Task Tolerance of MT Output in Integrated Text Processes

John S. White, Jennifer B. Doyon, and Susan W. Talbott
Litton PRC
1500 PRC Drive
McLean, VA 22102, USA
{white_john, doyon_jennifer, talbott_susan}@prc.com

Abstract

The importance of machine translation (MT) in the stream of text-handling processes has become readily apparent in many current production settings as well as in research programs such as the Translingual Information Detection, Extraction, and Summarization (TIDES) program. The MT Proficiency Scale project has developed a means of baselining the inherent "tolerance" that a text-handling task has for raw MT output, and thus how good the output must be in order to be of use to that task. This method allows for a prediction of how useful a particular system can be in a text-handling process stream, whether in integrated, MT-embedded processes, or less integrated user-intensive processes.

1 Introduction

Issues of evaluation have been pre-eminent in MT since its beginning, yet there are no measures or metrics which are universally accepted as standard or adequate. This is in part because, at present, different evaluation methods are required to measure different attributes of MT, depending on what a particular stakeholder needs to know (e.g., Arnold 1993). A venture capitalist who wants to invest in an MT start-up needs to know a different set of attributes about the system than does a developer who needs to see if the most recent software changes improved (or degraded) the system. Users need to know another set of metrics, namely those associated with whether the MT system *in situ* improves or degrades the other tasks in their overall process. Task-based evaluation of this sort is of particular value because of the recently envisioned role of MT as an embedded part of production processes rather than a stand-alone translator's tool. In this context, MT can be measured in terms of its effect on the "downstream" tasks, i.e., the tasks that a user or system performs on the output of the MT.

The assertion that usefulness could be gauged by tasks to which output might be applied has been used for systems and for processes (JEIDA 1992, Albisser 1993), and also particular theoretical approaches (Church and Hovy 1991). However, the potential for rapidly adaptable systems for which MT could be expected to run without human intervention, and to interact flexibly with automated extraction, summarization, filtering, and document detection calls for an evaluation method that measures usefulness across several different downstream tasks.

The U.S. government MT Functional Proficiency Scale project has conducted methodology research that has resulted in a ranking of text-handling tasks by their tolerance to MT output. When an MT system's output is mapped onto this scale, the set of tasks for which the output is useful, or not useful, can be predicted. The method used to develop the scale can also be used to map a particular system onto the scale.

Development of the scale required the identification of the text-handling tasks members of a user community perform, and then the development of exercises to test output from several MT systems (Japanese-to-English). The level of ease users can perform these exercises on the corpus reflects the tolerance that the tasks have for MT output of varying quality. The following sections detail the identification of text-handling tasks, the evaluation corpus, exercise development, and inference of the proficiency scale from the apparent tolerance of the downstream text-handling tasks.

2 Proficiency Scale Development

In order to determine the suitability of MT output for text-handling tasks, it was necessary to interview users of text-handling tools to identify the tasks they actually perform with translated material. It was necessary also to compile a corpus of translations and create exercises to measure the usefulness of the translations.

2.1 Task Identification

Expert user judgments were needed to ensure confidence in the resulting proficiency scale. The users who provided these judgments work monolingually on document collections that include translated material. Preliminary interviews were conducted with 17 users. During the preliminary interviews, users completed questionnaires providing information identifying the text-handling tasks that ultimately formed the proficiency scale.

2.2 Corpus Composition

For a 1994 evaluation effort, the Defense Advanced Research Projects Agency (DARPA) Machine Translation Initiative developed a corpus of 100 general news texts taken from Japanese newswires. These texts were translated into English and were incorporated into what is now known as the "3Q94" evaluation. A subset of these translations was used for the MT Functional Proficiency Scale project.

The 100 3Q94 Japanese source texts were translated into six English output versions, four from commercial and research MT systems (Systran (SY), Pivot (P), Lingstat (L), and Pangloss (PN)), and two from professional expert translations (E) used as baseline and control for the 3Q94 evaluations. Translations were selected from all of these sets for the proficiency scale corpus. For the purpose of validating the project's results, two additional systems' translations were added to its corpus. These included translations from a current version of Systran (SY2) and Typhoon (TY).

2.3 Exercise Definitions

The user exercises were designed to determine if users could successfully accomplish their regular tasks with translations of varying qualities, by eliciting judgments that indicated the usefulness of these translations. A variety of human factors issues were relevant to the development of the exercise sets. Since the texts to be seen by the users were general news texts, it was unlikely they would be relevant to the users' usual domains of interest (White and Taylor, 1998 and Taylor and White, 1998). This issue was handled by selecting texts related to domains that were thought to be similar, but broader, than those typically handled by users (White and Taylor, 1998 and Taylor and White, 1998). Additionally, the simple elicitation of a judgment (to a question such as "can you do your job with this text") is possibly biased by a predisposition to cooperate (Taylor and White 1998). Therefore, it was necessary to develop two complementary sets of exercises: the snap judgment exercise and the task-specific exercises. Detailed definitions of these two exercises can be found in Kathryn B. Taylor and John S. White's paper "Predicting What MT is Good for: User Judgments and Task Performance" in the Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA '98.

3 Results

3.1 Compilation of Responses

The user responses for the snap judgment exercise are shown in Exhibit 1. In the snap judgment exercise, the users were asked to look at 15 translations and categorize each as being of a good enough quality to successfully complete their text-handling task, i.e., "YES" or "Y," or if they could not use the translation to perform their task, i.e., "NO" or "N." The top row of Exhibit 1 lists the 15 translations by their document identification codes. Each document identification code includes a document number followed by the code of the MT system that produced it (MT system codes can be found in the Corpus Composition section above). The first column of Exhibit 1 contains a list of the users who participated in the snap judgment exercise separated by which text-handling task they performed. The users' responses of "Y" or "N" appear under each of the translations' document identification codes by user. The snap judgment scores for each of the text handling tasks was calculated as the percentage of "Ys" for the corpus of 15 translations by all users performing that task.

The user responses and results for the gisting exercise are shown in Exhibit 2. In the gisting exercise, each user was asked to rate decision points in a translation on a 1-5 scale. The top row of Exhibit 2 lists the seven documents seen by the users by their document identification codes. The first column of Exhibit 2 contains a list of users who participated in the gisting exercise. User ratings averaged for each translation appear under each of the translation codes for each of the users. The scores for each of the translations were calculated by totaling a user's ratings and dividing that total by the number of decision points contained in the document.

The user responses and results for the triage exercise are shown in Exhibit 3. In the triage exercise, each user was asked to order three separate stacks of translations by their relevance to a problem statement. The top row of Exhibit 3 lists the 15 translations seen by the users by their document identification codes. The first column of Exhibit 3 contains a list of users who participated in the triage exercise. User responses of ordinal number rankings appear under each of the document identification codes by user. Each of the category rankings was scored by comparing its results to that of a ground truth ranking of the same translations.

The user responses and results for the extraction exercise are shown in Exhibit 4. In the extraction exercise, each user was asked to identify named entities in each translation: persons, locations, organizations, dates, times, and money/percent. This extraction exercise was modeled after the "Named Entity" task of the Message Understanding Conference (MUC) (Chinchor and Dungca, 1995). Exhibit 4 contains two charts. The top row of both charts contain a list of users who participated in the extraction exercise. The first column of both charts lists seven documents seen by the users by their document identification codes. In the top chart, recall scores appear under each of the users for each translation. In the bottom chart, precision scores appear under each of the users for each translation. Recall was calculated by the number of possible named entities in a translation the user identified. Precision was calculated by the number of items the user identified as being named entities that were actually named entities.

The user responses and results for the filtering exercise are shown in Exhibit 5. In the filtering exercise, each user was asked to look at 15 documents to determine if a document fit into any one of the three categories of Crime, Economics, or Government and Politics, i.e., "YES" or "Y," none of the three categories, i.e., "NO" or "N," or if they could not make a decision either way, i.e., "CANNOT BE DETERMINED" or "CBD." Exhibit 5 contains two charts. The top row of both charts lists the 15 translations seen by the users by their document identification codes. The first column of both charts contains a list of users who participated in the filtering exercise. The users' responses of "Y," "N," or "CBD" appear under each of the translations' document identification codes by user. The results of the filtering exercise were calculated with the measure of recall. Recall was calculated by the number of translated documents related to the three categories of Crime, Economics, and Government and Politics the user identified.

The user responses and results for the detection exercise are shown in Exhibit 6. In the detection exercise, each user was asked to look at 15 documents to determine if the document belonged to the category of Crime (C), the category of Economics (E), the category of Government and Politics (G&P), none of the three categories, i.e., "NO" or "N," or if they could not make a decision either way, i.e., "CANNOT BE DETERMINED" or "CBD." Exhibit 6 contains three charts. The top row of all three charts lists the 15 translations seen by the users by their document identification codes. The first column of all three charts contains a list of users who participated in the detection exercise. User responses of "C," "E," "G&P," "CBD," or "NOTA" appear under each of the translations' document identification codes by user. The results of the detection exercise were calculated with the measure of recall. Recall was calculated by the number of translated documents related to each of the three categories of Crime, Economics, and Government and Politics the user identified.

3.2 Mapping Results onto Tolerance Scale

The results of the snap judgment exercise are shown in Exhibit 7. In the snap judgment exercise each user was asked whether a document was coherent enough that it could

	2062E	2005L	2088TY	2040SY	2079L	2053TY	2057PN	2019P	2020P	2032SY	2021P	2045SY	2043P	2031PN	2077L
GISTING															
User A	Y	Y	Y	N	N	N	Y	Y	N	N	Y	N	Y	N	N
User B	Y	Y	Y	N	Y	N	N	N	N	N	N	N	N	N	N
User C	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N
TRIAGE															
User D	Y	Y	N	Y	N	Y	N	Y	N	N	N	N	N	N	N
User E	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N
User F	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N
User G	Y	Y	Y	Y	Y	Y	Y	N	Y	N	N	Y	N	N	Y
EXTRACTION															
User H	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	N	N	Y	Y
User I	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N
User J	Y	N	N	N	N	N	N	N	N	N	N	N	N	N	N
FILTERING															
User K	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	N	Y	N	N	N
User L	Y	Y	Y	Y	N	Y	Y	N	Y	Y	Y	N	N	N	N
User M	Y	Y	Y	Y	Y	Y	N	Y	N	Y	N	N	Y	N	N
DETECTION															
User N	Y	Y	Y	Y	Y	N	N	Y	N	N	Y	N	N	Y	N
User O	Y	N	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	N	N	N
User P	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	N	N

Exhibit 1 - Snap Judgment Results

	2051E	2070SY2	2049L	2069PN	2082TY	2055P	2050SY
GISTING							
User A	4.46	2.47	2.13	2	1.73	1.71	2
User B	4.62	3.47	2.2	2.31	2.09	1.79	2.08
User C	4.85	3	2.13	2	2.18	2.29	1.46
AVERAGE	4.64	2.98	2.15	2.10	2.00	1.93	1.85
	MEAN(MEANS)	2.52					
ACCEPTABLE	YES	YES	NO	NO	NO	NO	NO

Exhibit 2 - Gisting Results

	CRIME	UOA=1.05						ECONOMICS	UOA=.676			GOVERNMENT & POLITICS	UOA=.236		
	2070	2069	2050	2049	2082	2055	2051	2056	2072	2023	2028	2078	2046	2012	2004
TRIAGE															
Ground Truth	1	2	3	4	5	6	7	1	2	3	4	1	3	4	2
User D	4	1	6	5	7	3	2	1	2	4	3	1	2	3	4
User F	3	1	2	5	CBD	6	4	1	CBD	2	3	1	3	4	2
User G	2	1	6	4	3	5	7	1	2	3	4	1	2	3	4
TOTAL DISTANCE	6	3	7	2	6	5	8	0	2	2	2	0	1	1	2
AVG DISTANCE	2	1	2.3333	0.6667	2	1.6667	2.667	0	0.667	0.66667	0.66667	0	0.3333	0.33333	0.666667
ACCEPTABILITY	NO	YES	NO	YES	NO	NO	NO	YES	YES	YES	YES	YES	NO	NO	NO

Exhibit 3 - Triage Results

	RECALL - User H	RECALL - User I	RECALL - User J	TOTAL RECALL		ACCEPTABLE:
EXTRACTION						
2082TY	87.4%	77.7%	77.9%	81%		YES
2051E	76.6%	70.5%	84.9%	77.3%	AV(R):	YES
2070SY2	63.9%	77.3%	57.0%	66.1%	62%	YES
2055P	69.2%	43.4%	72.0%	61.5%		NO
2050SY	57%	53%	57.6%	55.9%		NO
2049L	52.8%	57%	47.8%	52.5%		NO
2069PN	32.5%	34.9%	51.2%	39.5%		NO

	PRECISION - User H	PRECISION - User I	PRECISION - User J	TOTAL PRECISION		
EXTRACTION						
2055P	97.2%	97.6%	95.2%	96.6%		YES
2082TY	95.2%	100%	91.7%	95.6%		YES
2069PN	96.7%	81.7%	100%	92.8%	AV(P):	YES
2050SY	88.9%	95.8%	91.1%	91.9%	87.7%	YES
2051E	81.1%	71.1%	92.4%	81.5%		NO
2070SY2	76.3%	74.6%	87.2%	79.4%		NO
2049L	75.5%	74.1%	78.1%	75.9%		NO

Exhibit 4 - Extraction Results

YES-CRIME	2049L	2051E	2069PN	2070SY2	2050SY	2055P	2082TY	RECALL
FILTERING								
User K	Y	Y	Y	Y	N	Y	CBD	71.4%
User L	Y	Y	Y	Y	Y	Y	CBD	85.7%
User M	Y	Y	N	N	Y	CBD	N	42.9%
			AV(R)	66.7%				
ACCEPTABLE	YES	YES	YES	YES	NO	NO	NO	

NO-CRIME	2056P	2072L	2046PN	2078L	2023SY	2012SY	2028PN	2004P	RECALL
FILTERING									
User K	N	N	N	N	N	N	CBD	N	87.5%
User L	N	N	N	N	Y	Y	Y	Y	50%
User M	N	N	N	N	N	N	N	CBD	87.5%
			AV(R)	75.0%					
ACCEPTABLE	YES	YES	YES	YES	NO	NO	NO	NO	

Exhibit 5 - Filtering Results

CRIME	2049L	2050SY	2051E	2055P	2070SY2	2069PN	2082TY	RECALL
DETECTION								
User N	C	C	C	C	C	E	E	71.4%
User O	C	C	C	C	C	C	E	85.7%
User Q	C	C	C	C	C	C	E	85.7%
User P	C	C	C	C	C	C	NOTA	85.7%
				AV(R)	82.1%			
ACCEPTABLE	YES	YES	YES	YES	YES	NO	NO	

ECONOMICS	2028PN	2056P	2072L	2023SY	RECALL
DETECTION					
User N	E	E	E	E	100%
User O	E	E	E	CBD	75%
User Q	E	E	E	E	100%
User P	E	E	E	E	100%
		AV(R)	94%		
ACCEPTABLE	YES	YES	YES	NO	

GOV & POL	2078L	2046PN	2004P	2012SY	RECALL
DETECTION					
User N	G&P	G&P	CBD	E	50%
User O	G&P	NOTA	NOTA	G&P	50%
User Q	G&P	G&P	G&P	E	75%
User P	G&P	NOTA	NOTA	CBD	25%
	AV(R)	50%			
ACCEPTABLE	YES	YES	NO	NO	

Exhibit 6 - Detection Results

be used to successfully complete their assigned task exercise.

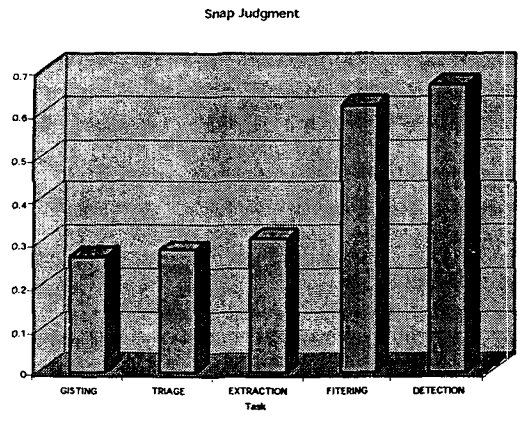

Exhibit 7 - Snap Judgment Results

The bars in Exhibit 7 represent the percentage of affirmatives for the corpus of 15 texts by all users.

The results for the user exercises needed be computed in a way which allowed their comparison across tasks, but which used the metrics relevant to each task at the same time. We address the computation of each of these in turn.

Gisting. Computing the acceptability cut-off for gisting follows the general pattern, except that the text scores are not recall or precision. Rather, since gisting judgments were elicited with an "adequacy" measure, each text for each user has an average of the scores for the decision points in that text. In turn, the average of these average scores gives the cutoff for acceptability for gisting, namely 2.52 out of a minimum of one and maximum of 5. By this means, 2 texts are identified as acceptable for gisting, indicated in Exhibit 2.

Triage. As shown in Exhibit 3, triage requires the comparison of ordinal rankings, with ordinal rankings from the ground truth set. Here, a uniformity of agreement measure was established, defined as the mean of the standard deviations for each text in each problem statement. Then the mean for each text in the user ranking was compared to the ground truth ranking, plus-or-minus the uniformity measure. A text is acceptable if it matches the ground truth within the uniformity measure. Based on this computation, 7 of 15, or 46.7%, of the texts are acceptable for gisting.

Extraction. Extraction was computed using both recall and precision measures. As with filtering and detection, average recall is computed (62%), which is used as the cut-off for acceptability, and identifies 3 texts as acceptable. Similarly, the average precision, 87.7%, creates a cut-off at 4 texts. To show extraction as a single value, the total acceptable in precision and in recall are averaged, equaling 3.5, or 50% of the texts in the 7-text set. These are shown in Exhibit 4.

Filtering. For filtering, user responses are computed on two tables conforming to the ground truth values for each text ("Y" or "N", i.e., whether the text was relevant to crime or not). The average recall over all users and all texts is 66.7% for Y and 75% for N. These averages create for the Y and N chart the respective cutoff boundaries for "YES" (text output is acceptable for filtering) and "NO" (it is not). The total number of YES's from the Y and N tables is 8 or 53% of the texts in the corpus acceptable for filtering. These results are illustrated in Exhibit 5.

Detection. As shown in Exhibit 6, there are three tables in detection, corresponding to the three domain areas of Crime, Economics, and Government and Politics. As with filtering, the average recall is computed for each domain over all users and texts, and this average establishes the cut-off boundary of acceptability of text outputs for detection. For the Crime domain, the average is 82.1%, for Economics 94%, and for Government and Politics 50%. The total number of texts thus identified as acceptable is 10, or 67% texts acceptable for detection.

Exhibit 8 shows the results of the task exercises.

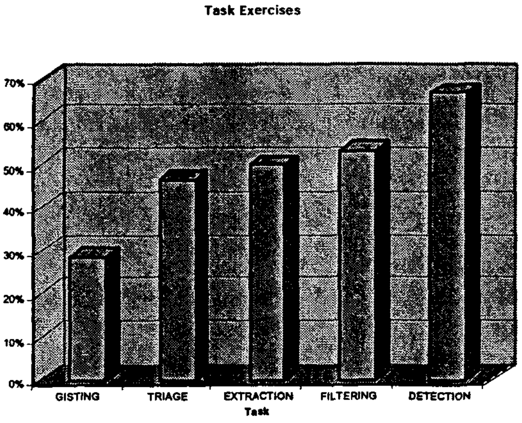

Exhibit 8 - Task Exercises Results

At the inception of this project, we established a heuristic scale of task tolerance, based on common understanding of the nature of each of these tasks. This scale – filtering, detection, triage, extraction, and gisting, in order of tolerance – was not a hypothesis per se; nevertheless, it is rather surprising that the results vary from the heuristic significantly. The results showed detection to be the most tolerant task, rather than filtering. The presumption had been that the filtering task, which simply requires a "yes" if a document is related to a specific topic or "no" if it is not, could be performed with higher accuracy than the task of detection that requires classifying each document by subject matter. In fact, when precision measures are factored in for filtering and detection (as they were for extraction), filtering appears to be even less tolerant than extraction. This outcome seems plausible when we consider that detection is often possible even when only small quantities of key words can be found in a document.

Also surprising, the triage task was less tolerant of MT output then expected. It was supposed that the ability to rank relevance to a particular problem could be done with sufficient keywords in otherwise unintelligible text; rather, a greater depth of understanding is necessary to successfully complete this task.

4 Future Research

There are at least two evaluation techniques that can use the task tolerance scale to predict the usefulness of an MT system for a particular downstream task. The set of exercises used to elicit the task tolerance hierarchy reported here can also be used to determine the position on the scale of a particular system. The system translates texts from the corpus for which ground truth has already been established, and the user exercises are performed on these translations. The result is a set of tasks for which the system's output appears to be suitable. The pre-existing scale can help to resolve ambiguous results, or can be used to make scale-wide inferences from a subset of the exercises: it may be possible to perform just one exercise (e.g., triage) and infer the actual position of the system on the scale by the degree of acceptability above or below the minimum acceptability for triage itself.

A second technique offers more potential for rapid, inexpensive test and re-test. This involves the development of a diagnostic test set (White and Taylor 1998, Taylor and White 1998), derived from the same source as the proficiency scale itself. For every task in the exercise results, there are "borderline" texts, that is, texts acceptable for one task but not for the next less tolerant task. These texts will exhibit translation phenomena (grammatical, lexical, orthographic, formatting, etc.) which are diagnostic of the difference between suitability at one tolerance level and another. The text will also contain phenomena that are not diagnostic at this level but are at a less tolerant level. By characterizing the phenomena that occur in the border texts for each task, it is possible to determine the phenomena diagnostic to each tolerance level.

A pilot investigation of these translation phenomena (Taylor and White 1998, Doyon et al. 1999) categorized the translation phenomena in terms of pedagogy-based descriptions of the contrasts between Japanese and English (Connor-Linton 1995). This characterization allows for the representation of several individual problem instances with a single suite of pair-specific, controlled, source language patterns designed to test MT systems for coverage of each phenomenon. These patterns may be tested by any MT system for that language pair, and the results of the test will indicate where that system falls on the proficiency scale by its successful coverage of the diagnostic patterns associated with that tolerance level.

The purpose of the user exercises is to establish a scale of MT tolerance for the downstream text handling tasks. However, the same method can be used to determine the usefulness of a particular system for any of the tasks by performing these exercises with the system to be tested. It is possible, for example, to isolate the performance of systems in the set used here, though the sample size from each system is too small to draw any conclusions in this case. We hope to perform this exercises with larger samples both to validate these findings and to execute evaluations on candidate MT systems.

Among other validation steps in the future will be confirmation of the exercise approach from an empirical perspective (e.g., whether to include "cannot be determined" as a choice), and a validation of the ground truth in the triage exercise.

Finally, we continue to refine the application of the methodology to reduce time and increase user acceptance. In particular, we have developed a web-based version of several of the exercises to make the process easier for the user and more automatic for scoring.

5 Conclusion

The MT Functional Proficiency Scale project has not only demonstrated that it is possible for poor MT output to be of use for certain text-handling tasks, but has also indicated the different tolerances each such task has for possibly poor MT output.

This task-based methodology developed in the MT Functional Proficiency Scale project using Japanese-to-English corpora should prove useful in evaluating other language pair systems. There is also potential for evaluating other text-handling systems, such as summarization, information retrieval, gisting, and information extraction, in the context of the other tasks that might process their output.

Task-based evaluations provide a direct way for understanding how text-handling technologies can interact with each other in end-to-end processes. In the case of MT systems, it is possible to predict the effective applicability of MT systems whose output seems far less than perfect.

6 References

Albisser, D. (1993). "Evaluation of MT Systems at Union Bank of Switzerland." Machine Translation 8-1/2: 25-28.

Arnold, A., L. Sadler, and R. Humphreys. (1993). "Evaluation: an assessment." Machine Translation 8-1/2: 1-24.

Chinchor, Nancy, and Gary Dungca. (1995). "Four Scorers and Seven Years Ago: The Scoring Method for MUC-6." Proceedings of Sixth Message Understanding Conference (MUC-6). Columbia, MD.

Church, Kenneth, 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. Pp. 147-157.

Connor-Linton, Jeff. (1995). "Cross-cultural comparison of writing standards: American ESL and Japanese EFL." World Englishes, 14.1:99-115. Oxford: Basil Blackwell.

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

Japanese Electronic Industry Development Association. (1992). "JEIDA Methodology and Criteria on Machine Translation Evaluation." Tokyo: JEIDA.

Taylor, Kathryn B., and John S. White (1998). "Predicting what MT is Good for: User Judgments and Task Performance." Proceedings of Third Conference of the Association for Machine Translation in the Americas, AMTA'98. Philadelphia, PA.

White, John S., and Kathryn B. Taylor. (1998). "A Task-Oriented Evaluation Metric for Machine Translation." Proceedings of Language Resources and Evaluation Conference, LREC-98, Volume I. 21-27. Grenada, Spain.