Trial and Error:
An Evaluation Project on Japanese <> English MT Output Quality

Maki Darwin
Mendez, Inc.
5095 Murphy Canyon Rd., #300, San Diego, CA 92123
U.S.A.
Maki.Darwin@lhsl.com

Abstract

This paper describes a small-scale but organized attempt to evaluate output quality of several Japanese MT systems. The project also served as the first experiment of the implementation of the in-house MT evaluation guidelines created in 2000. Since time was limited and the budget was not infinite, it was launched with the following compact components: Five people; 300 source sentences per language pair; and 160 hours per evaluator. The quantitative results showed noteworthy phenomena. Although the test materials had been presented in a way that evaluators could not identify the performance of any particular system, the results were quite consistent. The scoring ratio that the two E-to-J evaluators employed was almost identical, while that of the J-to-E evaluators was similar. This indicates that high-quality output has universal appeal. Additionally, the evaluators noted that stronger systems, regardless of language pair, tended to be superior in source sentence analysis, target sentence arrangement, word choice, and lexicon entries whereas weaker systems tended to be inferior in these areas. As for language-pair comparison, the results indicate that English-to-Japanese systems may require more improvement than their counterparts, judging from the scores given and the number of unfound words recorded.

Keywords

MT evaluation techniques and results; A case study of MT evaluation.

Part 1: Evaluation Environment

1. Evaluation Goals

This project aimed to evaluate the **translation quality** of leading machine translation systems on the Japanese market by following the in-house evaluation guidelines created in June 2000. The language pairs tested include English-to-Japanese and Japanese-to-English. Another purpose was to assess the translation quality of L&H's Japanese-English MT systems in comparison with competitors' products.

The evaluation team consisted of the following five people: Coordinator (Maki Darwin); two E-to-J Evaluators, both native speakers of Japanese (Evaluator A and Evaluator B); and two J-to-E Evaluators, including one native speaker of English (Evaluator C) and one equivalent of English speaker, a native speaker of Japanese (Evaluator D). All of them have experience in either E-to-J or J-to-E translation as freelance translators or business personnel. Each evaluator was assigned 160 hours to complete his/her evaluation.

2. MT Systems Evaluated

The project evaluated eight systems for each language pair. They are marked as [EJsys-1]…[EJsys-8] for the E-to-J direction and as [JEsys-1]…[JEsys-8] for the J-to-E direction. Most of the products tested were bi-directional, each direction sharing the same system number (e.g.: [EJsys-8] and [JEsys-8]); in other cases, mono-directional products from the same maker shared the same system number. To attain the greatest possible objectivity, the evaluators were not shown the identities of these products throughout the project; all the systems were shown as code names only, both when the evaluators scored the output and analyzed their evaluation results.

As a rule, the systems were tested in "as is" condition. Therefore, the **default settings** for their translation output remained unless a setting contradicted the purpose of the project (e.g.: Use "Document Type-General" instead of the default setting "Document Type-Technical"). Some systems might have performed better if they had had more favorable settings. Default settings, however, are the conditions that the developers believe work best for most cases and that are shipped with the products. Thus, this evaluation project also tested the systems in such an environment.

3. Nature and Presentation of Test Materials

300 sentences (60 paragraphs, each containing 5 sentences) from general "real-world" text (newspaper articles, business documents, letters, etc.) were used as test materials for each language pair. Some sentences were short; others were long and complicated.

In preparation, each source paragraph was translated by the systems. Then the translation by one system was divided into five sentences. They were mixed with the sentences translated by the other systems in the same direction. All the target sentences (along with their source sentences) from one source paragraph were shuffled and anonymously listed in random order in a spreadsheet. For example, Source Sentence 1 would list eight different translations, coded as Trans-1...Trans-8, which did not correspond to [EJsys-1]…[EJsys-8] or [JEsys-1]…[JEsys-8]. In addition, each spreadsheet had a different "random" order of source sentences. Since the test materials were presented to the evaluators in spreadsheets in this way, it was almost impossible to identify which translation came from which system.

When the evaluators analyzed their evaluation results,

they were shown modified spreadsheets in which translations by the same system were listed together. In this way, the evaluators were able to point out strengths and weaknesses of each system.

The average evaluation pace was 1 minute per sentence for each category (three categories in total). The evaluators wrote their reports during the rest of their assigned hours.

4. Evaluation Categories

Each system's output was evaluated for the following categories described in the in-house guidelines:

- **Intelligibility** (incl. subcategory **Misspellings**)
- **Accuracy** (incl. subcategory **Unfound Words**)
- **Other Issues** to record

Intelligibility is how clear and understandable the target text is. Measured in no more than two readings, Intelligibility was evaluated regarding the coherence of the target sentence, without referring to the source sentence. Note that ambiguity in the target text should not affect its Intelligibility score since the source text itself can be ambiguous.

The following 5-point scale was used to evaluate Intelligibility[1]:

1. Hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence.

2. Masquerades as an intelligible sentence, but actually is more unintelligible than intelligible. Nevertheless, the idea can still be vaguely apprehended. Word choice, syntactic arrangement, and/or alternative expressions are generally bizarre, and critical words may remain untranslated.

3. The general idea is intelligible only after considerable study, but after this study one is fairly confident that he/she understands. Poor word choice, grotesque semantic arrangement, untranslated words, and similar phenomena are present, but constitute mainly "noise" through which the main idea is still perceptible.

4. Generally clear and intelligible, but style and word choice and/or syntactical arrangement are poorer than for translations rated 5. Poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements definitely interfere with full comprehension. Post-editing could leave this in nearly acceptable form.

5. Perfectly or almost perfectly clear and intelligible; may contain minor grammatical or stylistic infelicities, and /or mildly unusual word usage that could, nevertheless, be easily "corrected."

Misspellings (subcategory): Evaluators also recorded any Misspellings they found in the target text.

Accuracy is how precisely the target text conveys the meaning of the source text. Primarily, this was measured by additional information the source sentence provided after the evaluator scored Intelligibility of the target sentence; the higher the amount of information, the lower the level of Accuracy. Additionally, omissions of actual word or sentence segments from source to target were checked.

The following describes the 5-point scale used in the measurement of Accuracy[2]:

1. (Almost) all relevant information has been lost in the translation. Reading the source text makes "all the difference in the world" in comprehending the meaning intended. (A rating of 1 should always be assigned when the translation completely changes or reverses the meaning conveyed by the source text.)

2. The translation lost a large amount of the information conveyed in the original text. Reading the source text contributes a great deal to the clarification of the meaning intended. Correcting sentence structure, words, and phrases significantly changes the reader's impression of the meaning intended, although not enough to change or reverse the meaning completely.

3. The translation lost some information about semantic relationships implied by the sentence structure. Reading the source text may also correct minor misapprehensions about the general meaning of the sentence or the meaning of individual words.

4. Due to one or a few minor errors, the translation did not capture all of the source information with 100% accuracy. Correcting one or two possibly critical meanings, chiefly on the word level or the grammatical level, gives a slightly different "twist" to the meaning conveyed by the translation. The source text adds no new information about sentence structure, however.

5. The translation conveyed every piece of information contained in the source text accurately; reading the source does not enhance the reader's confidence in his/her understanding of the meaning.

Unfound Words (subcategory): These are words that should be translated but are not found in the engine's lexicon. Proper nouns and acronyms are not included in this category. Evaluators also recorded any Unfound Words they found in the target text.

[1] Adapted from the "Scale of Intelligibility" in *Language and Machines: Computers in Translation and Linguistics* (Washington, D.C.: National Academy of Sciences, 1966), p. 69.

[2] Adapted from the "Scale of Informativeness" in *Language and Machines*, p. 70.

Other Issues that the evaluators also recorded include:

- Stock phrases (e.g.: *How do you do?* in English; *Okage sama de* in Japanese)
- Low-level entities (e.g.: dates, times, numbers, salutations, parenthetical material, quoted material)

5. Evaluation Steps

The evaluation project proceeded in the following order:

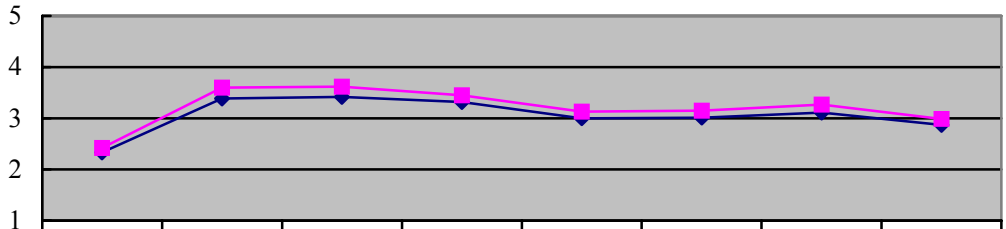
(1) Instructions for Evaluators
(2) Pre-test: As practice, score 20 random sentences one by one for *Intelligibility* and *Accuracy*, for all the systems simultaneously, as well as recording anything noticeable, including *Misspellings* and *Unfound Words*
(3) Feedback from Pre-test
(4) Preliminary Measurement and Rating: Score 100 random sentences in the same method as Pre-test
(5) Analysis and Feedback from Preliminary Measurement and Rating
(6) Full Measurement and Rating: Score 200 random sentences in the same method as Pre-test
(7) Analysis and Feedback from Full Measurement and Rating
(8) Final Analysis: Write *Final Analysis Reports*
(9) Organize the evaluation results [by Coordinator]

Part 2: Evaluation Results

6. Quantitative Results

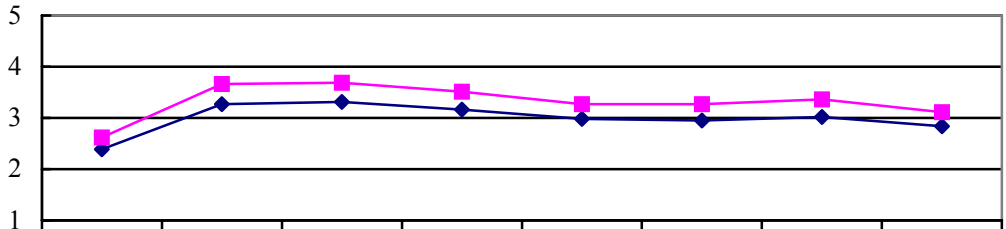
The following Figures and Tables include average **Intelligibility (I)** and **Accuracy (A)** scores for each system. Figures 1 and 4 show the overall average scores by each system, combining results by the two evaluators of each language pair. Scores are shown in points (5 is the highest possible score). Figures 2, 3, 5, and 6 show individual evaluators' average scores. Tables 1 – 4 show how individual evaluators scored the test sentences in each evaluation phase.

The Tables indicate that each evaluator gave quite consistent scores in all the evaluation phases. In addition, according to the Figures, the scoring ratio by the two E-to-J evaluators was almost identical, while that by the J-to-E evaluators was close. These phenomena are noteworthy because the test materials were presented randomly and anonymously throughout the project, making it almost impossible to identify the performance of any particular system.



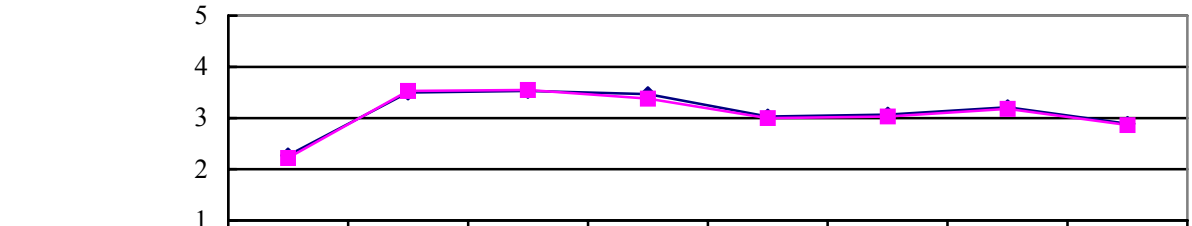
	[EJsys-1]	[EJsys-2]	[EJsys-3]	[EJsys-4]	[EJsys-5]	[EJsys-6]	[EJsys-7]	[EJsys-8]
Intelligibility	2.33	3.39	3.42	3.32	3	3.01	3.11	2.87
Accuracy	2.42	3.6	3.62	3.45	3.13	3.15	3.27	2.99

Figure 1: Overall E-to-J Average Scores (Possible Score 5 Points)



	[EJsys-1]	[EJsys-2]	[EJsys-3]	[EJsys-4]	[EJsys-5]	[EJsys-6]	[EJsys-7]	[EJsys-8]
Intelligibility	2.39	3.27	3.31	3.16	2.98	2.95	3.02	2.84
Accuracy	2.62	3.66	3.69	3.51	3.27	3.27	3.36	3.11

Figure 2: E-to-J Average Scores by Evaluator A (Possible Score 5 Points)



	[EJsys-1]	[EJsys-2]	[EJsys-3]	[EJsys-4]	[EJsys-5]	[EJsys-6]	[EJsys-7]	[EJsys-8]
Intelligibility	2.27	3.5	3.53	3.47	3.03	3.07	3.21	2.89
Accuracy	2.22	3.53	3.55	3.38	3	3.03	3.18	2.86

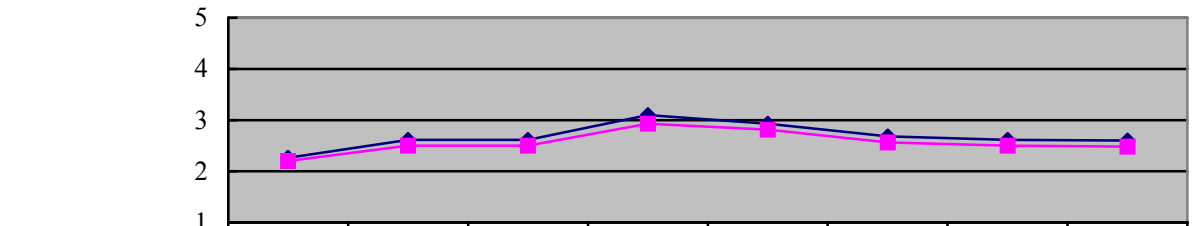
Figure 3: E-to-J Average Scores by Evaluator B (Possible Score 5 Points)

	[EJsys-1]		[EJsys-2]		[EJsys-3]		[EJsys-4]		[EJsys-5]		[EJsys-6]		[EJsys-7]		[EJsys-8]	
	I	A	I	A	I	A	I	A	I	A	I	A	I	A	I	A
Sen# 1-100	2.38	2.62	3.25	3.56	3.3	3.54	3.14	3.48	3.1	3.29	2.97	3.26	3.08	3.33	2.81	3.04
Ranking	8	8	2	1	1	2	3	3	4	5	6	6	5	4	7	7
Sen# 101-200	2.67	2.83	3.53	3.87	3.58	3.91	3.32	3.65	3.17	3.45	3.17	3.53	3.33	3.69	3.14	3.43
Ranking	8	8	2	2	1	1	4	4	5	6	5	5	3	3	7	7
Sen# 201-300	2.11	2.41	3.02	3.54	3.05	3.61	3.01	3.4	2.67	3.06	2.71	3.02	2.65	3.07	2.56	2.86
Ranking	8	8	2	2	1	1	3	3	5	5	4	6	6	4	7	7
All 300	2.39	2.62	3.27	3.66	3.31	3.69	3.16	3.51	2.98	3.27	2.95	3.27	3.02	3.36	2.84	3.11
Ranking	8	8	2	2	1	1	3	3	5	6	6	5	4	4	7	7

Table 1: E-to-J Average Scores by Evaluator A (Phase by Phase)

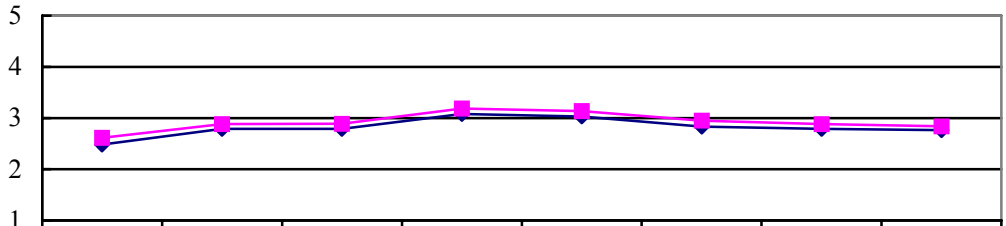
	[EJsys-1]		[EJsys-2]		[EJsys-3]		[EJsys-4]		[EJsys-5]		[EJsys-6]		[EJsys-7]		[EJsys-8]	
	I	A	I	A	I	A	I	A	I	A	I	A	I	A	I	A
Sen# 1-100	1.91	1.76	3.15	3.08	3.08	2.98	3.08	2.87	2.73	2.55	2.78	2.65	2.83	2.75	2.48	2.39
Ranking	8	8	1	1	2	2	2	3	6	6	5	5	4	4	7	7
Sen# 101-200	2.65	2.6	3.86	3.86	3.89	3.9	3.74	3.6	3.32	3.29	3.42	3.35	3.59	3.53	3.31	3.22
Ranking	8	8	2	2	1	1	3	3	6	6	5	5	4	4	7	7
Sen# 201-300	2.25	2.29	3.5	3.66	3.61	3.77	3.6	3.68	3.03	3.15	3.02	3.09	3.2	3.25	2.89	2.97
Ranking	8	8	3	3	1	1	2	2	5	5	6	6	4	4	7	7
All 300	2.27	2.22	3.5	3.53	3.53	3.55	3.47	3.38	3.03	3	3.07	3.03	3.21	3.18	2.89	2.86
Ranking	8	8	2	2	1	1	3	3	6	6	5	5	4	4	7	7

Table 2: E-to-J Average Scores by Evaluator B (Phase by Phase)



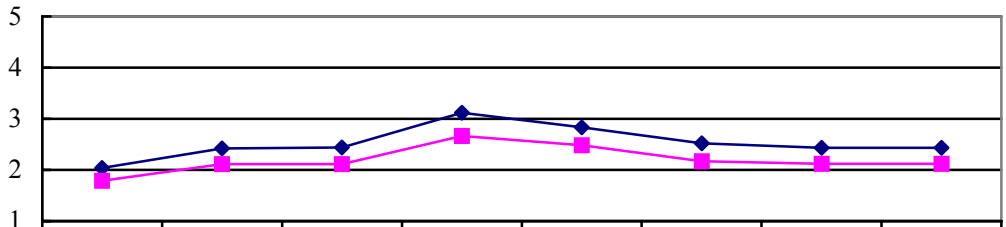
	[JEsys-1]	[JEsys-2]	[JEsys-3]	[JEsys-4]	[JEsys-5]	[JEsys-6]	[JEsys-7]	[JEsys-8]
Intelligibility	2.26	2.61	2.61	3.1	2.93	2.68	2.61	2.6
Accuracy	2.2	2.5	2.5	2.93	2.81	2.56	2.5	2.48

Figure 4: Overall J-to-E Average Scores (Possible Score 5 Points)



	[JEsys-1]	[JEsys-2]	[JEsys-3]	[JEsys-4]	[JEsys-5]	[JEsys-6]	[JEsys-7]	[JEsys-8]
Intelligibility	2.48	2.79	2.79	3.08	3.03	2.83	2.79	2.76
Accuracy	2.61	2.88	2.89	3.19	3.14	2.95	2.88	2.84

Figure 5: J-to-E Average Scores by Evaluator C (Possible Score 5 Points)



	[JEsys-1]	[JEsys-2]	[JEsys-3]	[JEsys-4]	[JEsys-5]	[JEsys-6]	[JEsys-7]	[JEsys-8]
Intelligibility	2.04	2.42	2.44	3.12	2.83	2.52	2.43	2.43
Accuracy	1.79	2.11	2.11	2.66	2.48	2.17	2.12	2.12

Figure 6: J-to-E Average Scores by Evaluator D (Possible Score 5 Points)

	[JEsys-1]		[JEsys-2]		[JEsys-3]		[JEsys-4]		[JEsys-5]		[JEsys-6]		[JEsys-7]		[JEsys-8]	
	I	A	I	A	I	A	I	A	I	A	I	A	I	A	I	A
Sen# 1-100	2.37	2.57	2.81	2.86	2.81	2.87	3.08	3.24	2.99	3.05	2.81	3	2.81	2.87	2.77	2.84
Ranking	8	8	3	6	3	4	1	1	2	2	3	3	3	4	7	7
Sen# 101-200	2.59	2.64	2.74	2.72	2.74	2.74	3.07	3.09	3.12	3.2	2.87	2.89	2.74	2.71	2.7	2.71
Ranking	8	8	4	5	4	4	2	2	1	1	3	3	4	6	7	6
Sen# 201-300	2.48	2.61	2.82	3.07	2.82	3.06	3.1	3.25	2.97	3.18	2.82	2.95	2.82	3.06	2.8	2.98
Ranking	8	8	3	3	3	4	1	1	2	2	3	7	3	4	7	6
All 300	2.48	2.61	2.79	2.88	2.79	2.89	3.08	3.19	3.03	3.14	2.83	2.95	2.79	2.88	2.76	2.84
Ranking	8	8	4	5	4	4	1	1	2	2	3	3	4	6	7	7

Table 3: J-to-E Average Scores by Evaluator C (Phase by Phase)

	[JEsys-1]		[JEsys-2]		[JEsys-3]		[JEsys-4]		[JEsys-5]		[JEsys-6]		[JEsys-7]		[JEsys-8]	
	I	A	I	A	I	A	I	A	I	A	I	A	I	A	I	A
Sen# 1-100	2.07	1.75	2.49	2.2	2.49	2.2	3.19	2.72	2.64	2.39	2.55	2.15	2.5	2.2	2.47	2.17
Ranking	8	8	5	3	5	3	1	1	2	2	3	7	4	3	7	6
Sen# 101-200	2.22	1.99	2.31	2.01	2.31	2.01	3.07	2.76	2.96	2.72	2.53	2.2	2.29	2.02	2.32	2.02
Ranking	8	8	5	6	5	6	1	1	2	2	3	3	7	4	4	4
Sen# 201-300	1.82	1.63	2.47	2.12	2.51	2.12	3.1	2.5	2.88	2.32	2.47	2.15	2.5	2.13	2.51	2.17
Ranking	8	8	6	6	3	6	1	1	2	2	6	4	5	5	3	3
All 300	2.04	1.79	2.42	2.11	2.44	2.11	3.12	2.66	2.83	2.48	2.52	2.17	2.43	2.12	2.43	2.12
Ranking	8	8	7	6	4	6	1	1	2	2	3	3	6	5	5	4

Table 4: J-to-E Average Scores by Evaluator D (Phase by Phase)

7. Highlights of Analysis

The evaluators found that, regardless of translation direction, stronger systems tended to be superior and weaker systems tended to be inferior in the following areas:

- Source sentence analysis
- Target sentence arrangement
- Word choice
- Lexicon entries

As for language-pair comparison, the English-to-Japanese systems received higher scores than their counterparts on the surface. However, we cannot simply conclude that the Japanese-to-English systems were inferior because each direction used different evaluators; one pair may have been more critical. Nevertheless, the J-to-E systems generally require more improvement, especially in their lexicons, because the evaluators recorded a substantial number of Unfound Words.

Conclusion

Although this evaluation project produced worthwhile results, including consistent scores and a comprehensive analysis, its methods and techniques leave room for discussion and improvement.

The makeup of the evaluation team. (1) The number of evaluators: Was employing two evaluators for each direction ideal? Had we had more than two evaluators for each direction, would we still have had consistent results? (2) Qualifications of an evaluator: It may have been more ideal to have two native speakers of English for the J-to-E evaluation. The time for preparation was limited, so we instead hired one English speaker and one Japanese speaker, who, we decided, had enough qualifications for the task.

Evaluation guidelines. (1) The 5-point scales: Was the 5-point scale ideal for each category? Had it been a 3-point scale, would it have been easier for the evaluators? Or, what if it would have been a 6-point scale (or lager number)? (2) Terminology: The evaluators sometimes found it difficult to tell to what issue a particular phenomenon belonged. It was true that there had been always gray areas and the guidelines did not discuss the terminology in-depth. Thus, it was often up to the evaluator to decide what issue covered a particular phenomenon. Related to this concern, one evaluator wrote in his feedback that he felt a trained linguist might have done a better job than a translator in order to more closely understand the descriptions of the evaluation categories. "Yes and no" to his comment, because in order to evaluate output of MT systems, an evaluator must know translation as well. (3) Unfound Word vs. Wrong Word Choice: It was often difficult to tell whether a strange translation was a result of the system lexicon's not containing a word, or merely a wrong word choice. Since the guidelines did not discuss it thoroughly, decision was up to the evaluator's judgment.

Evaluation schedule. (1) Evaluation pace: Another evaluator felt the evaluation pace (average: 1 minute per sentence for each evaluation category) was too fast. She felt she needed more time to write comments.

In any case, the project served as an interesting case study and a worthwhile experiment for the in-house guidelines. For future evaluation projects, we will include the criticism and comments by the evaluators to improve the methods and techniques of the evaluation.

Acknowledgements

My special thanks go to Monika Forner, who encouraged me to submit the paper and gave me critical advice on how it should be presented.

Reference

Pierce, John R., and Carroll, John B., et al. (1966). *Language and Machines: Computers in Translation and Linguistics* (A Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council). Washington, D.C.: National Academy of Sciences.