Towards Fairer Evaluations of Commercial MT Systems on Basic Travel Expressions Corpora

Hervé Blanchon†, Christian Boitet†, Francis Brunet-Manquat†,
Mutsuko Tomokiyo†, Agnès Hamon‡, Vo Trung Hung†, Youcef Bey†

† Laboratoire CLIPS
BP 53
38041 Grenoble Cedex 9, France
{first.last}@imag.fr

‡ Laboratoire de Statistique (Université Haute Bretagne)
Place du recteur H. Le Moal, CS 24307
35043 Rennes Cedex, France
Agnès.Hamon@uhb.fr

Abstract

We compare the performance of several SYSTRAN systems on the BTEC corpus. Two language pairs: Chinese to English and Japanese to English are used. Whenever it is possible the system will be used "off the shelf" and then tuned.
The first system we use is freely available on the web. The second system, SYSTRAN Premium, is commercial. It is used in two ways: (1) choosing and ordering available original dictionaries and setting parameters, (2) same + user dictionaries.
As far as the evaluation is concerned, we competed in the unlimited data track.

1. Introduction

We first give our motivations for participating in this campaign with a commercial system. Then we briefly describe the system tested (SYSTRAN® 5.0) and the ways it can be parametrized. The bulk of this paper describes the evaluation procedure. We finish by presenting and analyzing the results.

2. Rationale

MT evaluation is a hot topic since 1960 or so. The literature on evaluation may even be larger than that on MT techniques proper. MT evaluation may have several goals:

- help buyers buy MT (or CAT) system best suited to their needs,
- help funders decide on which technology to support, and
- help developers measure various aspects of their systems, and measure progress.

The MT evaluation campaign organized by the C-STAR III consortium falls in the latter category. Its aim is to measure the "quality" of various MT systems developed for speech-to-speech translation when applied to the BTEC corpus [7]. Another goal is to compare the MT systems developed by the C-STAR partner not only between them, but also with other systems, notably commercial systems.
In the past, we often got the impression that, in similar campaigns, the commercial system used as a "baseline" were tested in quite biased ways. From what was reported, the experimenters submitted the input texts to free MT web servers, and evaluated the results. But that method is quite unfair, which makes all the conclusions scientifically invalid.
For example, long ago, the CANDIDE system trained intensively on the Hansard corpus, was compared with an off-the-shelf version of SYTSRAN without any tuning. SYSTRAN clearly won, but the margin might have been far bigger (or perhaps not, this should have been studied!), if SYSTRAN had been tuned to this totally unseen corpus, at the level of dictionaries, of course, but perhaps also of grammars.
Another example is given by MSR [5] on comparison between their French-English system, highly tuned to their documents (actually, the transfer component was 100% induced from 150000 pairs of sentences and their associated "logical forms" or deep syntactic trees). They used also SYSTRAN, this time slightly tuning it by giving priority to SYSTRAN dictionaries containing computer related terms[1]. However, they apparently did not invest time to produce a user dictionary containing the MicroSoft computer science dictionary. Technical terminology varies a lot from firm to firm and even from product to product. What is then the value of the conclusion that their system was (slightly) better than SYSTRAN? And when they tried to do the same comparison on the Hansard, SYSTRAN ("general") won.
As members of C-STAR III not engaged in developing J-E or C-E systems (although we worked on prototypes in the 80's), we felt it was interesting to take part in this evaluation campaign to establish a "most faithful baseline" for a commercial system.
Which commercial system(s) to use? We choose SYSTRAN because:

- it offers the two pairs C-E and J-E,
- these pairs have been recently slightly improved (although more work has been done on E-C and E-J),
- SYSTRAN agreed to give us free access to the latest version (v5), in its Premium packaging (Windows interface, with many tunable parameters, and the possibility to create a user dictionary),
- it can be considered as a kind of "medium" baseline, when compared with the other commercial MT systems for C-E and J-E (some are far worse, and some far better, e.g. ATLAS-II for E-J and ALT/JE for J-E).

There is another worry with the current evaluation campaigns: objective evaluation is performed using "reference" human translations, but the measures, based on n-gram co-occurrence counts, correlate only (very) weakly with human judgment [9], and that human judgment is too often not task oriented. As a result:

- the evaluations produce tables of figures with no decisive, clear interpretation about intrinsic quality (relative to some precise goal);
- real translation results are never shown side by side for subjective evaluation by the readers themselves;
- no task-oriented measure is computed.

To compute such a measure, one should measure the time it takes a human to produce a "reference translation" from an MT output, as done by [8].

[1] This is not in the paper but what answered to a question.

As a byproduct, one also gets new reference translations, for cheaper than with usual human translation: o, 510 sentences of the BTEC, the second author spent about 12 mn per page (of 250 words) using SYSTRAN English-French output while 3 colleagues each spent 59 mn per page usig no machine help.
If the goal is to produce good translations, the intended use of the MT system is to help human translators, and that measure is perfect. If the goal is to help readers understand text in a foreign language, it is also a very good indicator, if the human "judge" is asked not to look at the source text before having really tried (hard) to understand the MT output.
In the framework of this campaign, we worked only on the first aspect (how to test a commercial system as faithfully as possible) and not on the second (how to improve the evaluation itself), although we produced parallel presentations of source (J/C), reference (E), and MT outputs for human direct inspection and subjective, global evaluation.
Let us now describe in more detail the SYSTRAN systems used, before moving to the evaluation protocol and its results.

3. SYSTRAN systems used

3.1. Version and usage

The SYSTRAN systems involved are:
- SYSTRAN 5.0, freely available on the web
- SYSTRAN 5.0 tuned by some parameters settings, and with a user dictionary

Those systems use there own linguistic resources, we took then part in the unlimited track of the evaluation.

3.2. Language pairs considered:

- Japanese to English
- Chinese to English

Those are by far not the best SYSTRAN pairs. They have been slightly improved from earlier ones in the context of a side project with CISCO, the main project concerning English to Chinese, Japanese, and Koréan. However, using Systran output to help human translation still gives a clear productivity increase, as a sizable pat of the translations need 1 or 0 changes only.

3.3. SYSTRAN systems linguistic components

SYSTRAN Linguisic components comprise three modules:
1. Source Language Analysis
2. Language Pair Transfer
3. Target Language Synthesis

3.3.1. Source language analysis

The morphosyntactic analysis module examines each sentence in the text input, noting all uncertainties and errors. This examination allows for reanalysis and decision-making on alternate translations in later processing.
The program flow and basic algorithms for the syntactic analysis module is essentially the same for all systems sharing the same source language, and the system design and architecture are the same for all language pairs. However, in the case of lexical and syntactic ambiguities, decisions are often taken with respect to the target language.

3.3.2. Language pair transfer

It is the only module that is unique to each language pair. It restructures the syntactic structure (a kind of chart) as necessary, and selects the correct target lexical equivalents of identified words and expressions. Regardless of the fact that restructuring and selection are different, the basic architecture and strategy are similar for all language pairs.
That architecture is a kind of "descending transfer" [1], because the only independent phase in generation is the morphological generation (there are actually very few real "horizontal" transfer systems).

3.3.3. Target language synthesis

The module makes all necessary assignments of case, tense, number, etc. according to the rules of the target language in order to generate the target language output.

3.3.4. Dictionaries

Two dictionaries are used: a stem dictionary and an expression dictionary. The stem dictionary contains terminology and base forms. An expression dictionary contains phrases and conditional expressions.
A dictionary manager tool provides a mean for improving translation results through the formation of multilingual dictionaries. Multilingual dictionaries are user-created collections of subject-specific terms that are analyzed prior to being integrated directly into the translation process.

4. Evaluation protocol

4.1. Tuning SYSTRAN

The dictionaries of the SYSTRAN Japanese to English and Chinese to English were updated in the following way.

4.1.1. Choosing the SYSTRAN dictionaries available

We choose two *original* dictionaries of the SYSTRAN premium 5.0: the *business* and *colloquial language* dictionaries.

4.1.2. Dictionary update for the Chinese to English system

SYSTRAN system with original dictionaries found 178 unknown words in the Chinese training corpus. So, we created a Chinese *user* dictionary containing these words and their English translation. Furthermore, the SYSTRAN system associated with this user dictionary found 4 unknown words in the test corpus. These words were added to the user dictionary. This final user dictionary is used to improve the SYSTRAN premium 5.0 system afterward.

4.1.3. Dictionary update for the Japanese to English system

We also create a Japanese user dictionary with the same method. We found 304 unknown words in the Japanese training corpus and 13 unknown words in the Japanese test corpus. The new Japanese user dictionary is also used to improve the SYSTRAN premium 5.0 system afterward.

4.2. Evaluation methods

4.2.1. Subjective evaluation

Subjective evaluation was conducted using the NIST protocol[2]. Both *fluency* and *adequacy* were evaluated with a set of three judges.
For the translation of each sentence, judges make the *fluency* judgment before the *adequacy* judgment. *Fluency* refers to the degree to which the target is well formed according to the

[2] http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess02.pdf

rules of Standard Written English. A fluent segment is one that is well-formed grammatically, contains correct spelling, adheres to common use of terms, titles and names, is intuitively acceptable, and can be sensibly interpreted by a native speaker of English. A *fluency* judgment is one of the following:

	How do you judge the fluency of this translation? It is:
5	Flawless English
4	Good English
3	Non-native English
2	Disfluent English
1	Incomprehensible

Where English translations retain source language characters or words, judges are instructed to give a score between "1: Incomprehensible" and "3: Non-native English" depending upon the degree to which the untranslated characters, among the other factors, affect the *fluency* of the translation.
Having made the *fluency* judgment for a translation of a segment, the judge is presented with one of four reference translations. Comparing the target translation against the reference translation, judges determine whether the translation is adequate. *Adequacy* refers to the degree to which information present in the original is also communicated in the translation. Thus for *adequacy* judgments, the reference translation will serve as a proxy for the original source language text. An *adequacy* judgment is one of the following:

	How much of the meaning expressed in the gold-standard translation is also expressed in the target translation?
5	All
4	Most
3	Much
2	Little
1	Non

Where English translations retain Chinese and or Japanese characters from the original sentences, judges are instructed to give a score between "1: None" and "4: Most" depending upon the degree to which the un-translated characters, among the other factors, affect the *adequacy* of the translation.
A simple statistical approach to quantify inter-judge agreement and concordance for the three pairs of judges is to compute the Cohen Kappa and Gamma coefficients [6]. The overall agreement between the three judges is assessed with an extension of the Kappa. The extension of the Kappa coefficient to evaluate agreement among the three judge is based on the assumption of identical marginal ratings, that is all the judge have the same level of grading severity. A valid interpretation of the overall Kappa coefficient is only possible when this hypothesis holds. Indeed, if the raters do not use the same scores in the same proportions, substantial agreement cannot be expected.

4.2.2. Objective evaluation

Five automatic scoring techniques have been used: BLEU, NIST, WER, PER, and GTM.

BLEU

One first compute a modified n-gram precision p_n ($1 \le n \le N$)

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count(n\text{-}gram)} \quad \text{(B 1)}$$

where $Count_{clip}(n\text{-}gram)$ is the maximum number of *n-grams* co-occurring in a candidate translation and a reference translation, and $Count(n\text{-}gram)$ is the number of *n-grams* in the candidate translation. In order to prevent very short translations to try to maximize their precision scores a brevity penalty, *BP*, is used:

$$BP = \left\{ \begin{array}{ll} 1 & if\ |c| > |r| \\ e^{(1-|r|/|c|)} & if\ |c| \le |r| \end{array} \right\} \quad \text{(B 2)}$$

Here |c| is the length of the candidate translation and |r| is the length of the reference translation. Then:

$$BLEU = BP \bullet \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad \text{(B 3)}$$

The weighting factor, w_n, is set at 1/N.
[2] quoted several limits of BLEU. First, the geometric mean of co-occurrence over n induces a counterproductive variance due to low co-occurrences for the larger values of n. In other words, a lack of long n-gram match will have a strong impact on the score. Second, it may be better to weight more heavily the more informative n-grams. Those n-grams are those that occur less frequently. Third, small variations of translation length impact sensibly the score. This impact should be minimized with a different Brevity Penalty.

NIST

The NIST score implements these proposals.
The Information weights are computed as follows:

$$Info(w_1, \ldots, w_n) = \log_2 \left(\frac{the\ \#\ of\ occurrences\ of\ w_1, \ldots, w_{n-1}}{the\ \#\ of\ occurrences\ of\ w_1, \ldots, w_n} \right) \quad \text{(N 1)}$$

An new brevity penalty is introduced to minimize the impact on score of small variations in the length of the translation.

$$BP_{NIST} = \exp\left\{\beta \log^2\left[\min\left(\frac{L_{sys}}{\overline{L}_{ref}}, 1\right)\right]\right\} \quad \text{(N 2)}$$

The NIST score is computed as follows:

$$NIST = BP_{NIST} \bullet \sum_{n=1}^{N} \left\{ \frac{\sum_{all\ w_1, \ldots, w_n\ that\ co\text{-}occur} Info(w_1, \ldots, w_n)}{\sum_{all\ w_1, \ldots, w_n\ in\ sys\ output} (1)} \right\} \quad \text{(N 3)}$$

where

$$L_{sys} = \left[\begin{array}{l} the\ number\ of\ words\ in\ the \\ translation\ being\ scored \end{array} \right.$$

$$\overline{L}_{ref} = \left[\begin{array}{l} the\ average\ number\ of\ words\ in\ a \\ reference\ translation\ averaged\ over \\ all\ reference\ translations \end{array} \right.$$

$$\beta = -4.22\ so\ that\ BP = 0.5\ when\ L_{sys} = 2/3 \cdot \overline{L}_{ref}$$

WER/PER

The WER is the standard technique used to evaluate speech recognition modules based on the Levenstein edit distance [4]. [3] proposed a Levenstein-like measure independent form the words position. (PER – Position-independent Error Rate) ; A sentence is seen as a bag of words and the distance between a sentence and any of its permutations is null[3].

[3] This measure is thus not a distance.

GTM

At the sentence level, the GMT [9] score is the harmonic mean (F-measure) of a new proposed precision and recall measures based on a maximum match size.

4.3. Evaluation Parameters:

- case insensitive (lower-case only)
- no punctuation marks (remove '.' ',' '?' '!' '"') ; periods that are parts of a word should not be removed, e.g., abbreviations, like "mr.","a.m.", remain as they occur in the corpus data
- no word compounds (substitute hyphens '-' with blank space)
- spelling-out of numerals

5. Results

5.1. Submitted runs

5.1.1. Chinese to English

We submitted three runs for the Chinese-to-English language pair. These runs were produced by SYSTRAN :

C_1	SYSTRAN web 5.0
C_2	SYSTRAN premium 5.0 with original dictionaries
C_3	SYSTRAN premium 5.0 with original and user dictionaries

5.1.2. Japanese to English

We submitted four runs for the Japanese-to-English language pair. Three runs were produced by SYSTRAN :

J_1	SYSTRAN web 5.0
J_2	SYSTRAN premium 5.0 with original dictionaries
J_3	SYSTRAN premium 5.0 with original and user dictionaries

The last run was made of the original translations produced for the **J_3** run revised by a human translator instructed to produce an adequate translation out of the SYSTRAN English translation minimizing the changes. Out of the 500 utterances, 50 (10%) were left unchanged.

J_4	**J_3** manually revised translation

5.2. Objective evaluation results for C-E

	BLEU	GMT	NIST	PER	WER
C_3	0.1620 **1**	0.5845 **1**	6.0061 **1**	0.5429 **2**	0.6581 **2**
C_1	0.1600 **3**	0.5802 **3**	5.9143 **3**	0.5423 **1**	0.6474 **1**
C_2	0.1620 **1**	0.5841 **2**	6.0039 **2**	0.5429 **2**	0.6581 **2**

Table 1: Objective evaluation results for the CLIPS C-E runs

5.3. Objective evaluation results for J-E

	BLEU	GMT	NIST	PER	WER
J_3	0.1320 **1**	0.5687 **1**	5.6476 **1**	0.5978 **1**	0.7304 **1**
J_2	0.1311 **2**	0.5672 **2**	5.6096 **2**	0.6012 **2**	0.7349 **2**
J_1	0.0810 **3**	0.5116 **3**	4.1935 **3**	0.7179 **3**	0.8726 **3**

Table 2: Objective evaluation results for the CLIPS J-E runs

5.4. Objective evaluation results for J_4

We were expecting far better results with the revised translations. The results confirmed our intuition.

	BLEU	GMT	NIST	PER	WER
J_4	0.4691	0.7777	9.9189	0.3236	0.3711

Table 3: Objective evaluation results for the CLIPS J_4 run

5.5. Subjective and objective competitive evaluation results for C-E

9 systems took part in the competitive evaluation. SYSTRAN is the **C_1** system with original and user dictionaries.

5.5.1. Subjective evaluation

	Fluency	Adequacy
CE_8	3.7760 **1**	3.6620 **1**
CE_3	3.0360 **4**	2.9960 **6**
CE_7	2.9340 **6**	3.2540 **3**
CE_5	3.7760 **1**	3.5260 **2**
CE_9	3.4000 **3**	2.8000 **8**
CE_2	2.6480 **8**	3.1880 **4**
CE_6	2.9540 **5**	2.7840 **9**
C_1	**2.5700** 9	**2.9600** 7
CE_4	2.7180 **7**	3.0820 **5**

Table 4: Subjective evaluation for the C-E Unlimited runs submitted by all participants

5.5.2. Objective evaluation

	BLEU	GMT	NIST	PER	WER
CE_8	0.5249 **1**	0.7482 **1**	9.5603 **1**	0.3198 **1**	0.3795 **1**
CE_3	0.3505 **3**	0.6849 **2**	7.3691 **3**	0.4428 **4**	0.5255 **3**
CE_7	0.2753 **5**	0.6669 **4**	7.5002 **2**	0.4276 **3**	0.5313 **4**
CE_5	0.4409 **2**	0.6720 **3**	7.2413 **4**	0.3930 **2**	0.4570 **2**
CE_9	0.3113 **4**	0.5639 **8**	5.9217 **7**	0.5310 **7**	0.5788 **6**
CE_2	0.2438 **6**	0.6119 **5**	6.1354 **5**	0.4872 **5**	0.5941 **7**
CE_6	0.2430 **7**	0.6023 **6**	5.4250 **8**	0.4998 **6**	0.5735 **5**
C_1	**0.1620** 8	**0.5845** 7	**6.0061** 6	**0.5429** 8	**0.6582** 8
CE_4	0.0798 **9**	0.3862 **9**	3.6443 **9**	0.7650 **9**	0.8466 **9**

Table 5: Objective evaluation for the C-E Unlimited runs submitted by all participants

5.6. Subjective and objective competitive evaluation results for J-E

4 systems took part in the competitive evaluation scheme. SYSTRAN is the **J_2** system with original and user dictionaries.

5.6.1. Subjective evaluation

	Fluency	Adequacy
JE_1	4.3080 **1**	4.2080 **1**
JE_3	4.0360 **2**	4.0660 **2**
JE_4	3.6500 **3**	3.3160 **3**
J_3	**2.4720** 4	**2.6020** 4

Table 6: Subjective evaluation for the J-E Unlimited runs submitted by all participants

5.6.2. Objective evaluation

	BLEU	GMT	NIST	PER	WER
JE_1	0.6306 **1**	0.7967 **2**	10.7201 **2**	0.2333 **1**	0.2631 **1**
JE_3	0.6190 **2**	0.8243 **1**	11.2541 **1**	0.2492 **2**	0.3056 **2**
JE_4	0.3970 **3**	0.6722 **3**	7.8893 **3**	0.4202 **3**	0.4857 **3**
J_3	**0.1320** 4	**0.5687** 4	**5.6476** 4	**0.5978** 4	**0.7304** 4

Table 7: Objective evaluation for the J-E Unlimited runs submitted by all participants

	BLEU	GMT	NIST	PER	WER
JE_1	0.6306 **1**	0.7967 **2**	10.7201 **2**	0.2333 **1**	0.2631 **1**
JE_3	0.6190 **2**	0.8243 **1**	11.2541 **1**	0.2492 **2**	0.3056 **2**
J_4	***0.4691*** *3*	***0.7777*** *3*	***9.9189*** *3*	***0.3236*** *3*	***0.3711*** *3*
JE_4	0.3970 **4**	0.6722 **4**	7.8893 **4**	0.4202 **4**	0.4857 **4**
J_3	**0.1320** 5	**0.5687** 5	**5.6476** 5	**0.5978** 5	**0.7304** 5

Table 8: Objective evaluation for the J-E Unlimited runs submitted by all participants and J_4

When comparing the previous results, the ***J_4*** run is ranked as third. That may seems not that high for humane revised translations!

6. Discussion

6.1. Results on C-E

Surprisingly, the web C-E version of SYSTRAN [**C_1**] performed better than SYSTRAN Premium 5.0 [**C_2**] (Table 1). This is because the Premium version was frozen in May-June, while the web version was updated later and keeps changing. With this evaluation, its rank is 8th/9 (Table 4 and Table 5).

6.2. Results on C-J

Runs **J_3**, **J_2**, and **J_1** ranks are in agreement with the intuition that the better the system is tuned, the better are the results (Table 2). With this evaluation, the rank of **J_3** is 4th/9 (Table 6 and Table 7).
Surprisingly, (perfect) human revised translation (**J-4**, Table 8) is ranked only 3rd and does not reach the first position! This proves that the measure or the evaluation method is flawed.

6.3. Inter-judge agreement in the subjective evaluation for both C-E and J-E language pairs.

All the Gamma coefficients (Table 9, third column) are greater than 0.6 indicating no disagreement in the ordering of ratings. Before computing agreement, we informally checked the assumption of identical marginal ratings among the three raters (not all the results are reported here). For the two evaluations concerning *fluency*, it appears that raters number 2 use the score 2 (Disfluent English) in about 46% of the ratings while the two others did not show such a behavior.
These two studies on *fluency* imply that it is not possible to interpret the value of the overall Kappa. For the two evaluations about *adequacy*, there were no notable differences among the marginal ratings.

		Gamma	Kappa
CE fluency	R1 & R2	0.803	0.384
	R2 & R3	0.703	0.197
	R1 & R3	0.712	0.317
	Overall		------
CE adequacy	R1 & R2	0.720	0.318
	R2 & R3	0.686	0.305
	R1 & R3	0.656	0.308
	Overall		0.309
JE fluency	R1 & R2	0.745	0.345
	R2 & R3	0.647	0.203
	R1 & R3	0.645	0.272
	Overall		------
JE adequacy	R1 & R2	0.724	0.320
	R2 & R3	0.747	0.340
	R1 & R3	0.695	0.298
	Overall		0.318

Table 9: Gamma and Kappa values for the inter-judge agreement evaluation

The pairwise Kappa values (last column) are between 0.19 and 0.38 indicating a moderate agreement between the raters. The overall Kappa values are 0.309 for the evaluation of Chinese-English *adequacy* and 0.318 for that of Japanese-English. These values indicate a moderate agreement between the 3 raters.

This first evaluation of agreement points out the concordance between judges on the ordering but a moderate agreement on the rating.

6.4. Analysis of SYSTRAN-produced translations[4]

All the Japanese utterances can be considered as polished transcribes of oral dialogues in the tourism domain. The language level is rather polite.
When the utterance is euphemistic (が), the particle is always translated by "but".
Some of the utterances do not make sense without any context (e.g. 切りますよ。➟ "it cuts" ?).
When the first person subject is omitted in Japanese, it is always translated as "it" (ここで降ります。➟ "It gets off here.").
The test set contains a lot of interrogative utterances. In the translations, the interrogative pronoun or adverb is always shifted at the end of the translation. The standard English word order is not respected (e.g. オペラ座はどこですか。➟ "Is the opera house where?").
A lot of spoken Japanese daily life idiomatic expressions are not present in the SYSTRAN dictionaries (e.g. どういたしまして。➟ "How doing." もしもし。➟ "It does." さようなら。➟ "Way if.").
Requests or invitations are not always well translated (e.g. 注文したいのです。➟ "It is to like to order." 一緒に行きましょう。➟ "It will go together.").
Lexicalized Japanese politeness is correctly analyzed (e.g. そのまま切らずにお待ち下さい。➟ "Without cutting that way, please wait.").
When the valency of the verb for two expressions in Japanese and English is different, the translation is almost always wrong (e.g. 寒気がする。➟ "Chill does.").
Finally, the aspect of the japanese predicate is not correctly rendered in English (e.g. 航空券を家に忘れてしまいました。➟ "The air ticket was forgotten in the house.").

7. Conclusion

Adding entries for unknown words in the SYSTRAN dictionary was not sufficient to raise the performance of the system at a score comparable with that of the other competing systems.
However, ranking of (perfect) translation obtained by postediting SYSTRAN outputs was only 3/4, which indicates that the evaluation method or experimental methodology used in the campaign may be flawed.
To investigate this, it will be absolutely necessary in future evaluation campaigns to produce side by side presentations of the various results in order to let a human compare the results.
Finally, we have observed, as always, that the rank of a system is not consistent over several objective evaluation techniques
As far as subjective evaluation is concerned, we have shown that agreement between the judges is not good.
With hindsight, it would appear that the definition of adequacy and fluency proposed by NIST are too much geared

[4] The appendix gives some example of real data submitted to the evaluation.

towards written style. In the context of speech-to-speech translation some of the severely evaluated utterances can be considered as perfectly OK.

8. References

[1] Boitet, C. (2003) *Machine Translation.* in Ralston, A., Reilly, E. and Hemmendinger, D. (ed.), Encyclopedia of Computer Science. John Willey & Sons. 10 p.

[2] Doddington, G. (2002) *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics.* Proc. HLT 2002. San Diego, California. March 24-27, 2002. vol. **1**/**1**: pp. 128-132 (note book proceedings).

[3] Leusch, G., Ueffing, N.*, et al.* (2003) *A Novel String-toString Distance Measure with Application to Machine Translation Evaluation.* Proc. MT Summit IX. New Orleans, U.S.A. September 23-27, 2003. 8 p.

[4] Levenshtein, V. I. (1966) *Binary codes capable of correcting deletion, insertions and reversals.* in Soviet Physics Doklady. vol. **10**(8): pp. 707-710.

[5] Pinkham, J. and Smets, M. (2002) *Traduction automatique ancrée dans l'analyse linguistique.* Proc. TALN'02. Nancy, France. 24-27 juin 2002. vol. **1/2**: pp. 287-296.

[6] Siegel, S. and Castellan, N. J. (1988) *Nonparametric Statistics for the Behavioral Sciences; 2nd ed.* McGraw-Hill. New-York. 400 p.

[7] Takezawa, T., Sumita, E.*, et al.* (2002) *Towards a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World.* Proc. LREC-2002. Las Palmas, Spain. May 29-31, 2002. vol. **1/3**: pp. 147-152.

[8] Tomás, J., Mas, J. À.*, et al.* (2003) *A Quantitative Method for Machine Translation Evaluation.* Proc. EACL 2003 – Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resouces reusable? Budapest, Hungary. April 14, 2003. vol. **1/1**: 8 p.

[9] Turian, J. P., Shen, L.*, et al.* (2003) *Evaluation of Machine Translation and its Evaluation.* Proc. MT Summit IX. New Orleans, U.S.A. September 23-27, 2003. pp. 386-393.

9. Appendix

Original	SYSTRAN Output	Revised translation	Comment
お 勘定 を おねがい し ます 。	We request calculation.	Give me the check, please.	Bad lexical choice for 勘定.
この 街 の 見どころ を 教え て ください 。	Please teach the places of interest of this town.	Tell me please the places of interest of this town.	Bad lexical choice for 教える.
あの 角 に あり ます よ 。	There is that angle.	There is one at the corner.	Bad lexical choice for 角. Object missing.
どれ が 私 の です か 。	Either one is my?	Which one is mine?	Question: good word order. Wrong lexical choice for どれ.
入場 料 は いくら です か 。	Is admission fee how much?	How much is the admission fee?	Question: wrong word order.
オペラ 座 は どこ です か 。	Is the opera house where?	Where is the opera house?	Question: wrong word order.
日本語 は 話せ ます か 。	You can speak Japanese?	Can you speak Japanese?	Question: wrong word order. Acceptable.
どういたしまして 。	How doing.	You are welcome.	Spoken fixed formula.
もしもし 。	It does.	Hello.	Spoken fixed formula.
それ に し ます 。	It makes that.	I'll take that.	Spoken fixed formula.
ここ で 降ろし て ください 。	Please lower here.	Drop me off here, please.	Subject: missing. Bad lexical choice for 降ろ.
ここ で 降り ます 。	It gets off here.	I'll get off here.	Subject: missing. Good lexical choice for 降ろ.
一緒 に 行き ましょ う	It will go together.	Let's go together.	Subject: missing.
切符 を 二 枚 持っ て い ます 。	It has two tickets.	I have two tickets.	Subject: missing.
医者 を 呼ん で ください 。	Please call the doctor.	Please call the doctor.	Subject: missing. Very good translation.
お 医者 さん に 行っ た 方 が いい かも よ 。	Whether the person where does to the doctor is good.	It's better to go to see a doctor.	Very bad translation.
薬 が 効き まし た か 。	Was the medicine effective?	Was the medicine effective?	Good translation. Impersonal sentence.
初心者 で も 大丈夫 です か 。	It is all right even with the beginner?	It is all right even with the beginner?	Very good translation. Impersonal sentence.
パスポート 、 トラベラーズチェック 、 航空 券 。	Passport, traveler's check and air ticket.	Passport, traveler's check and air ticket.	Very good translation. Noun phrases.
観光 ツアー は あり ます か 。	Is there a sightseeing tour?	Is there a sightseeing tour?	Very good translation: Impersonal sentence.