Communicative Strategies and Patterns of Multimodal Integration in a Speech-to-Speech Translation System

Susanne Burger — CMU-ISL, Pittsburgh, PA, U.S.A., sburger@cs.cmu.edu

Erica Costantini — ITC-irst, Trento, Italy, costante@itc.it

Fabio Pianesi — ITC-irst, Trento, Italy, pianesi@itc.it

Abstract

When multilingual communication through a speech-to-speech translation system is supported by multimodal features, e.g. pen-based gestures, the following issues arise concerning the nature of the supported communication: a) to what extend does multilingual communication differ from 'ordinary' monolingual communication with respect to the dialogue structure and the communicative strategies used by participants; b) the patterns of integration between speech and gestures. Building on the outcomes of a previous work, we present results from a study aimed at addressing those issues. The initial findings confirm that multilingual communication, and the way in which it is realized by actual systems (e.g., with or without the push-to-talk mode) affects the form and structure of the conversation.

1 Introduction

Multilingual human-human communication is the topic of many recent research work (e.g., NESPOLE!, C-STAR, LC-STAR) which have in some cases extended to the broader issue of multimodality. The level of communication effectiveness achieved (and achievable) remains largely untouched. In this respect, the following aspects deserve particular attention: a) To what extent does multilingual communication through a speech-to-speech translation (STST) system differ from 'ordinary' monolingual communication, with respect to its dialogue structure and to the participants' communicative strategies? b) Which patterns reflect the integration between speech and gestures, and what are the possible advantages of this integration for multilingual communication?

A better understanding of this issue would significantly contribute to research and development efforts in STST. It would help in modeling appropriate conversational structures for STST systems, and could emphasize on the importance of multilingual training corpora for STST systems.

The NESPOLE! project offered an opportunity to investigate the mentioned aspects. Jointly funded by EU/NSF, NESPOLE! was designed to provide fully functional STST capabilities within real-world settings for common users involved in e-commerce applications. It exploits a client-server architecture to allow an English-, French- or German-speaking user, who is browsing through the web pages of a service provider on the Internet, to connect seamlessly to an Italian speaking human agent. Commercially available PC video-conferencing technology is used to connect the two parties in real-time. The communication is mediated by NESPOLE!'s STST services, which exploit an Interlingua-based approach to translation, using the Interchange Format (Levin et al., 2002) as an intermediate representation. Several multimodal features such as video contact, the possibility of the transfer of picture material on web pages and maps and pointing on map details by transferring drawing gestures support the dialogue. More information on design principles of NESPOLE! can be found in Lavie et al., 2001.

Two user studies have been conducted so far within the NESPOLE! project. We will summarize the results of the first study (see also Costantini, Pianesi and Burger, 2002), and will especially focus on the second user study.

2 User Study 1: Multimodal vs Speech-Only Conditions

Previous research demonstrated that, when interacting with spatial tasks, the performances of

users sensibly improve when multimodal input is available (Oviatt, 1997). These results were obtained in highly controlled experimental conditions in a monolingual setting using the Wizard of Oz technique. We designed and conducted an experiment (Costantini, Pianesi and Burger, 2002) to test how far these results could be replicated by replacing the wizard with a "real" system for multilingual human-human communication via Internet.

14 German-speaking and 14 English-speaking novice users interacted with seven Italian-speaking travel agents through the first prototype of the Nespole! system using it in a push-to-talk mode.

We compared two conditions: SO (speech-only), which allowed only spoken input, and *MM (*multimodal*),* where users were allowed to use pen-based gestures to select or point at portions of a map to support the conversation.

The most relevant results were the following:

- Multimodal interaction did not affect the dialogue length, the number of spoken turns and words, and the number of disfluencies and spontaneous phenomena.
- When the dialogue partners talked about spatial information, dialogues with MM input were clearly more successful than SO dialogues: the number of ambiguities, repetitions and non-successful turns was decreased; misunderstandings were faster resolved, preserving the dialogue fluency.
- Subjects performed only a low number of gestures (one gesture per 10 spoken turns; almost all came from agents), not enough to have a significant impact on global dialogue variables.
- Pen-based gestures always followed the verbal contribution instead of occurring simultaneously. Few or no deictic expressions were used. Together, these suggest a low level of integration between speech and gestures.
- When the agents, who were involved in both parts of the experiment were explicitly asked to express a preference between the MM and the SO condition, they showed a clear preference for the MM.

Two main issues concerning the integration of multilingual and multimodal communication were left open by this study.

1) Impact of technique: how significant is the impact of the specific STST system itself with all its delays, translation errors and technical problems upon the way speech and gestures are integrated. How significant is the impact of the push-to-talk mode (PTT)?

2) Dialogue effectiveness: analyzing only dialogue length, number of disfluencies and of turns, and vocabulary counts, as well as "classical" measures such as task accomplishment and translation successfulness seemed not to be sufficient enough to show interesting differences on the level of dialogue structure.

These considerations built the basis for the design of a further user study.

3 User Study 2: Multilingual vs Monolingual Conditions

The second user study aimed at
a) explicitly comparing multilingual dialogues with monolingual dialogues, with and without PTT, and
b) adopting a more structured conversation analysis.
This resulted in the following three experimental conditions:

- **STST condition**: multilingual (English/ Italian), using the STST system as translation, push to talk mode;
- **PTT condition**: monolingual (Italian/Italian), push to talk mode;
- **Non-PTT condition**: monolingual (Italian /Italian), free talk without push to talk.

We did not extend the multilingual condition to other language pairs, since previous studies did not reveal any important cross-linguistic difference (Costantini, Pianesi and Burger, 2002).

We expected the multilingual condition to be different from the monolingual conditions with respect to dialogue length, spoken input features, dialogue structure and speech-gesture integration patterns. In addition we hypothesized that the PTT mode used in the multilingual condition could play a role in determining those results, so that differences could be found between the two monolingual conditions.

3.1 Scenario and Data Collection

The scenario featured a customer browsing the web pages of an Italian tourist board office, searching for information about winter holidays in Val di Fiemme, Trentino, Italy. Customers could access detailed information by clicking a special button, which opened a direct connection with a human agent. The customer's task was to choose

an appropriate location and an all-inclusive tourist package within the constraints specified a priori, concerning the relevant geographical area, the available budget, etc. The agent's task was to provide the requested information following the available descriptive cards. Customers and agents both received written information and instructions about the scenario, the task, system functionalities and interaction modalities.

For the STST condition seven English customers located in Pittsburgh interacted with three tourist agents located in Italy through the final version of the NESPOLE! system, resulting in seven recorded dialogues. Participants wore a head-mounted microphone, using it in a push-to-talk mode. Each participant could hear only the message of the party as translated by the system, and had no cues about the original.

The same three agents acted as agents again in 16 additional monolingual dialogues: half of these dialogues were recorded in PTT mode (PTT condition) and the other half in free speaking style (Non-PTT condition). The role of the customer in the monolingual dialogues was played by 16 native Italian volunteers. Since it was too difficult to get 16 Italians connected from Pittsburgh, customers and agents were recorded in Italy. This resulted in better network connections and very limited transfer delays.

The interface screen used by agents as well as customers displayed four windows: the Netmeeting® window, displaying a live video of the other party allowed visual contact; the WhiteBoard window, where images and pen-based gestures (to select, point at, or highlight portions of the displayed image) could be shared; and, for the multilingual condition, two windows providing visual and textual feedback concerning the translation process. A more detailed description of the interface is available in Taddei, Costantini and Lavie, 2002.

For each dialogue, an audio file containing the contributions of both speakers was recorded at each side. In STST condition, each file contained the original voice of the local speaker and the other party's translated and synthesized messages. All the audio files were transcribed according to the VERBMOBIL conventions[1], using the TransEdit[2] annotation tool. Aside from orthographic words, transcription files contained annotations for spontaneous phenomena. Gestures were manually annotated using videos of agents recorded in Italy. In addition, all dialogues were annotated following a dialogue structure annotation schema (see below). The speaker may repeat her utterance to overcome system errors or misunderstandings, and so turn repetitions were counted as well.

3.1.1 Dialogue Structure Annotation Schema

In order to assess the dialogue structure, we resorted to the Dialogue Structure Coding Scheme (DSCS) from the HCRC (Human Communication Research Centre[3]). DSCD differs from previous coding schemes by boasting higher task independence than other contemporary schemes (Carletta et al., 1996; Carletta et al., 1997). To this end, DSCS attempts to both classify single utterances according to their discourse goals and capture the higher-level structure of dialogues in terms of their so-called *game* structures. Conversational games are associated with mutually understood conversational goals, such as obtaining information or convincing a partner to perform an action. A dialogue game is a set of utterances. It starts with an 'initiation', i.e., a turn that sets up expectations, possibly followed by 'responses' which are turns fulfilling those expectations. A dialogue game encompasses all utterances until the purpose has been fulfilled, e.g., the requested information has been transferred or abandoned. DSCS allows structuring of games into nested sub-games. Finally, games consist of conversational *moves* which are different kinds of initiations and responses classified according to their purposes, e.g. opening, checking, affirmative replies, etc.

Although devised for the Map Task Corpus (Anderson, 1991), DSCS designers intended it to apply to other types of task-oriented dialogue but were also aware that it did not probably exhaust the speakers' repertoires and therefore can be extended. Since our complex scenario demanded coverage of a higher number of phenomena, we modified the DSCS by introducing new moves. The following table shows the modified schema. A star "*" marks those moves newly added to the DSCS schema. The *proposal, disposition, action*

[1] http://www.is.cs.cmu.edu/trl_conventions/

[2] For more information: sburger@cs.cmu.edu

[3] http://www.hcrc.ed.ac.uk/Site/

and information moves are subclasses of the former *information* move.

Move	Explanation
1. Initiating	introduces a new discourse purpose into the dialogue
Align	checks transfer successfulness
Check	checks confirmation of correct understanding or inference
Query-yn *Query-w*	yes/no questions (*yn*), open questions (*w*)
Request	requests (former *instruct* move), e.g. "could you show me a map?"
Proposal	proposal or offer
Disposition	needs or interests, e.g. "I'm interested in skiing"
Action	description of actions, e.g. "I selected the hotel with a circle"
Information	Not elicited, spontaneously provided information
2. Response	fulfils the expectations set up within the game
Acknowledge	confirming, communication success
Reply-y, *Reply-n*, *Reply-w*, *Reply-amp*	yes/no answers, answers to open questions (*w*), answers adding not requested information (*amp*, former *clarify* move)
**Problem*	negative feedback (notification of non-successful communication)
**Other*	answers where the speaker misunderstood the question and talked about different things
Preparation	expressing readiness to start
**Comment*	out of domain comments (partially overlapping with the former *uncodable* label).
**Noise*	turns with no linguistic content, e.g. made by words interrupted because of technical problems

Table 1. Dialogue Annotation Schema

Another secondary annotation was added to the moves: this annotation aimed to inform whether a move was continued, abandoned, repeated, reformulated, and if it concerned technical issues (e.g. bad audio) or multimodal issues.

4 Results: Speech Input

4.1 Dialogue Length, Turns and Words

The total number of spoken turns, word-tokens and word-types (used vocabulary) was counted for each dialogue. In STST and PTT condition, a turn was operationally defined as a speaker contribution between a switching-on and a switching-off of the microphone button. In Non-PTT condition a turn was defined as any speaker contribution. Speakers usually ended their contribution by showing prosodic cues and semantic features. Transcribers followed the definition of turn as given by the VERBMOBIL transcription scheme. In cases of ambiguity, there may still be a certain degree of freedom as to where a transcriber set a turn boundary. Word-tokens are occurrences of a given word-type, e.g. the sentences "Paul is the brother of John" and "John is the brother of Paul" contain 12 word-tokens and 6 word-types.

The collected corpus consists of a total number of 18100 word tokens. The average duration of a dialogue was 23 minutes for the STST condition, and 9.85 for PTT condition, and 8.87 minutes for Non-PTT condition. The difference in dialogue duration between monolingual and multilingual conditions is mainly attributable to two factors: (1) The time needed for the process of automatic translation and (2) the Internet's rate of information transfer. In the case of STST condition, silence, translation and speech synthesis account for 87% of the dialogue duration; in the monolingual PPT condition 49% of the dialogue duration shows silence and transfer. In Non-PTT dialogues this is reduced to only 19%. Clearly, the long waiting time significantly slowed down the conversation in STST. Moreover, an effect of PTT emerges.

Figure 1 shows the average number of word-tokens per speaker, per dialogue in the three conditions. Word tokens are divided into proper names (names), content words (content), and function words (func). Besides the lower number of tokens in STST condition, the diagram shows a clear tendency for agents to speak more than customers, which is more evident in the monolingual conditions. In addition, the results for PTT condition are somewhat intermediate between those for STST and Non-PTT condition, indicating that the PTT already has an effect in the

monolingual case, so that STST condition is affected both by the PTT mode, and by the characteristics of the STST system.

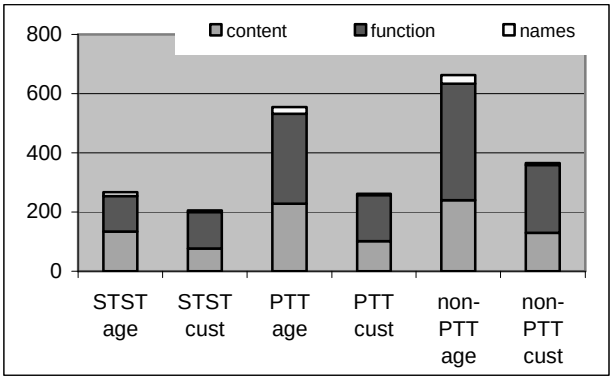

Figure 1. Average number of word-tokens for the three conditions, for agent and customer.

4.2 Results: Dialogue Structure

We counted the frequencies of *games* per each dialogue, finding an average number of 13 games per dialogue in the STST, 14 for the PTT and 17 for the Non-PTT condition. In addition, we calculated the number of *moves* per each *game*, finding an average of 4.6 moves per game for the STST, 4.6 for PTT and 5.6 for the Non-PTT condition: *games* tend to be shorter in the dialogues recorded with PTT procedure and longer in the monolingual dialogues without PTT. There is a trend towards fewer nested games (games embedded within another game) in the STST condition (10% of the games) than in the monolingual conditions (26% in PTT and 23% in the Non-PTT condition), revealing a more complex structure in the monolingual dialogues.

Moves with similar functions were grouped together in broader categories: five moves that included direct and indirect questions formed the category "query" (*query-yn*, *query-w*, *request*, *proposal*, *disposition*); six moves providing information of different types were classified under "information" (*reply-y*, *reply-n*, *reply-w*, *reply-amp*, *information*, *other*). Another category includes the two moves *check* and *align*, which aim to check for comprehension and transfer success, respectively. The moves *acknowledgement* (acceptation), *action* (actually description of an action or gesture) and *ready* (preparation) were kept as single moves. The other three moves (*noise*, *comment*, *problems*) occurred less frequently (under 5%) and were therefore classified as "other" (see figure 2).

Figure2 shows no relevant cross-conditional differences for categories with lower frequencies. The percentages for turns that provide information are also similar (around 30%) in all conditions. On the other hand, there is a clear trend towards a higher number of queries in STST condition (35%) than in the monolingual conditions, with intermediate values for PTT (23%) and a lower value for Non-PTT (14%). Noticeably, STST condition is the only condition having approximately the same number of moves that request information and moves that provide information, while in the monolingual conditions the frequency of the moves that request information is lower than that of moves that provide information. This suggests that the amount of spontaneously offered, not elicited information is higher in the monolingual than in the multilingual conditions. The picture is confirmed considering the frequencies for the *information* move (marking not elicited information): 8% of all the moves in STST, 12% in PTT and 15% in Non-PTT condition.

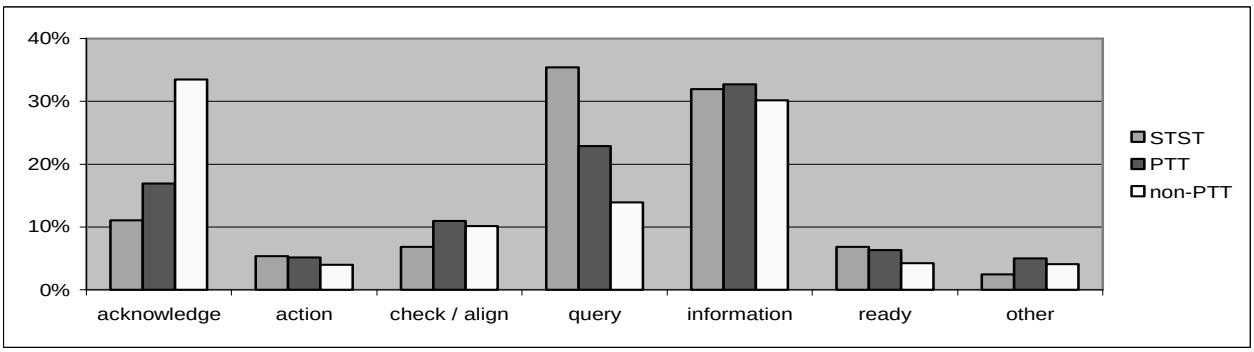

Figure 2. Percentages of move categories for the three conditions.

Figure 2 also shows that the acknowledge moves are more frequent in Non-PTT (33%) than in PTT (17%) or STST (11%). This could be mainly due to a higher preference for ending a game with an 'acknowledge' in Non-PTT condition. Indeed, 66% of the games of Non-PTT condition end with an acknowledgment while the figures for PTT and STST condition are 38% and 23%, respectively. The information moves show an opposite trend: 25% of the games end of non PTT condition ends with a reply or information move, 52% in PTT and 50% in STST condition. None of the remaining moves closed a game with a frequency higher than 8%.

In conclusion, these preliminary results show that there are specific features in multilingual communication that affect communication styles. Analysis techniques investigating dialogue structure are appropriate tools for revealing them.

5 Results: Gestures

The term gesture refers to all WhiteBoard (WB) commands concerning shared maps and web pages (Taddei, Costantini and Lavie, 2002): loading images, running a web browser, scrolling images, zooming images, free-hand strokes, selection of areas and lines on the map. The first four classes are multimedia commands that allow the exchange and exploration of visual information. The latter three are drawings marked by a pointing device that involve the deictic/referential use of image portions, indicating relevant locations, connecting different places, etc.: hence, they directly contribute to the contents of the interaction.

The average number of gestures per dialogue was similar in all three conditions (12.9 in STST, 13.6 in PTT, and 13.7 in Non-PTT condition); about half were drawings. Web pages were not used at all, most likely because the two available web pages contained information not seen as crucial. Zoom was also never used.

We annotated three classes of temporal integration patterns between gestures and speech: (a) immediately before, (b) during, or (c) immediately after the corresponding speech turn. The following table reports the percentages for each category.

The figures in the table 2 are not separated for agents and customers, since most of the gestures were performed by the agents (98% in STST, 92% in PTT and 86% in Non-PTT condition).

	STST	PTT	Non-PTT
Before	32%	8%	0%
During	14%	61%	96%
After	53%	31%	4%

Table 2. Percentages of turns performed before, during or after the corresponding turn.

In STST, about half of the gestures followed the speech (53%), with the content of the turn often anticipating the gesture, e.g., "I'm going to send you a map," "I'll show you the ice skating rink on the map." Then the switching-off of the microphone followed, and, finally, the gesture performance. In addition, a significant number of gestures (32%) were performed before speech: however, all but two were multimedia commands (map loading or closing and scrolling). The majority of these cases follows a certain pattern: The agent loads a map and eventually scrolls (one or two gestures *before* speech); she switches on the microphone to explain the map and verbally anticipates the subsequent drawing gestures, e.g., "This is the map of Val di Fiemme. There are three hotels in Val di Fiemme, I'm showing them to you on the map with black circles." Then the agent switches off the microphone and performs the anticipated drawing gestures. A limited number of gestures were performed during the ongoing turn (13%), specifically, while the subject was speaking, leaving the microphone switched on. All of those latter gestures were drawing gestures (elliptical and rectangular selection and lines).

Interestingly, in the monolingual dialogues the number of gestures performed during speech drastically increases. In particular, in the non-PTT condition (assumedly closer to a 'natural' dialogue condition) almost all the gestures were performed during speech. PTT condition is somewhat intermediate: a higher number of gestures during speech than STST, but a lower number than non-PTT condition. This confirms further that the presence of PTT requires adaptations by the users, resulting in multimodal integration patterns that are distinct from those found in 'natural' conversations (Non-PTT condition).

6 Additional Results for the STST System

For the multilingual dialogues a translation success index was calculated. We asked three bilingual graders to judge each spoken turn using three categories of success by comparing them with their translation and the relative reply: *successful* turns were turns with grammatically and semantically accurate translation; *non-successful* turns contained no comprehensible components from the original utterance, or no translation at all; *partially successful* turns had poor or bad translations, either because of grammatical or syntactical errors, or because some words were badly translated or not translated at all; at the same time, the translation conveyed enough of the original message to enable the targeted party to react acceptably.

We used a majority score for each category, i.e. for each turn we adopted the success category negotiated by at least two graders. In cases of total disagreement, the turn was labelled 'disagreement'. Graders did not reach an agreement on 3% of the graded turns. Among the remainder, successful turns constitute 33% of the original turns, partially successful turns 32%, and non-successful turns 35%.

In addition, we counted turn repetitions, turns during which the speaker repeated or reformulated an utterance to overcome misunderstandings or system failures. A low number of turn repetitions may be considered as a further index of turn success. Speakers repeated 15% of turns at least once to overcome system errors (repeated turns). Each repeated turn was repeated, on average, 1.6 times. Turn repetitions, the subsequent utterances of repeated turns, made up 24% of the turns (not counting the first instance of the turn): this means that almost one quarter of all spoken contributions were repetitions of already uttered turns. In the monolingual conditions the percentages of repetitions or reformulations of previously uttered turns were much lower: 6% in the PTT condition and 1.3% in the Non-PTT condition, suggesting that the high percentage found in the STST condition is mainly due to translation errors.

After being repeated, 32% of the repeated turns were successful and 47% were judged as partially successful. Another group of turns was still judged as non-successful even after being repeated (22% of the repeated turns). This means that the speaker had to surrender to system difficulties and gave up.

Most of the unsuccessfully repeated turns were due to limitations of the system in dealing with meta-communicative concepts. In particular, questions from the customer asking for clarification concerning the agent's previous turn were poorly managed, e.g. "Is the hotel selected in green?", "Is this the map of Cavalese?" (a kind of *check* move). These types of questions were mainly used to ask for confirmation when the content of the received translated turn was not completely understood; this condition is difficult to find in monolingual dialogues. NESPOLE!'s training set consisted exclusively of monolingual data, hence the trained system was unable to adapt. This illustrates the importance for STST systems of closely considering the phenomena arising in the real contexts of the interaction. Training data must be obtained from scenarios as close as possible to a scenario of effective use, here multilingual scenarios.

7 Discussion and Conclusions

By comparing multilingual (STST) dialogues with monolingual dialogues (both in PTT and in Non-PTT mode), we found that the STST system:

- dramatically slows down the conversation;
- reduces the number of words spoken per dialogue, especially for agents.

As for dialogue structure, the STST dialogues are characterized by:

- shorter dialogue games than in Non-PTT condition ;
- fewer nested games than in the monolingual conditions;
- more direct and indirect questions, and less spontaneously offered, not explicitly requested information;
- lower number of *acknowledgment* moves in the multilingual condition, which, in turn, is due to a preference to end games as soon as the information is provided, instead of adding an acknowledgment.

Those data suggest that in the STST dialogues the speakers focus on 'essential' information, reducing dialogue complexity (number of nested games) and try to adhere to a question/answering pattern.

As far as gestures are concerned, we observed:

- a similar number of gestures performed in all conditions;

- a clear trend for gestures to be more often associated with speech in the monolingual non-PTT condition than in the others.

This shows that strict speech-gesture integration is quite a delicate feature that can be lost as soon as more tasks are to be handled in parallel or the context of the interaction becomes more difficult because of the PTT, delays, etc

As a general remark, the overall results for the monolingual-PPT condition were ususally intermediate between those of the monolingual, free-speech condition and those of the multilingual condition, suggesting that the latter is affected both by the characteristics of the STST system itself, and by the PTT mode.

Despite its preliminary status, the reported results show the existence of adaptive communication strategies to the different context of multilingual communication. In this respect, methods addressing the dialogue structure can help us understand and clarify the phenomena. The exclusive usage of the rather classical evaluation methods (based on the number of errors made by users, word error rates, task completion time, etc.) seems inappropriate for evaluating the efficacy of systems such as STST systems, supporting complex communication, or the impact which specific features of these systems have upon communicational structures. Finally, the analysis of the communication styles may be of great interest to the STST research community, particularly regarding the choice of training materials. Within the scenarios covered by the NESPOLE! system, the least effectively translated turns were meta-communicative turns. This reflects a general avoidance of meta-communication that cannot be mended with data-driven approaches as long as the data do not contain corresponding concepts. Obviously, meta communicative concepts were not sufficiently enough represent by the monolingual conditions exploited for NESPOLE! as well as for other similar projects and, therefore, left unaddressed by the resulting system.

8 Acknowledgements

This work has been partially supported by the National Science Foundation under Grant number 9982227, and by the European Union under Contract number 1999-11562 as part of the joint EU/NSF MLIAM research initiative. Any opinion, suggestion and recommendation expressed are those of the authors and do not necessarily reflect the views of the EU or of the NSF.

The authors wish to acknowledge the contribution and support by the other participants in NESPOLE!, in particular Roldano Cattoni, Nadia Mana, Emanuele Pianta, Elisabetta Fauri, Franca Rossi, Robert Isenberg, Chad Langley, Loredana Taddei and Walter Gerbino.

9 Bibliographical References

Alon Lavie et al. 2001. Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-commerce Application. In *Proceedings of HLT'01*, San Diego, CA.

Lori Levin, et al. 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.

Sharon L. Oviatt. 1997. Multimodal Interactive Maps: Designing for Human Performance. *Human-Computer Interaction*, 12, pp. 93-129.

Sharon L. Oviatt. 1999. Mutual Disambiguation of Recognition Errors in a Multimodal Architecture. In *Proceedings of CHI '99*, ACM Press, New York, pp. 576-583.

Erica Costantini, Fabio Pianesi & Susanne Burger. 2002. The Added Value of Multimodality in the NESPOLE! Speech-to-Speech Translation System: an Experimental Study. In *Proceedings of ICMI'02*, Pittsburgh, PA.

Jean Carletta et al. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, 23 (1) 13-21.

Jean Carletta et al. 1996. HCRC Dialogue Coding Manual", *HCRC Technical Report*, HCRC/TR-82.

Anne H. Anderson et al. 1991. The HCRC Map Task Corpus, *Language and Speech* 34 (4), 351-366.

Loredana Taddei, Erica Costantini & Alon Lavie. 2002. The NESPOLE! Multimodal Interface for Cross-lingual Communication - Experience and Lessons Learned. In *Proceedings of ICMI'02*, Pittsburgh, PA.