The FAME Speech-to-Speech Translation System for Catalan, English and Spanish

Victoria Arranz
ELDA – Evaluation and Language Resources Distribution Agency
55-57, rue Brillat Savarin
75013 Paris, FRANCE
arranz@elda.org

Elisabet Comelles
TALP Research Centre, Universitat Politècnica de Catalunya
C/ Jordi Girona 1-3
08034 Barcelona, SPAIN
comelles@lsi.upc.edu

David Farwell
Institució Catalana de Recerca i Estudis Avançats
TALP Research Centre, Universitat Politècnica de Catalunya
C/ Jordi Girona 1-3
08034 Barcelona, SPAIN
farwell@lsi.upc.edu

Abstract

This paper describes the evaluation of the FAME interlingua-based speech-to-speech translation system for Catalan, English and Spanish. This system is an extension of the already existing NESPOLE! that translates between English, French, German and Italian. This article begins with a brief introduction followed by a description of the system architecture and the components of the translation module including the Speech Recognizer, the analysis chain, the generation chain and the Speech Synthesizer. Then we explain the interlingua formalism used, called Interchange Format (IF). We show the results obtained from the evaluation of the system and we describe the three types of evaluation done. We also compare the results of our system with those obtained by a stochastic translator which has been independently developed over the course of the FAME project. Finally, we conclude with future work.

1 Introduction

The FAME interlingual speech-to-speech translation system (SST) for Catalan, English and Spanish has been developed at the Universitat Politècnica de Catalunya (UPC), Spain, as part of the recently completed European Union-funded FAME project (Facilitating Agent for Multicultural Exchange) that focused on the development of multi-modal technologies to support multilingual interactions (see http://isl.ira.uka.de/fame/ for details). The FAME system is an extension of the existing NESPOLE! translation system (Metze *et al*., 2002; Taddei *et al*., 2003) to Catalan and Spanish in the domain of hotel reservations. At its core is a robust, scalable, interlingual speech-to-speech translation system having cross-domain portability that allows for effective translingual communication in a multi-modal setting. Although the system architecture was initially based on NESPOLE!, all of the modules have now been integrated on an Open Agent Architecture platform (Holzapfel *et al*., 2003, for details see http://www.ai.sri.com/~oaa). This type of multi-agent framework offers a number of technical features for a multi-modal environment that are highly advantageous for both system developers and users, specially when considering the complex number and nature of the modules that needed to be integrated within the full FAME project (e.g., modules handling image and video processing, information retrieval, topic detection, etc.).

Broadly speaking, the FAME system consists of an analysis component and generation component. The analysis component automatically transcribes spoken source language utterances and then maps that transcription into an interlingual representation. The generation component then maps from interlingua into natural language text and then produces a synthesized spoken version of that text in the target language. The central advantage of this interlingua-based architecture is that in adding additional languages to the system, it is only necessary to develop new analysis and generation components for each new language in order to be able to translate into and out of all the other existing systems.

For both Catalan and Spanish speech recognition, we used the JANUS Recognition toolkit (JRTk) developed at UKA and CMU (Woszczyna *et al*., 1993). For the text-to-text component, the analysis side utilizes the top-down, chart-based SOUP parser (Gavaldà, 2000) with full domain action level rules to parse input utterances. Natural language generation is done with GenKit, a pseudo-unification based generation tool (Tomita, *et al*. 1988). For both Spanish and Catalan, we use a Text-to-Speech (TTS) system fully developed at the UPC, which uses a unit-selection based, concatenative approach to speech synthesis.

The Interchange Format (Levin, *et al*. 2002), the interlingua used by the C-STAR Consortium (see http://www.c-star.org for details), has been adapted for this effort. Its central advantage for representing dialogue interactions such as those typical of speech-to-speech translation systems is that it focuses on identifying the speech acts and the various types of requests and responses typical of a given domain. Thus, rather than capturing the detailed semantic and stylistic distinctions, it characterizes the intended conversational goal of the interlocutor. Even so, in mapping to and from IF it is necessary to take into account a wide range of structural and lexical properties related to Spanish and Catalan.

For the initial development of the Spanish analysis grammar, the already existing NESPOLE! English and German analysis grammars were used as a reference point. Because of this, some efforts needed to be made to overcome important differences between English and German and the Romance languages dealt with. The Catalan analysis grammar, in turn, has been adapted from the Spanish analysis grammar and, in this case, the process has been rather straightforward. The generation grammar for Spanish was mostly developed from scratch, although some of the underlying structure was adapted from that of the NESPOLE! English generation grammar. Language-dependent properties such as word order, gender and number agreement, etc. have needed to be dealt with representationally but on the whole starting with existing structural descriptions has been useful. On the other hand, the generation lexica play a very major role in the generation process and these had to be developed from scratch. Again, however, the Catalan generation grammar has been adapted from the Spanish generation grammar directly with almost no significant complication.

2 Interchange Format

In this section we describe the Interchange Format (IF), the interlingua used in our system. IF is based on Searle's Theory of Speech Acts (Searle, 1969). It tries to represent the speaker's intention rather than the meaning of the sentence per se. In the hotel reservation domain there are several speech acts such as giving information about a price, asking for information about room type, verifying a reservation, etc. Because domain concepts such as prices, room type and reservation are included in the representation of the act, in our interlingua, such speech acts are referred to as Domain Actions (DAs) and they are the type of actions expressed by the sentence. These DAs are formed by different combinatory elements expressing the semantic information that needs to be communicated.

Generally speaking, an IF representation has the following elements:

Speaker's Tag + DA + Arguments

The Speaker's Tag may be *a* for the agent's contributions, or *c* for the client's.

Inside the DA we find the following elements:

- Speech Act: an obligatory element that can appear alone or followed by other elements. Examples of Speech-Acts include *give-information*, *negate*, *request-information*, etc.
- Attitude: an optional element that represents the attitude of the speaker when explicitly described. Some examples are *+disposition*, *+obligation*, etc.
- Main Predication: a compulsory element that represents what we talk about. Examples of these elements are *+contain*, *+reservation*, and so on.
- Predication Participant: optional elements that represent the objects we talk about, for instance, *+room*, *+accommodation*, etc.

The DA is followed by a list of arguments. These elements are expressed by argument-value pairs positioned inside a list and separated by a ",".

By way of example, an IF representation of the sentence in (1) contains all the elements mentioned above.

(1) *Would you like me to reserve a room for you?*

IF: *a*: *request-information+disposition+ reservation+room* (*for-whom=you, who=i, disposition=*(*who= you, desire*), *room-spec=*(*quantity=1, room*))

From this representation we know that the speaker is the agent and that he or she is asking for some information. The attitude expressed here is a desire of a second person singular, i.e., the client. The main predication is to make a reservation and the predication participant is one room.

Having briefly introduced the formalism, we continue with a discussion of the use of IF for a Machine Translation System for English, Spanish and Catalan, and its application to new languages.

3 Evaluation

The evaluation performed was done on real users of the SST system, in order to:

- Examine the performance of the system in as real a situation as possible, as if it were to be

used by a real tourist trying to book accommodation in Barcelona,
- Study the influence of using Automatic Speech Recognition (ASR) in translation.
- Compare the performance of a statistical approach and an interlingual approach in a restricted semantic domain and task of this kind.
- Investigate the relevance of certain standard evaluation methods used in statistical translation when applied to evaluate interlingual translation.

3.1 Evaluation: Data recording and treatment

Prior to the evaluation task *per se*, a number of tasks had to be done to obtain the necessary data. These included dialogue and data recording during real system usage; adapting the translation system to register every utterance from the two different translation approaches and from ASR; recruiting people to play the roles of the users; designing the scenarios for the users; designing the sequence of events for the recording sessions; transcribing all speech data; etc.

Conversations took place between an English-speaking client and a Catalan- or Spanish-speaking travel agent. The number of dialogues to be carried out was 20, and a total of 12 people were recruited for that purpose. Out of these 12 speakers, 10 had no prior knowledge of the task or the system and 2 were familiar with the system. The former (the 10 speakers) participated in 2 dialogues each and the latter (the other 2) participated in 10 dialogues each. That way, each dialogue would resemble a real-situation dialogue where one of the speakers would always be familiar with the task and the system while the other one would not. It should also be added that all English speakers recruited for the evaluation were non-native speakers of the language. We consider this realistic as most of the potential clients to use such a system would actually be from non-English speaking countries. Although some of them had a very high proficiency of English, this was not the case with all of them, and it should be taken into account that the results from Automatic Speech Recognition and Translation have suffered from this.

5 different scenarios were designed per speaker (agent or client) and they were available in all relevant languages (agents' scenarios in Catalan and Spanish and client's in English). Before starting the actual evaluation, speakers were presented with very basic knowledge about the system, like where to click to start/stop recording, where to find the necessary information regarding the scenarios, etc. Computer screens only showed them their respective scenarios and system interface. The system interface provided them with their own ASR output as well as the translation output (both from the interlingual and statistical systems) of the other user's utterances. The former allowed the speakers to check if the ASR had recognised their utterances properly and, if appropriate, to allow the translation to go on or otherwise to intervene before communication failure took place. The latter allowed them to have the two written translations of the other speaker's utterances on their screen even though the synthesiser only provided a spoken version for one of them (the choice of which is based on a very simple algorithm).

Dialogue recording took place in a room set up for that purpose. Speakers were situated separately with their respective computers in such a way that they can only view their own computer screen. Once finished and all the conversations recorded, the following steps were taken to prepare the data for evaluation:

- All speech files were transcribed and concatenated into dialogue units. That is, all utterances were grouped according to the dialogue they belonged to. During transcription, speech disfluencies were also marked and all utterances tagged.
- Reference translations were created for each speaker+dialogue file, so as to be able to evaluate the translations using BLEU and mWER metrics.

3.2 Task-Oriented Evaluation Metrics

A task-oriented methodology has been developed to evaluate both the end-to-end system (with ASR and TTS) and the source language transcription to target language text subcomponent. An initial version of this evaluation method had already proven useful during system development since it allowed us to analyse content and form independently and, thus, contributed towards practical system improvements.

The evaluation criteria used were broken down into three main categories (*Perfect, Ok* and *Unacceptable*), while the second was further subdivided into *Ok+, Ok* and *Ok-*. During the evaluation these criteria were independently applied to *form* and to *content*. In order to evaluate *form*, only the generated output was considered by the evaluators. To evaluate *content*, evaluators took into account both the input utterance or text and the output text or spoken utterance. Accordingly, the meaning of the evaluation metrics varies if they are being used to judge either *form* or *content*:

- **Perfect**: well-formed output (*form*) or full communication of speakers' information (*content*).
- **Ok+/Ok/Ok-**: acceptable output, grading from only some minor *form* error (e.g., missing determiner) or some minor non-communicated information (*Ok+*) to some more serious *form* or *content* problems (*Ok-*).
- **Unacceptable:** unacceptable output, either essentially unintelligible (*form*) or simply totally unrelated to the input (*content*).

3.2.1 Task-Oriented Evaluation Results

The results obtained from the evaluation of the end-to-end translation system for the different language pairs are shown in Tables 1, 2, 3 and 4. After studying the results we can conclude that many of the errors obtained are caused by the ASR component. However, it should be pointed out that results remain rather good since, for the worst of our language pairs (English-Spanish), a total of 62.4% of the utterances were judged acceptable in regard to content. This is comparable to evaluations of other state-of-the-art systems such as NESPOLE! (Lavie *et al*., 2002), which obtained slightly lower results and were performed on Semantic Dialog Units (see below) instead of utterances (UTT), thus simplifying the translation task. The Catalan-English and English-Catalan pairs were both quite good with 73.1% and 73.5% of the utterances being judged acceptable, respectively, and the Spanish-English pair performs very well with 96.4% of the utterances being acceptable.

SCORES	FORM	CONTENT
PERFECT	70.59%	31.93%
OK+	5.04%	15.12%
OK	6.72%	9.25%
OK-	9.25%	16.80%
UNACCEPTABLE	8.40%	26.90%

Table 1: Evaluation of End-to-End Translation (with ASR) for the Catalan-English Pair. Evaluation based on 119 UTTs.

SCORES	FORM	CONTENT
PERFECT	92.85%	71.42%
OK+	4.77%	11.90%
OK	1.19%	7.14%
OK-	0%	5.96%
UNACCEPTABLE	1.19%	3.58%

Table 2: Evaluation of End-to-End Translation (with ASR) for the Spanish-English Pair. Evaluation based on 84 UTTs.

SCORES	FORM	CONTENT
PERFECT	64.96%	34.19%
OK+	15.39%	11.97%
OK	8.54%	14.52%
OK-	5.12%	12.82%
UNACCEPTABLE	5.99%	26.50%

Table 3: Evaluation of End-to-End Translation (with ASR) for the English-Catalan Pair. Evaluation based on 117 UTTs.

SCORES	FORM	CONTENT
PERFECT	64.80%	17.60%
OK+	4.80%	10.40%
OK	12.00%	18.40%
OK-	8.80%	16.00%
UNACCEPTABLE	9.60%	37.60%

Table 4: Evaluation of End-to-End Translation (with ASR) for the English-Spanish Pair. Evaluation based on 125 UTTs.

3.3 Statistical Evaluation Metrics

Evaluation of the end-to-end interlingual speech-to-speech translation system was also carried out along side an evaluation of a stochastic translation system using statistical metrics such as BLEU (Papineni, *et al*. 2001) and mWER. The latter translation system was developed independently but in parallel with the interlingual system over the course of the FAME project. Briefly, the system produces a translation by maximizing the *joint* probability between source and target languages, which is equivalent to a language model of an special language with bilingual units (called tuples). It implements this tuple language model by means of a Finite-State Transducer (FST) considering an Xgram memory, that is, a variable length N-gram model which adapts its length to evidence in the data. Given such a bilingual FST, the search for a translation becomes the search for the best-scoring path among the transducer's edges. For a detailed description of the system and the core processes for training it from a parallel corpus, see (Gispert, *et al*. 2004).

With respect to the interlingual system, we anticipated that results would drop drastically when compared to the manual evaluation presented in Section 3.2 and they did. This considerable drop is due to a number of factors:

- The resulting translation is compared to a single reference translation, which does not cover language variety and flexibility and, thus, worsens results.

- Metrics like BLEU and mWER penalise all diversions from the reference translation, which may be caused by the use of minor form errors that do not harm the end results badly.
- Generally, results drop when the input language is English. This is clearly due to: a) work focused mostly on Catalan and Spanish and these two languages had their modules further developed for the task domain; b) English-speaking volunteers for the evaluation were not native speakers of English, which complicated the task of speech recognition considerably at some points.

3.3.1 Statistical Evaluation Results

Results obtained both from the statistical approach and the interlingua-based approach are shown below in Tables 5 and 6:

Language Pairs	# sentences	mWER	BLEU
CAT–ENG	119	74.66	0.1218
ENG–CAT	117	77.84	0.1573
SPA–ENG	84	61.10	0.1934
ENG–SPA	125	80.95	0.1052

Table 5: Results of the Statistical Translation System

Language Pairs	# sentences	mWER	BLEU
CAT–ENG	119	78.98	0.1456
ENG–CAT	117	81.19	0.2036
SPA–ENG	84	60.93	0.3462
ENG–SPA	125	86.71	0.1214

Table 6: Results of the Interlingua-based Translation System

The results are consistent with respect to the relative performance of the different systems in terms of language pairs. The Spanish-to-English systems, both statistical and rule-based, performed best. The English-to-Spanish systems, both statistical and rule-based, performed the worst. This seems to be due to the more advanced development of the Spanish analysis modules with respect to those of Catalan and English. The Catalan-to-English and English-to-Catalan systems performed somewhere in between with the latter slightly outperforming the former, according to the BLEU scores, but the former outperforming the latter in terms of mWER scores.

As for the relative performance of the statistical systems as opposed to the rule-based systems, the results are entirely contradictory. The mWER scores of the statistical system are consistently better than those of the rule-based systems (apart from the Spanish-to-English case where the two systems essentially performed equally). On the other hand, the BLEU scores of the rule-based system are consistently better than those of the statistical systems. It is unclear how this happened although it is likely that since the BLEU metric rewards overlapping strings of words (as opposed to simply matching words) that the rule-based systems produced a greater number of correct multiword sub-strings than the statistical systems did. In any case, were it not for the low performance of all the systems and the very limited size of the test corpus, this would be a very telling result with regard to the validity of the evaluation metrics.

3.4 User Satisfaction Evaluation

Finally, a used-oriented evaluation of a system that combined both interlingual and statistical systems was also carried out from both a quantitative as well as a qualitative perspective.

3.4.1 Quantitative Study

The quantitative study of the user satisfaction consists in measuring the results obtained from the end-to-end translation system according to a number of metrics established for that purpose. Metrics 1 and 3 are used as reference point for the assessing results obtained for metrics 2 and 4, respectively:

1. *Number of turns per dialogue*: This is the total number of contributions by both participants per dialogue.
2. *Success in communicating the speaker's intention/Successful turns per dialogue:* This is the total number of successful contributions.
3. *Number of items of target information per dialogue*: For each turn a speaker may wish to communicate 1 or more items of information. This number is always higher than the number of turns. An *item of target information* roughly corresponds to a speech act. For instance, for an utterance such as "*Good morning*, *I'd like to make a hotel reservation*," there are two items of target information: a *greeting* and *reservation request*. These items of target information have been established according to the criteria followed by the interlingua used (cf. Section 4.1).
4. *Successful items of target information obtained*: This measures the number of speech acts that were successfully performed by both users (agent and client).

5. *Number of disfluencies per dialogue*: This refers to the number of infelicitous expressions uttered by the users, covering mostly clicks (wrong clicks of the mouse when using the system platform), pauses, hesitations and mistakes.
6. *Number of repetitions per dialogue*: This is the number of turns in which users must repeat themselves in order to achieve their goal.
7. *Number of abandoned turns per dialogue*: This is the number of goals that end up unattained, usually after several repetitions. However, sometimes users abandon goals after just one utterance possibly because they feel uneasy about repeating themselves for a machine or because they simply feel nervous. As can be seen below, this number is very low.

Before showing the table, results obtained from metrics 2 and 4 should be further explained given that they appear to provide much lower results than they actually do. The success obtained both at a turn level and at an item of target information level are shown in a global way, that is, taking into account the full number of repetitions (which are included in reference metrics 1 and 3). Thus, a dialogue may be successful by means of some repetitions while the number of successes measured by metrics 2 and 4 is relatively low. A more complete way to evaluate the success of a given dialogue is to also look at metric 7, which indicates the number of abandoned turns and, thus, reflects the number of goals the speakers failed to attain.

Last but not least, and as already mentioned in Section 3.1., users playing the role of the English-speaking client were not native speakers of English, which certainly makes the task of speech recognition even more difficult. This is particularly so in some dialogues where the speakers do not master the language at all. Table 7 shows the results obtained with the above metrics:

Dialogues	M-1	M-2	M-3	M-4	M-5	M-6	M-7
Eng/Spa-1	7	7	11	10	0	0	0
Eng/Spa-2	24	13,5	33	21,5	0	4	2
Eng/Spa-3	24	19	36	27	2	1	1
Eng/Spa-4	12	7,5	22	15	1	2	1
Eng/Spa-5	22	16,5	36	28	2	4	1
Eng/Spa-6	18	7	18	8	5	9	0
Eng/Spa-7	9	6	14	10	3	1	0
Eng/Spa-8	32	14,5	42	21	0	10	3
Eng/Cat-9	23	9,5	33	15,5	1	12	1
Eng/Cat-10	40	20,5	52	26,5	3	17	0
Eng/Cat-11	20	9	34	16	1	8	1
Eng/Cat-12	37	16	44	18,5	1	15	1
Eng/Cat-13	6	5,5	12	11	1	1	0
Eng/Cat-14	37	18,5	48	27	0	14	4
Eng/Spa-15	25	8	35	15	0	13	2
Eng/Spa-16	37	24	52	35,5	2	9	2
Eng/Cat-17	11	5,5	23	12	0	5	0
Eng/Cat-18	31	15	43	24	1	14	1
Eng/Cat-19	7	4,5	13	9	1	2	0
Eng/Cat-20	23	14,5	32	20,5	1	9	1

Table 7: User Satisfaction Results

As it can be observed in M-7, 7 dialogues have successfully communicated all information; 8 dialogues have only given up on one turn, and 3 dialogues on 2. The remaining 2 dialogues have abandoned 3 and 4 turns, respectively. This is not an important loss, bearing in mind that after analysing the results, it was observed that often problems are related to non-central tasks such as greetings and thanking.

3.4.2 Qualitative Study

The quantitative study presented above has been supplemented by a qualitative evaluation based on users' responses to a brief questionnaire. Below we present a breakdown of users' opinions of the system on both a per question and per questionnaire basis. The average response is 3.4 points on a 5 point scale where 0 is least agreement and 5 is greatest agreement.

Results per question

Question 1: I understood the information the system passed on to me.

3 points out of 5 → 30%
4 points out of 5 → 40%
5 points out of 5 → 30%
} Average 4.0 pts

Question 2: The system understood what I told it to pass on.

2 points out of 5 → 10%
3 points out of 5 → 60%
4 points out of 5 → 30%
} Average 3.0 pts

Question 3: At each point during the interchange I understood what I could say.

2 points out of 5 → 20%
4 points out of 5 → 30%
5 points out of 5 → 50%
} Average 3.6 pts

Question 4: The dialogue was normal and natural.

2 points out of 5 → 30%
3 points out of 5 → 20%
4 points out of 5 → 10%
5 points out of 5 → 40%
} Average 3.5 pts

Question 5: I succeeded in getting what I wanted done.

3 points out of 5 → 50%
4 points out of 5 → 20%
5 points out of 5 → 30%
} Average 4.0 pts

Question 6: I would use this system again to help reserve a hotel room.

1 point out of 5 → 20%
2 points out of 5 → 10%
3 points out of 5 → 40%
4 points out of 5 → 20%
5 points out of 5 → 10%
} Average 3.0 pts

Question 7: The system behaved as expected.

2 points out of 5 →	20%	Average 3.0 pts
3 points out of 5 →	40%	
4 points out of 5 →	40%	

Question 8: The system allowed me to easily correct any errors that arose.

4 points out of 5 →	70%	Average 4.5 pts
5 points out of 5 →	30%	

Question 9: The dialogue was very long.

1 point out of 5 →	40%	Average 2.0 pts
2 points out of 5 →	40%	
3 points out of 5 →	20%	

Question 10: I had trouble with turns about:

Hotel names →	12.50%
Hotel categories →	25.00%
Room Types →	18.75%
Dates →	25.00%
Prices →	6.25%
Other – Greetings →	12.50%

Average for questions 1-9: **3.4 pts out of 5.0**

Questions 1, 2 and 3 address the user's impression of the adequacy of the process of information exchange. While users generally thought the system's contributions were coherent (4.0), they were less convinced that the system understood what they were saying (3.0). This is likely due to the need to repeat themselves when the system fails to provide a translation for their contribution. Still, on the whole the users knew where they were during the interchange and how they should be saying what they wanted to say (3.6). The results to these questions also appear to reflect that 2 of the 10 users were less comfortable and had greater difficulty during the exchange.

Questions 4, 7, 8 and 9 address the user's sense of the fluidity of the interchange and their ability to recover from errors. Again, on the whole they thought the interaction was natural (3.5) and straightforward (2.0 on Question 9). In addition, it appears that the users were quite satisfied with their ability to overcome system errors through some sort of corrective dialogue (4.5). However, given that in general the system did not behave completely as expected (3.0), users were clearly not completely comfortable with the fluidity of the interchange.

Question 10 is designed to identify the user's impression of their ability to carry out particular types of subtasks within the reservation task. The results indicate that hotel categories (one-star, two-star, etc.) and dates gave them the greatest problems (25%) while hotel names (12.5%) and hotel prices (6.75%) seemed to be the least problematical.

Finally, Questions 5 and 6 address the user's impression of their ability to successfully carry out the reservation task using the system. Most felt they accomplished the task before them (4.0) but clearly their impression of the utility of the system was less than enthusiastic. When asked if they would use the system again to make a hotel reservation the response varied a good deal and the average was a mere 3.0. It must be remembered that the Spanish and Catalan users were quite familiar with English and that the English users, as mentioned above almost all spoke Spanish and/or Catalan. If Chinese or Bantu or some other less well known language had been selected for Client, the willingness of the users to use the system in future would no doubt have been rather greater.

Results per questionnaire:

- Questionnaire 1 →	3.5 points out of 5
- Questionnaire 2 →	3.1 points out of 5
- Questionnaire 3 →	2.5 points out of 5
- Questionnaire 4 →	3.6 points out of 5
- Questionnaire 5 →	4.3 points out of 5
- Questionnaire 6 →	3.4 points out of 5
- Questionnaire 7 →	3.2 points out of 5
- Questionnaire 8 →	3.4 points out of 5
- Questionnaire 9 →	3.7 points out of 5
- Questionnaire 10 →	2.8 points out of 5

An informal inspection of the results per questionnaire indicates that the reaction of the users as a whole was consistent and weakly positive (taking 3.0 as a median). There was one user (Questionnaire 5) who had a very positive experience and two users (Questionnaires 3 and 10) who had a somewhat negative experience.

4 Conclusions

This article has described the FAME interlingua-based speech-to-speech translation system for Catalan, English and Spanish and the different evaluations performed on real users. Three different types of evaluation have been carried out so as to check a) the performance of the system, b) the influence of ASR in translation, c) the comparison in performance of the interlingua and the stochastic systems developed within FAME for the domain and task set, and d) the relevance of certain standard evaluation metrics used in statistical translation when applied to interlingual translation.

The different evaluations prove that the system is already at an interesting and promising stage of development. In addition to these evaluations, a

public demonstration of the system took place during which non-familiar users participated and tested the system. Results from this open event were also very satisfactory.

Now that this basic level of development has been reached, our next step is to solve remaining technical problems and expand the system both for this domain and for others. As for the technical problems, we need to focus on improving the ASR component, which seems to be an important source of errors. To this end, further domain-specific data is to be collected so as to develop better language models. We also need to deal better with degraded translations. One option is to incorporate certain recovery strategies within the dialogue model which will allow speakers to request repetitions or reformulations from their counterparts.

Last but not least, a detailed study of the problems that have arisen during the process of applying IF to the Romance languages used in this system was carried out (Arranz, *et al*. 2005). It has resulted in a number of proposed changes and improvements in the IF which we expect to implement as part of the next stages of system development.

5 Acknowledgements

This research has been partially financed by the FAME (IST-2001-28323) and ALIADO (TIC2002-04447-C02) projects. We would also like to thank, very specially, Climent Nadeu and Jaume Padrell, for all their help and support in numerous aspects of the project. We are also grateful to other UPC colleagues, José B. Mariño and Adrià de Gispert, for fruitful exchanges during the development of both SST systems. Last but not least, we are strongly indebted to our colleagues at CMU, Dorcas Alexander, Donna Gates, Lori Levin, Kay Peterson and Alex Waibel, for all their feedback and help.

References

Arranz, V., Comelles, E., Farwell, D. 2005. Speech-to-Speech Translation for Catalan. To be presented at the Conference on Lesser-Used Languages, Bolzano, Italy, Oct 27-28, 2005.

Gavaldà, M. 2000. SOUP: A Parser for Real-world Spontaneous Speech. In *Proceedings of the 6th International Workshop on Parsing Technologies (IWPT-2000)*, Trento, Italy.

Gispert, A., Mariño, J.B., Crego J.M., 2004. TALP : Xgram-based Spoken Language Translation System. In *Proceedings of the International Workshop on Spoken Language Translation*. Kyoto, Japan.

Holzapfel, H., Rogina, I., Wölfel, M., and Kluge, T. 2003. FAME Deliverable D3.1: Testbed Software, Middleware and Communication Architecture.

Lavie, A., Metze, F., cattoni, R., Constantini, E. 2002. A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System. In *Proceedings of ACL-2002 Workshop on Speech-to-Speech Translation: Algorithms and Systems*. Philadelphia, PA.

Levin, L., Gates, D., Wallace, D., Peterson, K., Lavie, A., Pianesi, F., Pianta, E., Cattoni, R., Mana, N. 2002. Balancing Expressiveness and Simplicity in an Interlingua for Task based Dialogue. In *Proceedings of ACL-2002 workshop on Speech-to-speech Translation: Algorithms and Systems*. Philadelphia, PA.

Metze, F., McDonough, J., Soltau, J., Langley, C., Lavie, A., Levin, L., Schultz, T., Waibel, A., Cattoni, L., Lazzari, G., Mana, N., Pianesi, F., Pianta, E. 2002. The NESPOLE! Speech-to-Speech Translation System. In *Proceedings of HLT-2002*. San Diego, California.

Papineni, K., Roukos, S., Ward, T., Zhu, W. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Division.

Searle, John. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press: Cambridge, UK.

Taddei, L., Besacier, L., Cattoni, R., Costantini, E., Lavie, A., Mana, N., Pianta, E. 2003. NESPOLE! Deliverable D17: Second Showcase Documentation. See the NESPOLE! Project web site: http://nespole.itc.it.

Tomita, M., Nyberg, E.H. 1988. Generation Kit and Transformation Kit, Version 3.2, User's Manual. *Technical Report CMU-CMT-88-MEMO*. Center for Machine Translation, Carnegie Mellon University, Pittsburgh, PA.

Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., Ward, W. 1993. Recent Advances in JANUS: A Speech Translation System. In *Proceedings of Eurospeech-1993*. Berlin.