05-0111

# Evaluation of Speech-to-Speech Translation Systems

**John Aberdeen**
**Christine Doran**
MITRE
202 Burlington Rd.
Bedford, MA 01730
{aberdeen,cdoran}
@mitre.org

**Beatrice Oshika**
MITRE
Santa Barbara, CA
bea@mitre.org

**Sherri Condon**
**Lisa Harper**
**Jon Phillips**
MITRE
7515 Colshire Dr.
McLean, VA 22102
{scondon,lisah,
jphillips}@mitre.org

## Abstract

The DARPA program CAST funded the development of four two-way speech-to-speech translation systems for four different language pairs across two domains. We describe the evaluation methodology, metrics and lessons learned from a non-comparative evaluation of the systems in one of these domains. Because the goal of the systems was to provide translation support for the performance goal-oriented tasks, the emphasis for assessment was on "live" real-time interactions between English subject matter experts and foreign language speakers engaged in realistic, unscripted scenario-based interactions. We measured performance in terms of how well these systems supported dialogue interaction between the participants. We discuss these metrics as well as some ideas for improvements in future evaluations.

## 1 Introduction

In August 2004, a final evaluation for DARPA's CAST speech translation program was held by MITRE in McLean, Virginia. Four development teams contracted by DARPA contributed two-way speech-to-speech translation systems for Thai, Farsi, Mandarin, and Pashto. The evaluation built upon a dry run assessment of methodology and logistics that had been held six months earlier. An Evaluation Advisory Committee (EAC) made up of the development teams, MITRE as evaluator, and government representatives met regularly to discuss and plan the evaluation. The dry run and the EAC meetings ensured that all parties had a common understanding of the assessment methodology and the goals of the final evaluation. There have been relatively few evaluations of speech-to-speech translation systems and even fewer emphasizing usability of the devices. Thus, the EAC had to design the CAST evaluation from the ground up. The only previous work we were able to build directly on was a field test of the Tongues speech-to-speech translation system using US Army officers and untrained Croatian speakers (Frederking et al. 2002).

Because of differences in system design, language structure and available resources for the various languages, it was agreed that the evaluation would not be comparative in the sense of scoring and ranking along simple dimensions. The intent was to provide a structured analysis of how well systems performed in mediating between two speakers, an English-speaking military medical Subject Matter Expert (SME) and a Foreign Language Expert (FLE) acting as a patient, who attempt to communicate about a medical condition and possible diagnosis.

The final evaluation was successful in several ways. Both the medical SMEs and the foreign language 'patients' recognized the potential utility of speech translation systems and, in general, adapted to using the systems. System performance improved over the baseline established during the dry run, and designers had attempted to address user interface concerns that had arisen in the dry run, such as the need for additional FLE user training on the systems. The audio and video data collected were very rich and provide insight into the nature of device-mediated conversations, as well as the diversity of strategies that participants use to overcome technical difficulties. Possibly most important, the framework for evaluation of such systems has now been well-tested and is available for future assessments.

## 2 Types of Assessment

The system evaluation included measures and methodologies for both component-level and system performance assessments. Testing of speech recognition and machine translation components was performed off-line by the developers themselves using a limited amount of prepared medical-domain-specific text and audio provided by MITRE (in English) and the developer sites (for the foreign languages).

More important was the live evaluation of overall system and task performance. As mentioned, no cross-system comparisons can legitimately be made because of the variability associated with the systems. Each system performs bi-directional translation for a unique language pair, and there are significant structural differences in the target languages plus widely varying language resources (availability of standard writing systems, lexicons, grammars, literacy rate of native speakers, previously existing speech and language technologies). In addition, the individual abilities and characteristics of the relatively small number of SMEs and FLEs had significant impact on the success or failure of the interactions.

Therefore, assessment of each system needs to be interpreted in the context of the resources and target populations for the system. A common framework and tools can be used for evaluation, and aggregate characteristics of usage of speech translation systems can be described, but a laboratory evaluation will always have limitations. The genuine utility and usability of such systems can be assessed more accurately in the field, where we they could be employed in many more (real) interactions.

The CAST systems were designed to conform to the expectation that the primary operator of the translation device is the English-speaking SME, while the FLE may or may not have physical or visual access to the device. Our evaluation methodology reflects this general assumption. Where appropriate, we allowed or disallowed visual access to the device by the FLE, to explore the effect of FLE access to visual feedback from the system. Some systems were not designed for the FLE to see the display, e.g., for Pashto, there is typically a low literacy rate among potential FLE users and such displays would add little value in real usage. In order to demonstrate the utility of systems under optimum conditions, as defined by individual system designs, and also under common conditions, the assessment protocol allowed for varying conditions to the extent reasonable.

## 3 Off-line System Assessment

The purpose of component testing is to provide a baseline for two core components of a speech translation system: the process of speech recognition and the process of translation. Word Error Rate (WER) is a widely-used ASR metric that is commonly understood. It is a measure of errors of insertion, deletion and substitution in a recognized string as compared with the spoken string. There is no similar 'standard' metric for translation because of the complexity associated with retaining the 'sense' of an utterance across languages. The CAST Evaluation Advisory Committee opted for BLEU, a similarity measure increasingly used for assessment of machine translation because it lends itself to automated analysis (Papineni et al., 2001). It measures similarity of an automatically translated string to any of N references that have been translated by foreign language experts (in most studies using BLEU, N=4). For speech translation systems, it is useful to run BLEU on both text-to-text translations (with well-formed input text) and speech-to-text translations (with 'noisy' input text generated by a possibly errorful ASR component).

MITRE provided the developers with five English dialogues representative of typical interactions between a medical professional and a patient. In addition to the English text, audio recordings of the medical professional English side of the conversations were created by MITRE employees and provided to the development teams. The total number of words for the original English versions of the five dialogues was only 1072. Each of the four English speakers recorded 801 words for the medical providers' utterances in the dialogues. The English version of the patients' utterances in the five dialogues was a total of 271 words.

Developers translated the patient side of the dialogues into the target foreign language (FL) and the translations were spoken by native FL speakers. This process provided English and FL text that could be used for text-based MT component testing, and English and FL speech that could be used for ASR component testing and MT component testing when the MT input was ASR output. Teams reported BLEU scores and word error rate for both the recorded and written interactions.

It is important to note that the English reference translations produced by FL speakers were not necessarily error-free. The human translators were not native speakers of English nor necessarily professionally trained translators. So English references used by BLEU could have included ungrammatical, awkward, or even inaccurate translations. The development teams and evaluators agreed to run BLEU scores as a component metric not as a means for evaluating system performance, but to see what we could learn about the metric itself since it was never intended for the particular kind of data that these systems output. For that reason, we are not reporting BLEU or WER scores here.

## 4 Live System Assessment

Real-time assessment was achieved via interactions between English-speaking medical professional SMEs and 'patients' role-played by FLEs. The FLEs were ordinary native speakers of the FL, who had no prior exposure to the systems. Participants role-played twelve scenarios illustrating illness and injury typical of the target FL environments. FLE patients were trained by a medical professional (via an interpreter if needed) in the symptoms/injuries they were supposed to simulate and describe, and SME medical personnel were asked to interview the patient using the speech translation system as a mediating device and to try to arrive at an expected diagnosis and treatment plan. Interactions were videotaped for later analysis, and both participants completed a questionnaire at the end of each session. These interactions were designed to determine the utility of the speech translation systems in supporting an actual task: a medical interview and expected diagnosis and treatment. Although component performance obviously affected system utility, such performance was not explicitly captured in the real-time interactions because of perceived labor-intensive post-processing required for transcription and translation. The emphasis was on participants' ability to complete the task or at least exchange relevant information using the speech translation system.

While reading system results, one should bear in mind that interface and effectiveness issues are always very sensitive to human factors. We have observed that many of the measures presented in this report are subject to behavioral differences caused by different styles and strategies, especially among the SMEs. In addition to measurable differences in specific behaviors, there were clear differences in attitude and comfort levels among both SMEs and FLEs.

### 4.1 Set-Up

Four systems were tested over three days, with two parallel evaluation sessions each day. Each session occupied two rooms, one room for the system interaction and the other for observation. There were also two rooms for patient training, plus separate waiting rooms for SMEs and FLEs who were not in a session. Interactions in both the evaluation rooms and patient training rooms were videotaped, and the evaluation room camera displays were broadcast to the observation rooms.

Participants each day included two Patient Trainers and three SMEs, as well as two FLEs per language (four total for day) and at least one interpreter per language.

The SMEs were all medical providers (doctors, corpsmen and physician's assistants) from the Naval Medical Center at Quantico, except for one from the School of Nursing at Boston College. Two of the SMEs were female and four were male. Three SMEs participated only on the first day, and the other three participated on both the second and third days.

For each language there were two FLEs acting as patients, plus a third who served as an interpreter for the monolingual FLEs. Each system and device included a pre-recorded audio or video presentation of instructions for the FLE, which was shown the first time that each FLE used a system each day. In some cases, FLEs viewed or heard the pre-recorded presentation again if their performance suggested that they were having problems.

Each system was evaluated for a day and a half. There was one hour for a development team to train the two SMEs on its system (with a FLE provided by the developers who was used only during SME training), and then 45 minutes for each scenario, for a total of eight scenarios per day. No FLE repeated the same scenario, and only one SME repeated one scenario. The variant condition (e.g., FLE could / could not see screen) was balanced across FLEs and SMEs.

Scenarios describing medical situations were adapted from 200 common development dialogues covering just under 100 different scenario topics. Twelve evaluation scenarios were selected to reflect situations that might occur in any of the regions where Farsi, Mandarin, Pashto and Thai are spoken. A topic involving the boarding of a ship would be inappropriate, for example, because some of the areas are landlocked. The selected scenarios dealt with common illnesses, such as gastrointestinal distress, and common injuries in military and civilian environments, such as shrapnel wounds or trauma from a vehicle accident.

The goal was to make the interactions as realistic as possible. Toward that end, developers were not allowed to be in the room during the formal evaluations, FLEs and SMEs could not observe training sessions, and SMEs did not have the details of a scenario until they elicited them from FLE patients. To reflect the expected use case, only SMEs could show FLEs how to use the system (i.e., there was no interaction between FLEs and development teams about device usage, unless there were technical problems).

For an evaluation of speech translation systems where the presumption is that FL speakers have no competency in English, it is important that no English mediates the interaction between the conversational participants. This meant that bilingual FLEs had to be acoustically isolated in some way so that they could not hear the English spoken by the SME or the English TTS from the system. We decided to recruit monolingual FLEs for each language, but it is virtually impossible to find a FL speaker living in this country who does not speak any English at all – especially for some of our target languages. We found one nearly monolingual FLE for each language, and used isolating headphones for bilingual FLEs.

### 4.2 Results and Analysis

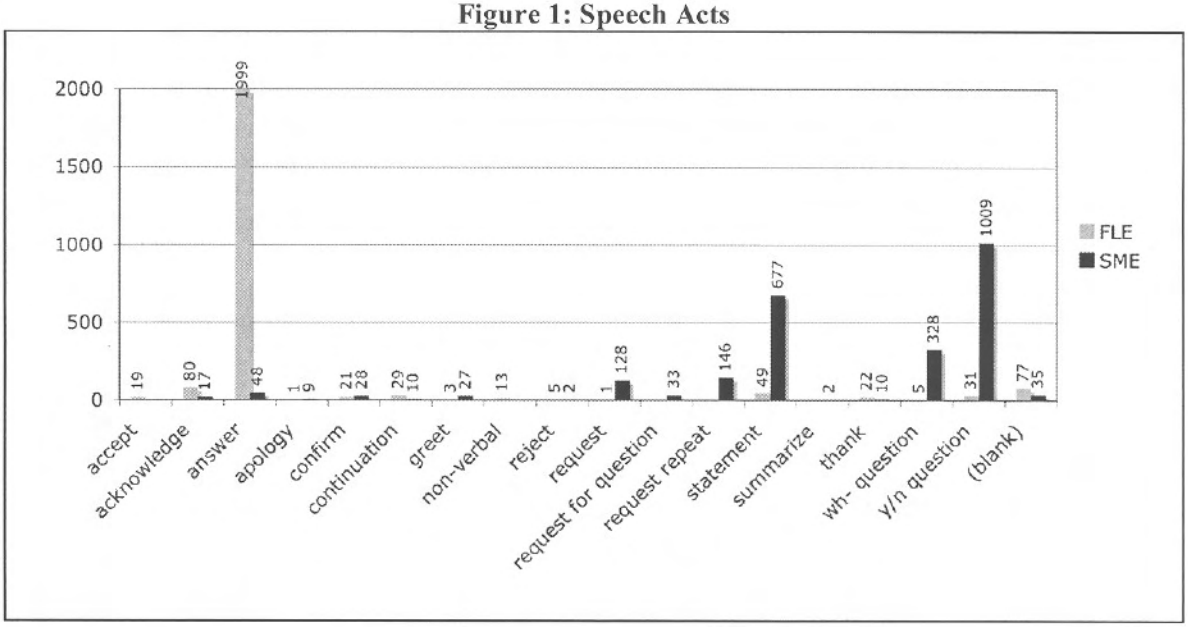
For each interaction, transcribers annotated at the turn level of interaction. We define a turn as a single segment handled by the system (i.e. literally, a unit of processing). Spoken English was transcribed word for word, and no attempt was made to punctuate the text or to represent intonation, partial words, incomplete words, fillers, pauses or filled pauses, or other features of pronunciation. In order to compare the English that was spoken with what was recognized by the system, we recorded the recognized string using the system logs and the videotaped view of the screen. Translations of FLE responses were recorded using the system log, the video of the screen, and the synthesized audio. Note that we did not have access to correct translations of FLE turns. In a comment field, we also noted nonverbal behavior and other relevant information from the videos.

Transcription of all interactions for a single language were made by a single MITRE linguist so that it was not necessary for all of the linguists to spend time becoming familiar with the idiosyncrasies of more than one system. However, all four MITRE linguists participated in linguistic annotation across each of the four systems. We worked in pairs and rotated pairs such that everyone worked with one another. This was important since we needed to share a common methodology for annotation that was consistent across annotators and languages.

For each of the twelve evaluation scenarios per language we annotated the following phenomena:

1) **Speech Acts**. The speech acts that occur in the interactions are determined in large part by the goals and structure of medical interviews, and the quantity and variety of acts can tell us a lot about a specific interaction. The set of seventeen speech acts that was used to annotate these interactions was selected by beginning with a list of common acts. The goal was to keep the categories simple and clear in order to preserve reliability, while discriminating types of acts that seemed significant in the interview context.

**Figure 1: Speech Acts**



2) **Translation**. We annotated whether FLE turns were communicated via translation, without translation, or before translation.

**Table 1: FLE Responses, non-cancelled turns**

| | Turns | Percentage |
|---|---|---|
| via Translation | 1089 | 86.16% |
| without Translation | 165 | 13.05% |
| Before Translation | 2 | 0.16% |
| No response | 8 | 0.63% |

3) **Relevance**. We annotated the relevancy of FLE turns that were interpreted to be second parts of an adjacency pair, such as answers to questions, acknowledgments of statements, and acceptance of requests. Translated responses were judged as complete and fully relevant, partial but relevant, interpretable as relevant, or irrelevant.

**Table 2: Relevance of FLE Responses**

| | Turns | Percentage |
|---|---|---|
| Complete and Relevant | 654 | 55.95% |
| Partially Relevant | 83 | 7.10% |
| Could be interpreted as relevant | 215 | 18.39% |
| Irrelevant | 217 | 18.56% |

4) **Repetition**. We annotated repetitions and reformulations, distinguishing whether these occurred before the partner's response or after the partner's response.

**Table 3: Repeats and Reformulations, non-cancelled turns**

| | Turns | Percentage |
|---|---|---|
| Repeat/Reformulation before Response | 671 | 19.31% |
| Repeat/Reformulation after Response | 180 | 5.18% |

5) **System problems**. We annotated system problems: system errors, unintended system actions, and turn cancellations. System errors occurred when the system produced error messages or some function failed to produce an action. Unintended system actions occurred when the system performed an action that wasn't the expected response to a SME command or occurred without a SME command. The majority of turns were cancelled by SMEs and FLEs when recognition was considered unsatisfactory (if the system and set up allowed them to do this).

**Table 4: System Problems**

| | Turns | Percentage |
|---|---|---|
| System Errors | 30 | 0.62% |
| Unintended System Actions | 18 | 0.37% |
| Turn Cancellations | 1390 | 28.58% |

6) **Turn timing for system input**. We annotated turns that were not appropriately synchronized with the system in the sense that the system was not in recognition mode or was in recognition mode for the wrong language. We observed several types of turns not synchronized with system: 1) the FLE could hear and understand the English and responded before being prompted; 2) more frequently, the FLE responded immediately to the translation before the SME was able to put the system in FL recognition mode; 3) some FLEs produced extremely long utterances that extended past the recognition mode interval (some systems had a time-out on recognition); 4) the SME did not expect the FLE to respond and did not put the system in FL recognition mode. Occasionally, the session was paused so that it could be communicated to the FLE that the SME would indicate when the FLE should speak. This was a problem encountered if the FLE did not understand the usage instructions at the beginning of the session.

**Table 5: Turns not Timed for System**

| | Turns | Percentage |
|---|---|---|
| Turns not timed for system | 199 | 4.09% |

7) **Task completion**. We tabulated the number of facts that were communicated of five basic facts that had been identified in advance as central. The facts were selected by one of the medical professionals who prepared the scenarios. Typically, two of the facts concern the main medical problem, such as chest pain, and when symptoms began. The remaining three anticipate questions that the SME would ask, such as *did you lose consciousness?, do you feel dizzy?* or *are you taking any medications?*

**Table 6: Task Completion**

| Average Task Completion | Percentage |
|---|---|
| 2.78 | 55.65% |

In addition to raw counts from annotations described above, we calculated the following:

8) **Length of interaction**. Many factors influence the number of turns that participants employ in each interaction. The scenario content, SME style, FLE characteristics, and system features can all affect the length of the interactions. There was great variability in the length of interaction for each system. The length of individual interactions does not necessarily correlate with the success of the interaction. A lengthy dialogue may indicate a successful interaction with much information communicated, or multiple repetitions caused by communication difficulties. Shorter interactions may indicate a smooth and quick exchange of information or, possibly, a quickly frustrated SME who gave up early. More detailed analyses of turn functions such as repetition are needed to determine the quality of the interactions.

9) **Average number of turns by SME.** Because features of SME interaction such as their persistence can have a large impact on the length of interaction, we wanted to measure the length of interactions for each SME.

10) **Interaction length, non-cancelled turns**. Because longer interactions may indicate either greater satisfaction with the device or more difficulty using the device, the interaction lengths of FLEs and SMEs were calculated to exclude cancelled turns. Of course, this still doesn't tell us whether there is greater satisfaction or more difficulty with using the systems. Repetitions and reformulations must be taken to account.

**Table 7: Interaction Length (turns)**

| | FLE | SME | Total |
|---|---|---|---|
| All turns | 49.06 | 52.27 | 101.33 |
| Non-cancelled turns | 30.19 | 41.19 | 72.38 |

11) **Speech Acts for each role (FLE or SME).** The asymmetry of SME and FLE roles is apparent in the large differences between the frequency of questions (*yes/no* and *wh-*) produced by SMEs and answers produced by FLEs (**Figure 1**). Also, SMEs produce many more requests and statements than FLEs. These patterns often reflect behavioral differences in the participants; for example, in the dry run one SME's strategy of summarizing responses from the FLEs led us to include an annotation for summarization, but this strategy was employed only twice in the final evaluation.

12) **Percentage of SME Questions Answered**. Counts of individual speech acts do not address the sequential relations among speech acts such as questions and answers or statements and acknowl-

edgments. We obtained an estimate of the proportion of successful translated FLE answers to SME questions by identifying FLE turns with the following three properties: First the turn was annotated as an answer (may or may not have been translated). Second, the turn was annotated as at least potentially relevant. Third, the turn was immediately preceded by a SME question. This information was calculated automatically, so it is possible that the answer was to some previous question. The number of these turns was divided by the number of all SME questions to obtain percentages.

**Table 8: FLE Answers to SME Questions**

| | SME Questions | FLE Answers | Answer Rate |
|---|---|---|---|
| yes/no questions | 1009 | 623 | 61.74% |
| wh-questions | 328 | 185 | 56.40% |

## 5 Discussion

Although there appeared to be discernible patterns among the measures, the scores obtained by the systems were often very similar. Number of turns per interaction, rates of turn cancellations, system errors and repetitions / reformulations were very close. Though they were more difficult to judge, annotations of relevance and speech acts provide more refined views of the interaction data and the effectiveness of the systems. The judgments of FLE turn relevance based on the English translations provide a reasonable estimate of useful information exchange that the SME was able to elicit using the system. However, the relevance measures are most informative when presented relative to the total number of FLE turns. Thus, the relevance measures do not provide an absolute score, in contrast to the task completion measures, which do provide an absolute score.

Speech act annotation is labor-intensive, but it can reveal significant patterns that are otherwise inaccessible. The presence of translated FLE questions suggests that the interaction was successful enough to proceed beyond the bare essentials of information required by the medical provider, although it may also be the case that cultural factors and the perception of medical personnel are likely to influence how free the FLE feels to take initiative.

Because each measure can be influenced by a variety of factors, our approach was to provide a broad set of measures that could reinforce our confidence in the results when they pattern together. Though the measures seem to agree on a global result of effectiveness, it would be helpful for developers, trainers, and users to devise measures that can evaluate specific functions or interface designs.

Traditional measures of translation accuracy are difficult to obtain: they require independent translations of each turn and judgments that are challenging and labor intensive. Like speech recognition accuracy, translation accuracy or adequacy might not be a good indicator of interaction success because traditional measures rarely take into account human ability to infer information, especially when motivation is high. There were several instances in which SMEs responded to nearly nonsensical translations with utterances that displayed surprisingly coherent interpretations. We are inclined to think this may be one of the "good applications for crummy machine translation" discussed by Church and Hovy (1991).

Most difficult is separating the influence of basic functions such as recognition and translation from design choices in the interface, but this specific evaluation does not allow us to get at such a comparison. Measures that might reflect the effectiveness of the interface design were not captured. For example, it would be possible to record the frequency with which SMEs selected options available in the interface or whether having selected an option, the SME actually used it instead of just switching back to the defaults. The latter was observed on several occasions in these interactions. The important of detailed system log files for this kind of analysis is discussed further in the lessons learned section.

Interface and effectiveness issues are always very sensitive to human factors, and we have observed that many of the measures presented in this report are subject to behavioral differences caused by different styles and strategies, especially among the SMEs. In addition to measurable differences in specific behaviors, there were clear differences in attitude and comfort levels among both SMEs and FLEs.

Clearly the best way to avoid effects of idiosyncratic differences among speakers is to employ a large sample of FLEs and SMEs, but this strategy was impractical for us: it required extensive effort and planning to recruit and schedule the 2 SMEs and 3 FLEs that were obtained for each system. Also, FLE recruitment difficulty varied among languages. For the final evaluation, we were able to increase the number of SMEs available, but recruiting FLEs, especially monolingual FLEs, created many problems.

Possibilities for improving future evaluation metrics for comparative evaluation include the following:

- Estimates of dialogue initiative;
- Amounts of information participants attempt to exchange;
- Concept transference metrics including concept weighting;
- Some notion of relative value of information transferred by adding weights to turns.

## 6 Summary

The final evaluation of speech translation systems developed under DARPA's CAST program included both offline component testing using representative canned data prepared by MITRE and real-time interactions between medical personnel and foreign language speakers.

The real-time interactions consisted of English-speaking medical providers interviewing speakers of Mandarin, Pashto, Farsi and Thai who were trained to act as patients in well-defined medical scenarios. Patients described symptoms of injury or illness and the task for the medical professional was to diagnose the condition through an interview mediated by the speech translation device (on both laptop and hand-held platforms). The interactions were videotaped and transcribed, and each speaker's turn was annotated for features characterizing the conversations, including turn-taking, types of responses, requests for repetition or clarification, and task completion.

## 7 Conclusion

System development and evaluation under the CAST program has demonstrated the feasibility of using speech translation systems to perform relatively complex but well-defined tasks such as medical diagnosis of common illness and injury. The interactions were characterized by more repetition and repair than would occur in a monolingual situation or with a human translator, and they were tedious and frustrating, but in many cases sufficient information was exchanged for medical professionals to make diagnoses and suggest a plan of action.

The CAST program also illuminated the qualitatively different nature of translingual 'dialogues' mediated by a device, and the need to support commonly used modes of communication such as gesture and to augment speech recognition and translation capabilities with graphical displays.

## References

K. Church and E. Hovy. 1991. "Good Applications for Crummy Machine Translation," in Proceedings of the 1991 Natural Language Processing Systems Evaluation Workshop. Rome Laboratory Final Technical Report RL-TR-91-362.

R. Frederking, A. Black, R. Brown, J. Moody, and E. Steinbrecher, 2002. "Field testing the tongues speech-to-speech machine translation system," in Proceedings of the Third International Conference on Language Resources and Evaluation (LREC).

K. Papineni, S. Roukos, T. Ward, and W-J. Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. IBM Research Report RC22176 (W0109-022) (http://citeseer.ist.psu.edu/papineni02bleu.html)