory. (2004). "Forward Area Language Converter (FALCon)."

Dunn, K. (2003). "Language tools- automated translation." Military Intelligence Professional Bulletin (Jan-March, 2003)

[Website online]. Available from: http://www.findarticles.com/p/articles/mi_m0IBS/is_1_29/ai_97822089; accessed 15 Sep. 2005.

Brain Connection. (2001). Paragraph Comprehension: The Connection to Reading Skills.

[Website online] http://www.brainconnection.com/content/5_1; accessed 15 Sep 2005

Human Intelligence and Counterintelligence (CI) Support Tools (HICIST). (2005)

[Website online] Available from: http://www.globalsecurity.org/intell/systems/hicist.htm; accessed 15 Sep 2005.

Myers, S. (1996). "Can Computers Translate?" Computing Japan Magazine

Swam, K. (1999). "FALCon: Evaluation of OCR and Machine Translation Paradigms." US Army Research Laboratory.

Tanner, S. (2004). "Deciding Whether Optical Character Recognition is Feasible." King's Digital Consultancy Services.

Authors

IAN A. MCCULLOH, MAJ. Department of Mathematical Sciences, United States Military Academy, West Point NY 10996 USA.

JILLIAN MORTON, 2LT, Department of Mathematical Sciences, United States Military Academy, West Point NY 10996 USA.

JENNIFER K. JANTZI, 2LT, Department of Behavioral Sciences and Leadership, United States Military Academy, West Point, NY 10996 USA.

AMY M. RODRIGUEZ, 2LT, Department of Behavioral Sciences and Leadership, United States Military Academy, West Point, NY 10996 USA.

JOHN GRAHAM, LTC, Department of Behavioral Sciences and Leadership, United States Military Academy, West Point, NY 10996 USA.

Applied Language Learning
Defense Language Institute
Foreign Language Center
Presidio of Monterey, CA 93944-5006

PB-65-08-1
United States Army
PIN: 084716-000
Approved for public release.
Distribution is unlimited.