# Acceptance Testing of a Spoken Language Translation System

**Rafael Banchs, Antonio Bonafonte, Javier Pérez**

Department of Communications and Signal Theory
Universitat Politècnica de Catalunya, Barcelona, Spain
{rbanchs, antonio, javierp}@gps.tsc.upc.edu

**Abstract**

This paper describes an acceptance test procedure for evaluating a spoken language translation system between Catalan and Spanish. The procedure consists of two independent tests. The first test was an utterance-oriented evaluation for determining how the use of speech benefits communication. This test allowed for comparing relative performance of the different system components, explicitly: source text to target text, source text to target speech, source speech to target text, and source speech to target speech. The second test was a task-oriented experiment for evaluating if users could achieve some predefined goals for a given task with the state of the technology. Eight subjects familiar with the technology and four subjects not familiar with the technology participated in the tests. From the results we can conclude that state of technology is getting closer to provide effective speech-to-speech translation systems but there is still lot of work to be done in this area. No significant differences in performance between users that are familiar with the technology and users that are not familiar with the technology was evidenced. This constitutes, as far as we know, the first evaluation of a Spoken Translation System that considers performance at both, the utterance level and the task level.

## 1. Introduction

Many works on translation and spoken language system evaluation have been reported in the literature (Polifroni et al., 1992; Arnold et al., 1993; Sikorski and Allen, 1995; Gates et al., 1997; Walker et al., 1997; Carter et al., 2000), among others. Some of them focus on evaluating translation quality, while others focus on task achievement.

In this work we present a two-test acceptance evaluation procedure based on the works of Somers and Sugita (2003) and Thomas (1999). First, an utterance-oriented evaluation for determining how the use of speech benefits communication is performed. This test allows for comparing relative performance of the different system components. Second, a task-oriented evaluation for determining the capability of users for achieving a given task is conducted.

Next section describes the spoken language system for which the acceptance test was performed. Then, section 3 explains the acceptance test procedure and presents a detailed description for both tests performed. Section 4, presents the results and discusses the most relevant issues related to them. Finally, some conclusions are presented in section 5.

## 2. Overview of the System

The different components of the spoken language system to be evaluated have been integrated into the *GAIA* framework. This platform was developed in the LC-STAR project with the objective of creating a distributed platform where different solutions to the three main aspects of a spoken translation system (speech recognition, machine translation and speech synthesis) could be integrated. We have used *GAIA* to demonstrate the project's experimental results on translation among three target language pairs (Catalan, Spanish and US-English) covering the tourist domain defined in the project. The platform can be configured to be used either for one channel (one person speaks in the source language and the systems provides the translation) or for two channels (two persons speaking through the platform, and the platform performs the translation). It can also be configured to acquire databases. The main part of the platform is the kernel, which handles the communication among the modules of the system:

- User modules: which collect the input of the user and provide the output. Three dual terminal servers have being developed: i.- telephone terminal, to interact using the telephone through Dialogic cards; ii.- speech console terminal, to interact using speech through and IP connection and iii.- text console terminal which is mainly used to test the translation engine. The demonstrator is based on the telephone terminal: two users can communicate using different languages, through a translation service provided by *GAIA* through the telephone.
- Technology modules: which are responsible for the speech processing tasks. The systems currently integrated into *GAIA* are:
  - Automatic speech recognition (ASR): provided by UPC (Bonafonte et al., 1998; Mariño et al., 2000)
  - Text to speech synthesis (TTS): provided by UPC (Bonafonte et al., 1998) and from the Festival project[1].
  - Spoken language translation (SLT): provided by RWTH (Och and Ney, 2002).
- Visualization modules: which allow for remote monitoring of the output of each technology in all the steps of the process.

Acoustic models for ASR for Spanish and Catalan have been trained using either the TALP-tourism corpus or a

[1] Available on-line at: "http://www.festvox.org/".

combination with SpeechDat databases[2]. Both, the acoustic databases and the software to train the acoustic models are provided by UPC. For English the MACROPHONE corpus has been used. However, this corpus is not adapted to the task. Furthermore, for this task there is not corpus in English available for development and testing. For this reason, in this work we have decided to avoid the use of English and perform the acceptance test in two directions only: Spanish-to-Catalan and Catalan-to-Spanish.
The language models for speech recognition are trained from the TALP-tourism corpus, using both the source sentences and the translated sentences. First, several classes are defined (hotels, names, cities of the world, etc.). Then, class n-grams are inferred using variable-length n-grams, linear discounting and back-off smoothing (Bonafonte and Mariño, 1998). The toolkit to estimate the n-grams is part of RAMSES, the UPC Continuous Speech Recognition System (Bonafonte et al., 1998). Several trials were done to add the Verbmobil corpus or tourist web pages to the training material but there was no significant decrease on the perplexity. The LC-STAR lexicons for Spanish and Catalan (UPC) and English (NSC) have been used for the speech recognition and speech synthesis engines. For a more detailed description of Gaia, refer to (Pérez and Bonafonte, 2004).

## 3. Acceptance Test Procedure

For performing the acceptance test of *GAIA*, two tests were designed. The first one focused on end-to-end evaluation at the utterance level and tried to evaluate if the use of speech, at the state of the technology, benefits communication. The second one was task oriented and aimed at evaluating if users can achieve a given task with the state of the technology.
For both tests, twelve subjects from two different groups were selected to participate in the evaluation: eight of the subjects were familiar with the technology (post-graduate students in language technologies) and four of the subjects were not familiar with the technology (administrative staff). Although most of them were bilingual, they were required to write/speak and read/listen in their native language during all the evaluation procedure.

### 3.1. Test-1: utterance-based evaluation

The aim of a spoken language translation system is to allow for the communication of two persons speaking different languages. Speech-to-speech communication is expected to be the more natural and comfortable way of communication. However, there are still some technological limitations in speech recognition and speech synthesis that can introduce additional errors degrading significantly the usability of the system. For this reason, this test evaluated four systems, all of which included translation but the input and the output modality could be either speech or text. Note that the translation engine has been trained from transcriptions from speech, not text. However, in our experience the use of written text in the same domain does not cause any degradation on the translation quality. Explicitly, the systems evaluated were:

- System 1: source text to target text (TT)
- System 2: source text to target speech (TS)
- System 3: source speech to target text (ST)
- System 4: source speech to target speech (SS)

#### 3.1.1. Production of the stimuli

Before starting, the twelve subjects selected to participate in the evaluation were asked to use *GAIA* from 10 to 20 minutes, so that they could become familiar with the system. During that time all of them were able to speak to the system and watch both the ASR output and its subsequent translation.
After they were ready, each subject was asked to utter 3 to 5 short sentences from two of the following four scenarios:

- Scenario A: hotel reservation (client)
- Scenario B: hotel snack bar service (client)
- Scenario C: flight reservation (client)
- Scenario D: hotel reservation (agent)

Afterwards, they were asked to write down 3 to 5 short sentences conveying the same meaning (same scenario).
In the case of the speech input, the utterances were first converted into text using the speech recognition server. Then, all text inputs, either written by the subjects or produced by ASR, were translated using the statistical speech translation server. And, finally, the translated text was converted into speech using the speech synthesis server.
Using this protocol, 4 versions (one for each system, TT, TS, ST and SS) of the 24 items (12 speakers x 2 scenarios) were generated. An item is a set of 3 to 5 sentences from a given scenario and a given subject.

#### 3.1.2. Evaluation

To evaluate the system usability, each resulting combination item/system was heard/read by one subject. The same 12 subjects who generated the stimuli were asked to evaluate all stimuli. As there were a total of 96 outputs (24 items x 4 versions) and each subject was asked to evaluate 16, each output was evaluated twice. Special care was taken so that: first, no subject evaluated any output produced by him/herself; and second, each subject evaluated no more than one version of each item.
Following Somers and Sugita (2003), the subjects were asked to paraphrase what they have understood using reported speech style: *She is asking for ...*.
Once all the reports had been produced, three judges read all the reports and compared them with their corresponding inputs. Based on this comparison, they rated the reports using the seven-point scale defined by Somers and Sugita (2003), which we reproduce here:

- Useful

  6: Clearly useful to communicate the intention of the utterance: the response matches was is intended in the original utterance. It contains the same concepts and all the necessary arguments.

[2] Available on-line at: "http://www.speechdat.org/".

5: Generally useful: the response nearly matches what is intended in the original utterance; may misrepresent or omit some detail that is not fatal.

- Borderline

  4: Useful but less informative compared with the above: basic match with what is intended, but some accompanying arguments are incomplete or inadequate.

  3: Useful but not wholly adequate: as 4 but some arguments are missing.

- Useless

  2: Almost useless but still informative and useful: the response does not match what is intended but nevertheless contains some partially useful information.

  1: Clearly useless: the response does not match what is intended in the original utterance at all.

- No response

  0: Blank or garbage

The judges evaluated all the reports and the average score was taken in order to obtain more consistent results.

### 3.2. Test-2: task-based evaluation

As already mentioned, this second experiment was task oriented and aimed at evaluating if users could achieve some predefined goals for a given task with the state of the technology. In this experiment, only system 4 (source speech to target speech) was evaluated.
Each evaluation consisted on a dialog between two participants. The same 12 subjects from the first experiment participated in this one, so a total of six dialogs were performed. The six pairs of subjects were required to interact following some given basic guidelines by using *GAIA*, in its speech-to-speech configuration. All the inputs to the platform (source speech in both languages) as well as the intermediate and final results (text before and after translation and final speech) were logged by the system to allow further analysis.

#### 3.2.1. Task description

For each dialog, participant 1, or "the client", was assigned the task of making a room reservation (scenario A from test 1), while participant 2, or "the agent", was assigned the task of booking the reservation (scenario D from test 1). The task involved achievement of the following eight specific goals:

- Goal 1: arrival date,
- Goal 2: number of nights to stay,
- Goal 3: type of room requested,
- Goal 4: price of the requested type of room,
- Goal 5: full name of the client,
- Goal 6: type of credit card to be used for making the reservation,
- Goal 7: credit card number, and
- Goal 8: credit card expiration date.

Both participants, the client and the agent, were given written guidelines with the information required by each of their roles and they were asked to write down what they had achieved for each of these eight specific goals.

#### 3.2.2. Evaluation

As already mentioned, each subject was asked to fill in a form reporting each accomplished goal according to the given guidelines. The analysis of the conversations and the filled forms revealed if the goals were actually achieved and how many repairs were needed to achieve them.
Scores where computed by using the prioritization-of-goals measure proposed by Thomas (1999), where the score is a function of the success in communicating the goal and the number of repairs attempts. In this experiment, successfulness of a given goal was determined by comparing the information written down by both users during the dialog. Then, the score for a successful goal was computed by using (1), where *repairs* refers to the number of times one user repeated the same goal related information. On the other hand, the score for an unsuccessful goal was computed by using (2).

$$score_{succ} = \frac{1}{1 + repairs}, \tag{1}$$

$$score_{fail} = \frac{1}{1 + repairs} - 1. \tag{2}$$

According to (1) and (2), goal scores range from 1 (when the goal is achieved without any repair) to -1 (when the goal is not achieved after many trials).
Three judges evaluated independently all the six dialog transcripts and computed a score for each of the eight goals in each dialog. The final score was obtained by averaging the three judges' scores.

## 4. Evaluation Results and Discussion

This section presents the results of both performed tests: the utterance-based evaluation (test 1) and the task-based evaluation (test 2). Only Spanish-to-Catalan and Catalan-to-Spanish translation directions were considered in these evaluations. As previously mentioned, a total of twelve subjects participated in evaluations. They were selected from two different groups: familiar with the technology and non-familiar with the technology; and although most of them were bilingual, they were required to write/speak and read/listen in their native language during all evaluations.

### 4.1. Test-1 results

This first test focused on end-to-end evaluation at the utterance level and tried to evaluate if the use of speech, at the state of the technology, benefits communication. Fig. 1 presents the ranking of the four evaluated systems: source text to target text (TT), source text to target speech (TS), source speech to target text (ST), and source speech to target speech (SS); according to the seven-point scale already described in sub-section 3.1.2. The ranking presented in

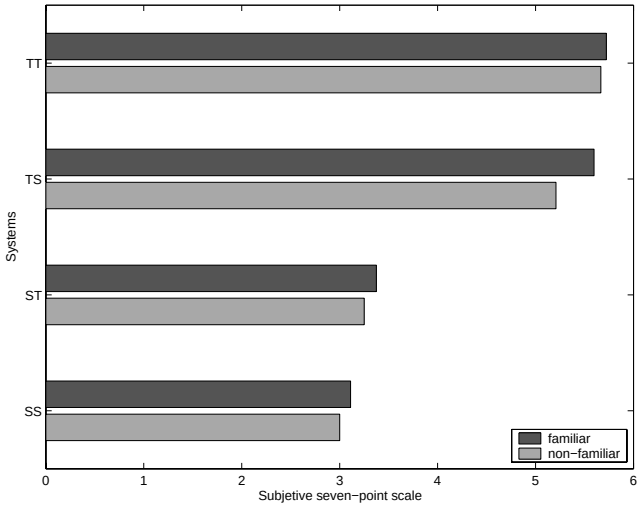


Figure 1: Ranking of the four systems considered (TT, TS, ST, SS) for both classes of users (familiar and non-familiar with technology) according to the seven-point subjective scale.

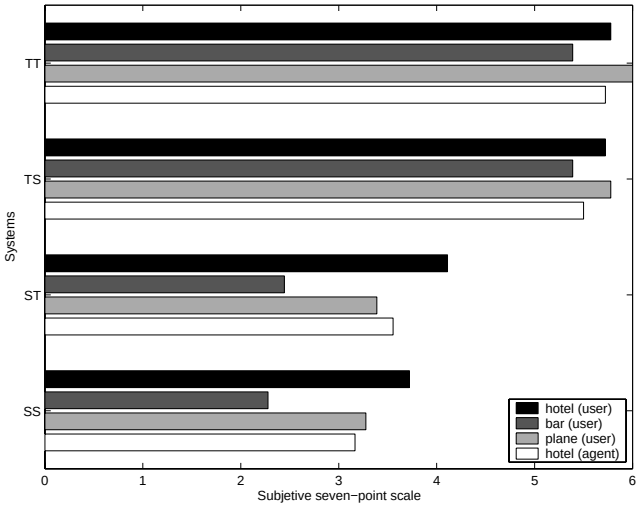


Figure 2: Ranking of the four considered systems (TT, TS, ST, SS) for each of four types of scenarios (room reservation: agent and client, snak bar: client, fly reservation: client) according to the seven-point subjective scale.

Fig. 1 makes the distinction between both classes of users, those familiar with the technology and those non-familiar.
Two important conclusions can be drawn from Fig. 1. First, it is evident that ASR is responsible for degrading the overall system performance from "useful" (scores of 6 and 5) to "borderline" (scores of 4 and 3). Systems 1 (TT) and 2 (TS) obtained scores of 5.72 and 5.60 respectively, while systems 3 (ST) and 4 (SS), which both involved the ASR, obtained scores of 3.38 and 3.11 respectively.
It should be noted that Catalan and Spanish are quite similar languages, so the translation system is much more accurate than it is expected for a different source-target language pair.
A second observation from Fig. 1 is that, with the exception of system 2 (TS), no significant differences are appreciated between the scores obtained by users that were familiar with the technology and those that were not. A possible explanation for the difference observed in case of system 2 (TS) can be that users which are familiar with the technology are more skilled when listening to the TTS output. On the other hand, in the case of system 4 (SS) which also implied listening to the TTS output, the degradation introduced by the ASR might result in an output as difficult to understand for users familiar with the technology as for those non-familiar.
A similar ranking of the four systems was performed for each of the four considered scenarios: hotel reservation/client (A), hotel snack bar service/client (B), flight reservation/client (C), and hotel reservation/agent (D). These results are presented in Fig. 2.
From Fig. 2, it can be seen that scenario B (hotel snack bar service) seems to be the less successful one. The difference between scenario B and the others is much more evident in the cases of systems 3 and 4, for which ASR is present. This clearly suggested, as was later confirmed, that scenario B was less represented on the training data than the other three given scenarios. On the other hand, ASR seems to be favoring scenario A (hotel reservation/client).

Finally, a cross-plot between the subjective seven-point evaluation and an automatic error metric was performed in order to see how well the subjective and automatic evaluations correlated to each other. The automatic metric used was the word error rate (WER), which was measured at the output of the MT system. The obtained regression, which is depicted in Fig. 3, happened to be significant and the value of the obtained R-squared was $0.71$.

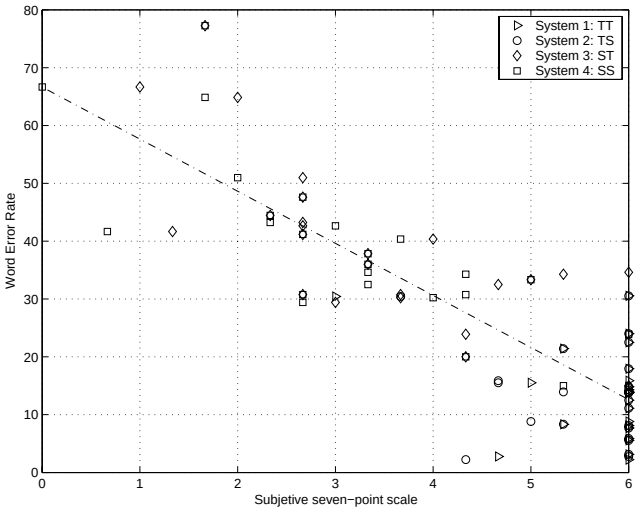


Figure 3: Cross-plot between subjective evaluation and the word error rate (WER).

### 4.2. Test-2 results

As already mentioned, this second test was a task-oriented evaluation and it aimed at evaluating if users could achieve some predefined goals for a given task with the state of the technology. In this experiment, only system 4 (SS: source speech to target speech) was evaluated. A total of six dialogs were evaluated.

Three judges evaluated independently all the six dialog transcripts and computed a score for each eight goals in each dialog by using the method described in sub-section 3.2.2. The final score was obtained by averaging the three judges' scores. The resulting goal and dialog scores are presented in Table 1 and Table 2, respectively.

| Dialog | arrival | nights | room | price |
|---|---|---|---|---|
| d-1 | 0.75 | 1.00 | 1.00 | -0.50 |
| d-2 | 0.60 | 0.38 | 1.00 | 1.00 |
| d-3 | 1.00 | 1.00 | 0.50 | 1.00 |
| d-4 | -0.50 | 1.00 | 0.50 | 0.60 |
| d-5 | 0.50 | 0.50 | 0.50 | 0.60 |
| d-6 | -0.63 | 1.00 | 0.43 | 0.27 |
| average | 0.29 | 0.81 | 0.65 | 0.50 |

| Dialog | name | cc-type | cc-num | cc-exp |
|---|---|---|---|---|
| d-1 | - | 1.00 | 0.00 | 1.00 |
| d-2 | 0.60 | 1.00 | -0.25 | -0.75 |
| d-3 | 1.00 | 1.00 | -0.50 | 0.60 |
| d-4 | 1.00 | 1.00 | -0.84 | 1.00 |
| d-5 | 0.50 | 0.75 | -0.57 | 1.00 |
| d-6 | 0.60 | -0.70 | 0.00 | -0.75 |
| average | 0.74 | 0.68 | -0.36 | 0.35 |

Table 1: Individual- and average-goal scores (in a scale from -1 to 1) for each of six dialogs in test 2.

From Table 1, goals can be ranked from more successful to less successful as follows:

- number of nights staying,
- full client name,
- credit card type,
- room type requested,
- price of room type,
- credit card expiration date,
- arrival date, and
- credit card number.

This result suggests, as we expected, that for the given state of the art the more restricted and closed answers and issues are the more successful the information communication is. Two important remarks must be addressed in the particular cases of goals 5 (client name) and 7 (credit card number).

In the case of the client names, both participants, "the client" and "the agent", did know in advance who they were talking to. So, although they were asked to report only based on what they actually could understand from the TTS, knowing who their interlocutor was might have boosted up this goal's score over more simple goals such as credit card type and room type. Also, it is important to mention that the number of names was limited to approximately 30 for the ASR system.

In the particular case of credit card number, this goal was consistently unsuccessful in all six dialogs. Two basic problems could be detected for this particular goal. First, most of "clients" said the sixteen digit credit card number making pauses after groups of four or six digits. Many of these small pauses were interpreted by the ASR as end of utterance, so "agents" received incomplete translated information which generated lot of confusions. Second, when long sequences of digits were correctly recognized and translated the output of the TTS was produced without any pause making it impossible for the "agent" to write down the whole sequence of digits. Actually, the reason of this second problem was a mismatch in format between the output of MT and the input of TTS.

| Dialog | Score | WER1 (asr) | WER2 (asr+mt) |
|---|---|---|---|
| d-1 | 0.61 | 26 | 39 |
| d-2 | 0.45 | 31 | 39 |
| d-3 | 0.70 | 18 | 30 |
| d-4 | 0.47 | 26 | 37 |
| d-5 | 0.47 | 39 | 61 |
| d-6 | 0.03 | 44 | 60 |
| average | 0.455 | 30.7 | 44.3 |

Table 2: Dialog scores (in a scale from -1 to 1) and WER (in %) measurements in test 2.

Table 2 presents the computed overall dialog scores along with the WER, which was measured at both the ASR output (WER1) and the MT output (WER2). From Table 2, it may be seen that dialog 3 was consistently the best ranked dialog by all the three given measurements. On the other hand, dialog 6 was the worst ranked dialog according to the computed score and WER1, and the second worst according to WER2. For the remaining dialogs (excepting dialog 5, for which the highest WER2 was obtained) the obtained scores were around 0.5, and WER1 and WER2 ranged from 26% to 31%, and from 35% to 40%, respectively. Notice, however, that automatic scores (WER) and manual scores are not perfectly correlated. In fact, dialogs 4 and 5 obtained the same manual score, but exhibit very different WER values.

## 5. Conclusions

This paper described the acceptance test procedure used to evaluate the LC-STAR speech-to-speech demonstrator platform. The procedure consisted of two independent tests. The first one was an utterance-oriented evaluation, and looked for evaluating how much the use of speech benefits communication. The second one was a task-oriented evaluation and aimed at evaluating if users could achieve a given task with the state of the technology.

From the results of test 1 it was made clear that the ASR constitutes the actual bottle neck of the whole demonstrator

platform. Additionally, results of test 1 also suggested that no significant differences in performance exists between users that are familiar with the technology and users that are not familiar with the technology. It was also clear, that with the exception of scenario B (snack bar service), the system performed similarly well for the other three evaluated scenarios.

From the results of test 2 it can be concluded that although the system performed fairly well in five out of six dialogs (notice from Table 2 that all scores for dialogs 1 to 5 were around 0.5) the usability of the system is not still acceptable since none of the six dialogs achieved the totality of goals. With the exception of the goal related to the credit card number, which was certainly affected by both problems mentioned in the previous section, it may be noticed from Table 1 that in only two of the six dialogs performed the other seven goals were totally achieved. According to this, we can say that state of technology is getting closer to provide effective speech-to-speech translation systems but there is still lot of work to be done in this area. It is also important to mention that results from test 2 allowed to identify two specific problems related to the ASR and the TTS that should be corrected.

## 6. Acknowledgements

This work was partly supported by the LC-STAR project by the European Community (IST project reference number 2001-32216), and by the Spanish Department of Education and Science (MEC).

The authors also want to thank to all the persons who participated in the evaluations.

## 7. References

D. Arnold, L. Sadler, and R. L. Humphreys. 1993. Evaluation: an Assessment. *Machine Translation*, 8:1–24.

A. Bonafonte, J. B. Mariño, A. Nogueiras, and J. A. Rodriguez. 1998. RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC. *VIII Jornadas de Telecom I+D*, pages 28–29, Madrid, Spain.

A. Bonafonte, I. Esquerra, A. Febrer, J. A. Rodriguez, and F. Vallverdu. 1998. The UPC Text-to-Speech System for Spanish and Catalan. *5th International Conference on Spoken Language Processing*, ICSLP'98, Sydney, Australia.

A. Bonafonte, and J. B. Mariño. 1998. Using X-Gram for Efficient Speech Recognition. *5th International Conference on Spoken Language Processing*, ICSLP'98, Sydney, Australia.

D. Carter, M. Rayner, R. Eklund, C. MacDermid, and M. Wiren. 2000. Evaluation. In Rayner, Carter, Bouillon, Digalakis and Wiren, editors, *The spoken language translator*, pages 297–312, Cambridge University Press.

D. Gates, A. Lavie, L. Levin, M. Gavalda, M. Woszczyna, and P. Zhan. 1997. End-to-end evaluation in JANUS: a speech-to-speech translation system. In Maier, Mast and Luperfoy, editors, *Dialog processing in spoken language systems*, pages 195–206, Springer, Berlin.

J. B. Mariño, A. Nogueiras, P. Paches, and A. Bonafonte. 2000. The demiphone: an efficient contextual subword unit for continuous speech recognition. *Speech Communication*, 32(3):187–197.

F. J. Och, and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics* (best paper award), pages 295-302, Philadelphia, PA.

J. Pérez, and A. Bonafonte. 2004. GAIA: Integrated Platform for the Development of Speech Translation Technologies. *UPC Internal Report*, Barcelona.

J. Polifroni, L. Hirschman, S. Seneff, and V. Zue. 1992. Experiments in evaluating interactive spoken language systems. *Proceedings of the DARPA Speech and NL Workshop*, pages 28–31.

T. Sikorski, and J. Allen. 1995. A task-based evaluation of the TRAINS-95 dialog system. Technical report, University of Rochester.

H. Somers, and Y. Sugita. 2003. Evaluating Commercial Spoken Language Translation Software. *Proceedings of the Ninth Machine Translation Summit*, pages 370–377, New Orleans.

K. Thomas. 1999. Designing a task-based evaluation methodology for a spoken machine translation system. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 569–308, University of Maryland.

M. Walker, D. J. Litman, C. A. Kamm, and A. Abella. 1997. PARADISE: a framework for evaluating spoken dialog agents. *AT&T Technical Reports*.