

XVI CICLO DEL
DOTTORATO DI RICERCA IN
PSICOLOGIA

Multimodality in Computer Mediated Communication

Dottoranda:
ERICA COSTANTINI



Relatore:
Prof. WALTER GERBINO

Correlatore:
Dott. FABIO PIANESI

Coordinatore:
Prof. CARLO SEMENZA

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XVI CICLO DEL
DOTTORATO DI RICERCA IN
PSICOLOGIA

Multimodality in Computer Mediated Communication

TORANDO:

CA COSTANTINI

76

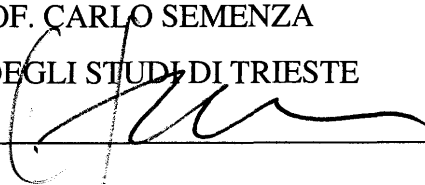
es

COORDINATORE DEL COLLEGIO DEI DOCENTI:

CHIAR.MO PROF. CARLO SEMENZA

UNIVERSITÀ DEGLI STUDI DI TRIESTE

FIRMA: _____



RELATORE:

CHIAR.MO PROF. WALTER GERBINO

UNIVERSITÀ DEGLI STUDI DI TRIESTE

CORRELATORE:

DOTT. FABIO PIANESI

ITC-IRST TRENTO

A Gian Piero

Acknowledgments

First of all, I wish to acknowledge my advisor, prof. Walter Gerbino, who believed in me and in my work.

I am very grateful to Fabio Pianesi and Susanne Burger, for their invaluable contributions during all the phases of this work.

Many thanks to all the other NESPOLE! project colleagues at ITC-Irst, CMU, UKA and Aethra S.p.A. for their direct and indirect contributions with recordings, transcriptions and annotations of dialogues, discussion, technical support and/or document revision - and for having contributed to make every working occasion so enjoyable. In particular I would like to mention Nadia Mana, Francesca Guerzoni, Emanuele Pianta and Roldano Cattoni, to whom I am deeply grateful for their constant and valuable support, and all the rest; working with them has been a great pleasure.

I would also like to give a warm thanks to all the other friends and colleagues in Trieste and in Trento, in particular my office mates, all people from the TCC division and the Gender&Science group at ITC-irst.

I acknowledge, in addition, the participants in the experiments, in particular the agents (including people working at APT) for their kind collaboration and great patience.

Thanks to Francesca for the cover and to Marco for the printing of this thesis, and again Nadia for late night revisions.

I am also deeply grateful to Gian Piero – he already knows all the reasons.

Finally, I acknowledge ITC-Irst for their kind hospitality during my PhD.

Table of Contents

INTRODUCTION	1
CHAPTER 1: MULTIMODALITY IN HUMAN-COMPUTER INTERACTION	3
1. Introduction	3
2. Multimodality in Human-Computer Interaction	4
2.1 Definitions	4
2.2 Integration of Different Modalities	5
2.3 Added Value and Myths of Multimodal Interaction	7
3. Computer-Mediated Communication	10
4. Analyzing Dialogues	12
4.1 Conversational Analysis	13
4.2 Pragmatics of Dialogue	14
4.2.1 <i>Speech Act Theory</i>	14
4.2.2 <i>Developments of Speech Act Theory</i>	15
CHAPTER 2: MULTILINGUALITY AND MACHINE TRANSLATION	17
1. Introduction	17
2. Machine Translation	17
2.1 Definition	17
2.2 History	18
2.3 Speech-to-Speech Translation	19
2.4 Evaluation of Machine Translation	21
3. The NESPOLE! Speech-to-Speech Translation Project	22
3.1 Principles and Design	23
3.1.1 <i>Scenario</i>	23
3.1.2 <i>Interlingua</i>	24
3.1.3 <i>Architecture and HLT Servers</i>	25
3.2 The NESPOLE User Interface: First Version	28
3.2.1 <i>Activation of the System</i>	28

3.2.2	<i>Microsoft NetMeeting</i>	29
3.2.3	<i>The AeWhiteboard</i>	30
3.2.4	<i>The Monitor Window</i>	30
3.3	Interface Improvements and Final Version	32
3.3.1	<i>AeWhiteboard</i>	32
3.3.2	<i>NetMeeting Window</i>	33
3.3.3	<i>NESPOLE! Monitor and Dialogue History Window</i>	33
CHAPTER 3:	FIRST EXPERIMENT	37
1	Introduction	37
2	Method	38
2.1	Experimental Design	38
2.2	Task and Instructions	38
2.3	Participants	40
2.4	Recordings, Transcriptions and Annotations	41
2.4.1	<i>Transcriptions: Conventions and Tools</i>	41
2.4.2	<i>Annotation of Gestures</i>	43
2.4.3	<i>Alignment of Transcription Files</i>	45
2.4.4	<i>Annotation for Turn Successfulness</i>	45
2.4.5	<i>Other Annotations</i>	46
3	Results	47
3.1	Turns, Tokens, Types, Dialogue Length	47
3.2	Disfluencies	50
3.3	Pen-Based Gestures	52
3.4	Dialogue Features	55
3.4.1	<i>Dialogue Fluency</i>	55
3.4.2	<i>Ambiguities</i>	56
3.4.3	<i>Successful Turns and Turn Repetitions</i>	57
3.5	Effectiveness	60
3.5.1	<i>Goal Attainment</i>	60
3.5.2	<i>System Usability Scale</i>	61
3.5.3	<i>User Preferences</i>	61
4	Conclusion	63

CHAPTER 4: SECOND EXPERIMENT	65
1 Introduction	65
2 Method	66
2.1 Experimental Design	66
2.2 Task and Instructions	67
2.3 Participants	68
2.4 Recordings, Transcriptions and Annotations	68
2.4.1 <i>Alignment of Transcription Files</i>	70
2.5 Dialogue Structure Coding Scheme	71
3 Results	74
3.1 Dialogue Length, Turns and Words	74
3.2 Gestures	75
3.3 Dialogue Structure	76
3.4 Turn Repetitions	78
3.5 Additional Results for the STST System	79
3.5.1 <i>Successful, partially successful and non-successful turns</i>	79
3.5.2 <i>Manual editing of the recognized string</i>	79
3.6 Questionnaires	80
4 Conclusions	82
CHAPTER 5: DISCUSSION AND CONCLUSIONS	83
1 Summary of Results	83
2 Discussion	86
2.1 Added Value of Multimodality	86
2.2 Speech-Gestures Integration	87
2.3 Push-to-Talk Procedure	88
2.4 Multilingual versus Monolingual Dialogues	89
3 Conclusion	90
REFERENCES	91

APPENDICES	97
APPENDIX 1: Customer's Instructions (Experiment 1)	99
APPENDIX 2: Customer's task (Experiment 1)	101
APPENDIX 3: Description Cards fir Agents (Experiment 1)	102
APPENDIX 4: Instructions for Agents (Experiment 1)	103
APPENDIX 5: Enrollment Form	107
APPENDIX 6: Questionnaire on Computer Literacy and Web Expertise	109
APPENDIX 7: List of Recorded Files	111
APPENDIX 8: Labelling of Spontaneous Phenomena	113
APPENDIX 9: Output of the TransEdit Annotation Tool	117
APPENDIX 10: Conventions for Annotation of Gestures	119
APPENDIX 11: System Usability Scale	124
APPENDIX 12: Description Cards	126
APPENDIX 13: Customer's Instructions (Experiment 2)	128
APPENDIX 14: Customer's task (Experiment 2)	132
APPENDIX 15: Agent's instructions (Experiment 2)	136
APPENDIX 16: HCRC Coding Scheme	138
APPENDIX 17: Decision Tree for Dialogue Annotation (NESPOLE! Version)	139
APPENDIX 18: Usability Questionnaire on the NESPOLE! System	141

INTRODUCTION

Human communication involves simultaneous use of multiple modalities, i.e. communication channels, to convey information. Meaning in conversation is conveyed by a complex and subtle combination of speech (including prosodic features), facial expressions, gestures. The redundancy and complementarity of multimodal information make human communication robust and flexible, and contribute significantly to its naturalness. A great effort has been spent in the last decades in the development of computer systems able to interact through different kinds of media and modalities, thus improving the effectiveness and naturalness of the interaction between humans and computers, and making it more and more similar to interaction between humans. In addition, faster and cheaper computers, and the great development of networks, lay the foundation for global communication infrastructures capable of supporting always more efficient remote communication, e.g. in the case of Web Phone, Video Call Centers, Videoconference Systems. In this context the main significant barrier to overcome is language, especially in the field of e-commerce and e-services, where the availability of a sufficient set of human operators covering all the relevant languages become too expensive, or even unfeasible (e.g. the case of highly technical services). Traditional approaches of machine translation started to move towards the issues of multimedia and multimodal human to human communication, where the integration of

different media and modalities can help overcoming the traditional limits of the translation provided by human language technologies.

For monolingual systems a large amount of studies and data are available about effects of different media and modalities on the kind of interaction that can be reached, both for human-computer interaction and for computer-mediated communication. There are few or no data describing what happens in multilingual scenarios, where translation is provided by a speech-to-speech translation system and multiple media and modalities are at play. This work aims at contributing to this point, by studying communicative strategies and speech-gestures integration in task oriented human to human conversation mediated by a multimedia and multimodal speech-to-speech translation system.

In chapter 1 we present an overview of the research concerning multimodality, in particular in Human-Computer Interaction and Computer-Mediated Communication, focussing on what is more strictly related to our work. We then explore issues related to analysis of dialogues, and describe into detail the dialogue coding scheme we resorted to for our work.

Chapter 2 is about multilinguality. We start giving the context of machine translation, and we then describe into details the NESPOLE! Speech-to-Speech translation system, that was used in our experiments.

Chapter 3 and 4 present objectives, method and results respectively for our first and second experiments.

In chapter 5 we discuss the results and draw our conclusions.

CHAPTER I

MULTIMODALITY IN HUMAN COMPUTER INTERACTION

1. Introduction

“Human-Computer Interaction (HCI) is a multidisciplinary field in which psychology and other social sciences unite with computer science and related technical fields with the goal of making computing systems that are both useful and usable” (Olson and Olson, 2003). HCI studies how people interact with computing technology. In the earliest days of computers, HCI was not a topic of interest, because very few people interacted with computer, and those who did generally were technical specialists. In the sixties the first papers appeared on the subject. As more and more people started using computers for a broadening number of tasks, HCI became a popular research topic. It has now been, for some years, a major area of research in computer science, human factors, engineering psychology, ergonomics and related disciplines (Nickerson and Landauer, 1997). The psychological and social aspects of HCI are many and diverse. Questions of interest range from aspects pertaining layout of keyboards and design of type fonts, to the configuration of virtual workspaces that have to be shared by geographically dispersed members of a work team. Those questions involve the effects that computer systems can have on their users, on work, on business processes, on furniture and

building design, on interpersonal communication, on society and social processes, or on the quality of life in general. Some research is motivated by an interest in making computer-based systems easier to use and in increasing their effectiveness as tools; some is driven by a desire to understand the role (or roles) this technology is playing in shaping our lives. Main issues explored in HCI include equity, security, function allocation, effects and impact, users' conceptions of the system they use, usefulness and usability, interface design, input devices, intelligent interfaces, augmentation, information finding, use and management, computer-mediated person-to-person communication (Nickerson and Landauer, 1997).

Here we focus on research about multimodal input. In particular we start with the definition of multimodality. We discuss the issue of integration of different modalities in computer systems, focussing on speech and pen-based gestures. We write about myths and evidences in HCI concerning the added value of multimodality. We then introduce multimodality in Computer-Mediated Communication (CMC), including the issue of dialogues analysis.

2. Multimodality in Human-Computer Interaction

2.1 Definitions

In a communication act, medium is a means of conveying a representation (to a human), e.g. a diagram or a text. Modality refers to the sensory or perceptual experience. Multimodality is based on the use of multiple sensory modalities by which humans receive information (tactile, visual, auditory, etc), and it also requests the use of at least two response modalities to present information (e.g. verbal, manual activity). So, for example, in a multimodal interaction a user may receive information by vision and sound and respond by voice and touch. Multimodality could be compared with 'unimodality', which would be based on the use of one modality only to receive or present information (e.g. watching a multimedia presentation and responding by pressing keys).

Nigay and Coutaz (1993) define multimodal systems as follows: "In the general sense, a multimodal system supports communication with the user through different

modalities such as voice, gesture, and typing. Literally, 'multi' refers to 'more than one' and the term 'modal' may cover the notion of 'modality' as well as that of 'mode'.

- Modality refers to the type of communication channel used to convey or acquire information. It also covers the way an idea is expressed or perceived, or the way an action is performed.
- Mode refers to a state that determines the way information is interpreted to extract or convey meaning.

In a communication act, both in the case it is between humans or between a computer system and a user, both the modality and the mode come into play. The modality defines the type of data exchanged whereas the mode determines the context in which the data is interpreted. Thus, if we take a system-centered view, multimodality is the capacity of the system to communicate with a user along different types of communication channels and to extract and convey meaning automatically. We observe that both multimedia and multimodal systems use multiple communication channels. But in addition, a multimodal system is able to automatically model the content of the information at a high level of abstraction. A multimodal system strives for meaning.'

2.2 Integration of Different Modalities

Most common multimodal interfaces combine speech recognition and lipreading, or speech and pen based interfaces (Oviatt and Cohen, 2000), but other combinations are also explored for instance: the integration of speech and gestures; of speech, eye-gaze and hand-gestures; face and gesture. It is not uncommon to find combination of speech and more conventional user interface modalities, such as keyboard and mouse related ones. What we have just described is multimodal input, but of course multimodality can also refer to output. This is usually associated with the use of animated characters (see Cassel, 2000), but can affect graphical user interface elements (display of textual information, or some graphical display). Nigay and Coutaz (1993) classify multimodal systems through a three-dimensional design space. The three dimensions are:

- temporal use of modalities: sequential versus parallel,
- fusion: combined versus independent;

- level of abstraction: meaning versus no meaning (it refers to different level of abstraction in processing, e.g. from just recording the signal to interpreting it as a meaningful parsed sentence).

According to Oviatt and colleagues (Oviatt et al., 2000) two main architectural types are employed in multimodal systems: early fusion system, that integrate signals at the feature level, and late fusion systems, that integrate information at a semantic level. Representative of the early fusion approach are the systems that employ Hidden Markov Models, in which the model is trained on the two modalities (e.g. lipreading and speech) simultaneously. The recognition process in one mode therefore influences the course of recognition in the other. While this architecture has proven effective for modalities that are closely coupled, such as lip movements and speech, they are harder to use when the modalities offer complementary, rather than overlapping information. Late fusion architectures, on the other hand, employ individual recognizers for each modality, whose output is then combined according to time and semantic constraints. The advantage of this latter type of architecture is that the independent recognizers can be acquired and trained individually, leveraging on existing techniques and recognizers. Late fusion also allows in principle the integration of more than one modality, and the addition of modalities to a system in an easier way than is generally possible when early fusion is used. Given the asynchronous nature of the individual recognizers used in late fusion architectures, and the heavy computational requirements associated with analyzing some of the modalities (e.g. speech and vision), the integration of modes and fusion of information constitutes a major problem.

Multimodality aims not only at making several modalities cohabit in an interactive system, but especially at making them cooperate together (Salisbury, 1990). For instance, if the user wants to move an object using a speech recognition system and a touch screen, she has just to say "*put that there*" while pointing at the object and at its new position (Bolt, 1980). In human communication, this task is very easy to achieve since the use of speech and gestures is completely coordinated. Unfortunately, and at the opposite of human communication means, the devices used to interact with computers have not been designed at all to cooperate. For instance, the difference between time responses of devices can be very large (a speech recognition system needs more time to recognize a word than a touch screen driver to compute the point coordinates relative to a pointing gesture). This implies that the system receives an information stream in an order which does not correspond to the real chronological

order of user's actions (like a sentence in which words have been mixed up). Consequently, this can lead to bad interpretations of user statements.

A main technological problem concerns the criteria that should be used to decide the type of fusion of an information with another one, and at what abstraction level this fusion should be done. On the one hand, a fusion at a lexical level allows for designing generic multimodal interface tools, though fusion errors may occur. On the other hand, a fusion at a semantic level is more robust because it exploits many more criteria, but it is in general application-dependent. It is also important to handle possible semantic conflicts between two modes, e.g. speech and gesture, and to exploit information redundancy when it occurs.

2.3 Added Value and Myths of Multimodal Interaction

"Multimodal systems represent a research-level paradigm shift away from conventional WIMP (windows-icons-menus-pointers) interfaces towards providing users with greater expressive power, naturalness, flexibility, and portability. Well designed multimodal systems integrate complementary modalities to yield a highly synergistic blend in which the strengths of each mode are capitalized upon and used to overcome weaknesses in the other. Such systems potentially can function more robustly than unimodal systems that involve a single recognition-based technology such as speech, pen, or vision" (Oviatt, 1999a).

However, in multimodal systems research it is often assumed that human-human communication is 'maximally multimodal and multimedia' (Bunt 1998). The 'added-value' of multimodal systems is often taken for granted. For instance, Bunt (1998) stated that "in natural communication, all the modalities and media that are available in the communicative situation are used by participants". But this is not always the case. Even providing to the users the whole range of modalities involved in human to human communication with a good level of integration and synchronization is not enough to have "natural" systems: "imitation" of human to human communication is not always possible, neither desirable. In fact, there are differences between human-human and human-computer interaction. In human-human interaction, for example, there is available a quite sophisticated system (human's mind), which indicates which

modality to be used and when, while current multimodal systems do not have such ability.

Two hypotheses can be made about combination of different modalities (Raisamo, 1999). The first is that the combination of human output channels effectively increases the bandwidth of the human-computer channel, thus improving the effectiveness of the interaction. Potential benefits of the multimodal interaction are increased efficiency, redundancy, perceptibility, naturalness, accuracy, synergy (Maybury and Wahlster, 1998). Several studies highlighted different benefits in specific scenarios and tasks. The second hypothesis is that adding extra output modality requires more neurocomputational resources and will lead to deteriorated output quality, resulting in reduced effective bandwidth. Two types of effects have been observed (Raisamo, 1999): a slow-down of all output processes, and interference errors due to the fact that selective attention cannot be divided between the increased number of output channels (e.g.: writing when speaking, or speaking when driving a car).

In 1999 Oviatt (1999b) identified 10 myths about multimodal interaction, which at that time were fashionable among computationalists, and discussed them from the perspective of contrary empirical evidence. The myths (and Oviatt's objections) were the following:

1. *If you build a multimodal system, user will interact multimodally:* Users like being able to interact multimodally, but they do not always do so.
2. *Speech and pointing is the dominant multimodal integration pattern:* Modes that transmit written input, manual gesturing and facial expressions are capable of generating symbolic information that is more richly expressive than simply object selection.
3. *Multimodal input involves simultaneous signals:* Beyond the use of deixis, users' spoken and pen-based input frequently do not overlap at all during multimodal commands to a computer.
4. *Speech is the primary input mode in any multimodal system that includes it:* Speech is neither exclusive carrier of important content, nor it has temporal precedence over other input modes.
5. *Multimodal language does not differ linguistically from unimodal language:* Multimodal language is different than unimodal forms of natural language, and in many respects it is substantially simplified.

6. *Multimodal integration involves redundancy of content between modes*: actual data highlights the importance of complementarity as a major organizational theme during multimodal communication.
7. *Individual error-prone recognition technologies combine multimodally to produce even greater unreliability*: In a well designed and optimized multimodal architecture, there can be mutual disambiguation of two input signals.
8. *All users' multimodal commands are integrated in a uniform way*: multimodal that systems can detect and adapt to a user's dominant integration pattern could lead to considerably improved recognition rates.
9. *Different input modes are capable of transmitting comparable content*: Different modes basically vary in the degree to which they are capable of transmitting similar information, with some modes relatively more comparable (speech and writing) and others less so (speech and gaze).
10. *Enhanced efficiency is the main advantage of multimodal systems*: there are other advantages of multimodal systems that are more noteworthy in importance than modest speed enhancement.

In separating myth from reality the nature of multimodality interaction has been made clearer, and some insights are given for guiding the design of multimodal systems. More research, in particular from cognitive science, is needed to understand the following (Raisamo, 1999):

- When is a multimodal system preferred to a unimodal system?
- Which modalities make up the best combination for a given interaction task?
- Which interaction devices are to be assigned to these modalities in a given computing system?
- How should these interaction devices be used, that is, which interaction techniques are to be selected or developed for a given task?
- How does the brain work and which modalities can best be used to gain the synergy advantages that are possible with multimodal interaction?

In this section we have considered some of the main issues concerning multimodality within the HCI framework. In this framework, the computer could be seen either as a tool or as a dialogue partner. In the first case the user is always responsible for initiating the operations and the machine is a passive tool that tries to understand the user through all different input modalities that the system recognizes. Multiple input

modalities are here used to enhance direct manipulation behaviour of the system. When the computer is seen as a dialogue partner, the user can have conversations with the computer (this is the case of agent-based conversational user interfaces). Here multiple modalities are used to increase the anthropomorphism in the user interface, for instance in talking heads. Another way of using computers is to support human to human communication. Even in this case, integration of different multimodalities plays a crucial role, as we describe in the following paragraph.

3. Computer-Mediated Communication

Computer-Mediated Communication (CMC) is the process by which people create, exchange, and perceive information using networked telecommunications systems (or non-networked computers) that facilitate encoding, transmitting, and decoding messages. Studies of CMC can view this process from a variety of interdisciplinary theoretical perspectives by focusing on some combination of people, technology, processes, or effects. Some of these perspectives include the social, cognitive/psychological, linguistic, cultural, technical, or political aspects; and/or draw on fields such as human communication, rhetoric and composition, media studies, human-computer interaction, journalism, telecommunications, computer science, technical communication, or information studies.

Most of research in CMC have been focussing on textual messages, synchronous (e.g. chat) or asynchronous (e.g. e-mail) (see Herring, 1996 and Lea 1992). Most of research concerning synchronous online speech-based communication, i.e. people talking to each other during remote connection mediated by computers, considers video conferencing scenarios, and investigates the impact of different features of video conferencing applications on communication. Here the multimodal aspect is given mainly by the presence of video, that makes available facial expressions and gestures of the speakers. For instance, several features of video have been considered: visual cues, audio-video synchronization, colour versus greyscale video, compression and video frame rate, image size and camera angles (for a review see Kies, J. K. and Williges, R. C.,1997). Video conferencing applications have been evaluated in comparison to telephone or face-to-face communication using measures as task performance, dialogue length, speech patterns, number of interruptions, back

channelling, dialogue structure (e.g. Anderson et al, 1996). For the evaluation of video conferencing applications, Monk et al. (1996) suggest distinguishing between measures that characterize the process of communication from measures of outcome. Whereas outcome measures, common in the human factors tradition, are solely concerned with how successfully the work was done, process measures, close to conversational analysis approaches, are concerned with the nature of the communication that took place i.e., the ways in which the work was done. Outcome measures are limited when it comes to evaluating technology. They are often insensitive and even when they do show effects they provide no real understanding of why those effects have occurred. Measures of process can help overcoming these limitations. These measures include global measures of dialogue efficacy such as: common ground and subjective effort; surface features of conversational content such as the use of personal pronouns and measures of conversational structure such as topic mention, overlapping speech and gaze.).

Different ways of analyzing CMC dialogues refer mainly to one of two classical approaches to dialogue: speech-act theory (from linguistics) and conversation analysis (from ethnomethodology). For instance Doherty-Sneddon et al., (1997) apply a dialogue structure scheme from the speech-act theory tradition to compare the structure of dialogues in face-to-face and video-mediated communication; Ruhleder and Jordan (2001) use Interaction Analysis (derived from Conversation Analysis and Ethnomethodology) to analyze one particular limitation of video-based teleconferencing (the impact of audio and video delay on distributed communication); Herring (2003) presents Computer-Mediated Discourse Analysis (CMDA), which adapts methods from both linguistic and ethnomethodology to analyze CMC dialogues. The two traditions are described in the next paragraph.

4. Analyzing Dialogues

Everyday humans interact, whether orally face-to-face, by telephone, through video-conferencing or through the written medium. Depending on the approach, communication has been investigated as social activity, as cognitive activity, as construction process for sharing of meanings and experiences. It has been studied as transmission of information through a channel (information theory, e.g. Shannon and Weaver, 1949); as interaction between text and context (Morris, 1938); as linguistic act (Austin, 1962); as intentional process building shared meanings (Sperber and Wilson, Grice), as “ways of being in communication” (Bateson, 1972); as mean for construction of identity or as result of relationship games.

Dialogue is the means through which a substantial amount of this communication is achieved. But conversations do not always consist of well-formed sentences and even when they do it is not obvious that it is the property of consisting of sentences that is important for the purpose of carrying out the conversation. Rather, successful conversation takes place despite the fact that speakers' utterances consists of disfluencies (false starts, interruptions, reformulations, laughter, etc.) and overlaps. A successful conversation is one where the rules of dialogue are followed and where the aim of the conversation is achieved, i.e. if the aim of the conversation is to exchange information, then the necessary information is exchanged; if the aim is to establish social relations, they are successfully established. Disfluencies and overlap do not necessary imply that a conversation will be unsuccessful, although in general it has been shown that speakers speak one at a time.

When computer systems are built to mediate, support or simulate human-to-human interaction, it becomes crucial to understand what it means for a dialogue to “work”, to be successful. Different methods, referring to classical approaches to dialogue, could help about this point. In particular *Conversation Analysis* (Sacks, 1992) was developed as an approach to dialogue analysis aiming at finding out if there are any regularities in conversation and if so, to attempt to formulate them. Conversation Analysis explores how participants collaborate in constructing the conversation, taking into account disfluencies, without applying a priori interpretation and/or annotation schemes defined. Speech Act Theory (Austin, 1975 and Searle, 1969) and related approaches were developed to give account for the functional meaning of an utterance, as well as for coherent sequences of verbal interaction. The latter approaches are

those commonly used in development of dialogue systems, and go usually under the broader name of Pragmatics of dialogue.

4.1 Conversational Analysis

Conversation Analysis (CA) is essentially a naturalistic, observation-based science of actual (verbal and non-verbal) behaviour, which uses audio and video recordings of naturally occurring interactions as the basic form of data.

Underlying the methodology of ca is the attempt to capture and document the back and forth, or processual, character of interaction. The analytic aim is to show how conversational and other interactions are managed and constructed in real time, through the processes of production and comprehension employed by participants in coordinating their activities when talking with each other. CA's methodology is naturalistic and largely qualitative, and is characterised by 4 key features:

1. CA's research is based on the study of naturally occurring data (transcriptions of audio visual data); data are not gathered through simulations, experimental or quasi-experimental tasks, and are not made-up.
2. phenomena in the data are generally not coded (coding tokens on the basis of certain manifest similarities runs the risk of collecting, in the same category, object which have in reality a quite different interactional significance).
3. CA's methodology is generally not quantitative (quantifying the occurrence of a certain object is likely to result in the truly interactional properties of that object being overlooked).
4. CA attempts to document and explicate how participants arrived at understanding each other's action during the back-and-forth interaction between them and how in turn they constructed their turn so as to be suitably responsive to prior turns. Therefore, CA focuses especially on those features of talk which are salient for participants' analyses of one another's turn at talk, in the progressive unfolding of interactions.

CA has developed a transcription system which aims to capture faithfully features of speech which are salient to the interaction between participants, including aspects of the relationship between turns at talk, as well as characteristics of speech delivery

(such as emphasis, loudness/softness, pitch changes, sound stretching and curtailment, etc.).

1.2 Pragmatics Of Dialogue

1.2.1 Speech Act Theory

Austin's 1955 Harvard lectures, first published in 1962, is the traditional starting point of speech act theory. Austin developed speech act theory as a reaction to traditional attitudes to language. It was commonly believed that the basic sentence type in language is declarative, that the major use of language is to describe states of affairs, and that the meaning of utterances can be described in terms of their truth or falsity. Austin observed that a lot of utterances in conversation are not statements (e.g. "excuse me") and that not all the utterances can be said to be true or false. Often even sentences with the grammatical form of declaratives are not used to make statements about states of affairs, e.g. "I name this ship Titanic". The alternative is the idea that sentences perform an action: speaking is rather viewed as a kind of action being performed by the speaker. The actions performed by sentences are called *acts*, hence the terms *speech acts*, for the unit of speech. Acts form the basis of Speech Act Theory.

A speech act is a complex unit. Austin offered an analysis of the concept of speech acts, which distinguishes between three aspects of a speech act:

1. locutionary act: it includes the phonetic act (producing noises), the phatic act (conforming the phonetic noises to a certain vocabulary and grammar), and the rhetic act (the use of phatic act with a special sense of reference) (Austin, 1975). The locutionary aspect is about saying something that makes sense in a certain language, and can thus be seen as connected to traditional semantics of language.
2. illocutionary act: it relates to the kind of action performed in saying something, i.e. asking or answering a question, giving information, etc. The illocutionary act is viewed as composed by illocutionary force, specifying the type of action (question, answer, etc.), together with the propositional content which specifies more closely

the action. The latter aspect can be said to mirror the speakers intentions behind a given utterance.

3. perlocutionary act: it is connected to the effects of the utterance, i.e. what effect a certain utterance provokes in a certain context. Examples are persuasion and surprise.

Austin made a classification scheme for speech acts, primarily based on illocutionary force. However, Austin (1975) does not seem to be completely happy over the classification. The theory was further developed by different authors, in different directions.

1.2.2 Developments Of Speech Act Theory

Searle (1969) developed the speech act theory, without focussing on the three different aspects of speech acts (he used only the concept of illocutionary act, but not those of locutionary and perlocutionary acts). He gave a more fine-grained and systematic description on speech acts. According to Traum (1999), Searle's most important contribution was the attempt to provide necessary and sufficient conditions for the performance of different types of illocutionary acts. In addition, he further developed Austin's taxonomy for the speech acts, basing the division in addition on the purposes/intentions behind the acts.

Both Austin and Searle were concerned with the description of the function of utterances in context, although, they did not address the issue of how these functional units and their meanings are related to each other in a longer sequence of speech acts. However, this issue came in focus when attempts were made to make use of speech acts in computational systems, e.g. in the field of Artificial Intelligence (AI) with the plan-based approaches to dialogue. For example Bruce (1975) worked on the connection of definitions of speech acts to more formalised and computationally useful criteria, using work on plans and actions in giving account of speech acts. Cohen and Perrault (1979) and Allen (1983) worked on language generation introducing a plan-based theory of speech acts. An attempt to connect explicit linguistic features to speech acts was done by e.g. Hinkelman and Traum (1989), who use linguistic cues to develop partial speech act templates. However, even though the definitions of speech

acts become more operational and implementable during the integration of speech act theory with theories of plan and intention recognition within AI, there is still a general lack of attempts giving account of complex aspects of verbal interaction, e.g. in dialogue.

Traum (1999) some attempts towards a better definition of speech acts in dialogue were done by Litman and Allen (1992), who extended Allen and Perrault's (1980) work to include connected dialogues rather than just single pairs of utterances. They also organised the dialogue in a hierarchical structure, based on the plans, which could be nested. In addition, Cohen and Levesque (1991) extended their work on the logic of speech acts to a theory on joint intention and multi-agent action. This work stressed the interactive, social aspect of communication.

The analyses of dialogue come to contain several levels or strata, similarly to the conversation analysis done by Sinclair and Coulthardt (1975) in their analysis of class room conversations. In this tradition the dialogue exchange is described as consisting of dialogue moves, dialogue games and dialogue transactions. The move is the smaller unit while the transaction is the largest unit. Agents are generally said to plan the dialogue at the level of game, but to execute them at the level of moves. Since the traditional speech acts were insufficient for describing or controlling the dialogue flow, new levels were introduced, including levels for turn-taking, repair, reference/information and attention, which could be viewed a discourse management (Traum, 1999). Traum and Hinkelman (1992) suggested a dialogue coding scheme comprising these new speech acts.

Other schemas were developed taking into account different levels of dialogue analysis. Dialogue coding has become a fundamental feature in all kinds of dialogue systems, question-answer systems, (e.g. LINDA, Ahrenberg et al., 1995), as well as translation systems (e.g. VERBMOBIL, Alexandersson et al., 1997). It has also been used in research on prosody in spontaneous speech (Stolke et al., 2000) as well as in attempts to improve speech recognition speech (Stolke et al., 2000).

CHAPTER II

MULTILINGUALITY AND MACHINE TRANSLATION

1. Introduction

Multilinguality is characteristic of tasks that involve the use of more than one natural language. In the modern world, there are always more tasks and situations that imply multilinguality. Hence, there is an increasing demand for translation services, and consequently interest in alternative ways of producing them. The principal alternatives that have been proposed include partially or fully automatic translation and machine aids for translators.

In the first part of this chapter we overview the history and challenges of machine translation, focussing on speech-to-speech translation (STST).

In the second part we describe the NESPOLE! STST system, used to perform the experiments described in the next chapters.

2. Machine Translation Systems

2.1 Definition

The term machine translation (MT) is normally taken in its restricted and precise meaning of fully automatic translation. In Hovy et al. (2001) this term is extended from fully automatic translation to “any computer-based process that transforms (or helps a

user to transform) written text from one human language into another. They distinguish between:

- *Fully Automated Machine Translation (FAMT)*: MT performed without the intervention of a human being during the processes;
- *Human-Assisted Machine Translation (HAMT)*: the computer system does most of the translation, appealing in case of difficulty to a (mono- or bilingual) human for help;
- *Machine-Aided Translation (MAT)*: human does most of the work but uses one of more computer systems, mainly as resources such as dictionaries and spelling checkers, as assistants.

Traditionally, MT has been used either to gather material written by others in a variety of languages and convert them all into his or her own language (assimilation), or to broadcast material, written in one language, in a variety of language to the world (dissemination). A third class of MT has also recently become evident, related to communication. This is the case of two or more individuals in more or less immediate interaction, typically via email or otherwise online, with an MT system mediating between them. Each class of translation has very different features, is best supported by different underlying technology, and is to be evaluated according to somewhat different criteria.

2.2 History

Machine Translation was the first computer-based application related to natural language, starting after World War II, when Warren Weaver suggested using ideas from cryptography and information theory. The first large-scale project was funded by the US Government to translate Russian Air Force manuals into English. After a decade of initial optimism, funding for MT research became harder to obtain in the US. However, MT research continued to flourish in Europe and then, during the 1970s, in Japan. Today, over 50 companies worldwide produce and sell translations by computer, whether as translation services to outsiders, as in-house translation bureaux, or as providers of online multilingual chat rooms. Ten years ago, the typical users of machine translation were large organizations such as the European Commission, the US Government, the Pan American Health Organization, Xerox, Fujitsu, etc. Fewer small companies or freelance translators used MT, although

translation tools such as online dictionaries were becoming more popular. However, ongoing commercial successes in Europe, Asia, and North America continued to illustrate that, despite imperfect levels of achievement, the levels of quality being produced by FAMT and HAMT systems were capable to address some users' real needs. Systems were being produced and sold by a small number of companies, and both the European Commission and the US government started investing in large MT projects in the 1980s and 1990s.

Thanks to ongoing commercial growth and the influence of new research, the situation is different today from ten years ago. There has been a trend toward embedding MT as part of linguistic services, which may be as diverse as email across nations, foreign-language web searches, traditional document translation, and portable speech translators with very limited lexicons (for travellers, soldiers, etc.). The use of tools for translation by freelancers and smaller organizations is developing quickly. Cheap translation assistants, often little more than bilingual lexicons with rudimentary morphological analysis and some text processing capability, are starting to help small companies and individuals write foreign letters, email, and business reports. MT services are offered via the Internet, often free for shorter texts, and it is increasingly being bundled with other web services (see the website of Altavista, which is linked to Systran).

Different approaches have been used for MT so far: Statistical versus Linguistic MT, Feature Symbolic Statistical MT, Rule-based vs. Example-based MT, Transfer vs. Interlingual MT, Multi-Engine MT (Hovy et al., 2001). Here we do not describe the different approaches. We focus on peculiarities of Speech-to-Speech translation (STSTS) from the point of view of communication, omitting technical details.

2.3 Speech-to-Speech Translation

Spoken Language Translation (SLT) is the ability of a machine to interpret a multilingual human-human spoken dialog. Speech-to-Speech Machine Translation is a multidisciplinary research area that addresses one of the most complex problems in speech and language processing.

Early speech translation systems implemented in the Eighties mainly had the purpose to demonstrate the feasibility of speech translation. Their main features included very restricted domains, severe limitations on fixed speaking style,

grammatical coverage, and limited size vocabulary. It has become increasingly clear that improving each component (e.g. speech recognition) was not enough: good speech translation cannot be achieved by mere combination of better speech recognition and machine translation. Over the last decade STST has benefited from advances in speech and language processing as well as from the availability of large multilingual databases. Speech recognition systems have been improved to handle the sloppy speech people produce when talking spontaneously to each other. The spontaneous phenomena of speech (e.g. interruptions, hesitations, noises) are automatically recognized, filtered and properly prepared for translation. Speech translation technology has matured to the point of allowing free, spontaneous dialogues using large vocabularies that can be translated into a variety of languages. However, this is possible only for very restricted domains: unrestricted simultaneous translation will remain impossible for the foreseeable future (Lazzari et al., 2001). In addition, the issues of dialogue efficiency still need to be addressed.

At the time there are many approaches to spoken language translation. They can roughly be divided in two classes: direct approaches that try to link speech recognition and machine translation techniques, and interlingual approaches that try to decode both recognition and understanding into a common consistent framework (Lazzari et al., 2001). We skip details here concerning the differences between the approaches. The NESPOLE! system that was used for the experiments described in this work, is interlingua-based. The NESPOLE! project directly benefited from the experience of the Verbmobil¹ and C-STAR², that follow the same interlingua approaches. Several institutions involved in C-STAR therefore stress an interlingual representation and the development of generation component from the given interlingual representation (CMU, UKA, ETRI, IRST, and CLIPS) (Angelini et al., 1997). Present activity has shifted toward a greater emphasis on interpretation of spoken language, i.e., the system's ability to extract the intent of a speaker's utterance (Bub and Schwinn, 1996). Discourse and domain knowledge and prosodic information are being explored, for more robust interpretation of ambiguous utterances.

One of the most interesting and challenging features of the speech translation system is that it does not need to give a complete correct translation, but just an expression in the target language conveying the relevant meaning of the original sentence. Some contextual cues could be used to disambiguate poor translations, so

¹ VERBMOBIL Project web site: [VERBMOBIL: http://www.dfki.de/verbmobil/](http://www.dfki.de/verbmobil/)

² C-STAR Project web site: [C-STAR: http://www.is.cs.cmu.edu/cstar/](http://www.is.cs.cmu.edu/cstar/)

that the conversation could be effective even in case of fairly bad translation (Lazzari, 2000). In such situation, it is supposed that not only conversational context can play this role, but also multimedia and even more multimodal features (included elements of non-verbal communication), could be effectively used to enrich communication and enhance dialogue effectiveness. However, there are few or not data available concerning how multimodal features could be integrated in a speech-to-speech translation system, and which could be the added value of multimodality in such scenarios.

2.4 Evaluation of Machine Translation

MT evaluations typically include features not present in evaluations of other NLP systems: the quality of the raw (unedited) translations, e.g., intelligibility, accuracy, fidelity, appropriateness of style/register; the usability of facilities for creating and updating dictionaries, for post-editing texts, for controlling input language, for customisation of documents, etc.; the extendibility to new language pairs and/or new subject domains; and cost-benefit comparisons with human translation performance. Adequacy evaluations by potential purchasers usually include the testing of systems with sets of typical documents. But these are necessarily restricted to specific domains, and for diagnostic and performance evaluation there is a need for more generally applicable and objective test suites (which have been under development since late 1980s).

Despite some methods and benchmarks for the evaluation of MT systems have been defined and spread, there is still much discussion about which are the most reliable methods and measures. As in other areas of NLP, three types of evaluation are recognised: adequacy evaluation to determine the fitness of MT systems within a specified operational context; diagnostic evaluation to identify limitations, errors and deficiencies, which may be corrected or improved (by the research team or by the developers); and performance evaluation to assess stages of system development or different technical implementations. Adequacy evaluation is typically performed by potential users and/or purchasers of systems (individuals, companies, or agencies); diagnostic evaluation is the concern mainly of researchers and developers; and performance evaluation may be undertaken by either researchers/developers or by

potential users. In the case of production systems there are also assessments of marketability undertaken by or for MT system vendors.

The initial intentions of most evaluation experiences to measure the productivity of systems for potential users was abandoned because it introduced too many variables. Evaluation has concentrated on the performance of the core MT engines of systems, in comparison with human translations, using measures of adequacy (how well a text fragment conveys the information of the source), fluency (whether the output reads like good English, irrespective of accuracy), and comprehension or informativeness. However, user studies started to appear more frequently in the MT evaluation field. For instance, in the last Machine Translation summit (New Orleans, 23-27 September 2003) a special session of the conference has been dedicated to user studies.

3. The NESPOLE! Speech-to-Speech Translation Project

NESPOLE! (NEgotiating through SPoken Language in E-commerce) is the name of a Speech-to-Speech Translation (STST) project. It was designed to provide a fully functional Speech-to-Speech machine Translation system working in real-world settings of common users involved in e-commerce applications. The project addressed four languages: Italian, German, English and French. Four research groups have been involved: ITC-IRST in Trento, Italy, ISL at Universität Karlsruhe (TH), Germany; CLIPS at Université Joseph Fourier in Grenoble, France, and ISL, at Carnegie Mellon University in Pittsburgh, PA, US. In addition, two industrial partners took part in the project: APT Trentino (the Trentino provincial tourism board), Trento, Italy; and AETHRA S.p.A. (a telecommunications company), Ancona, Italy. The project started in January 2000 and ended in December 2002. It was funded jointly by the European Commission and the USNSF (National Science Foundation).

The scenario involves an Italian-speaking agent located in an APT, and an English-, German- or French-speaking customer at an arbitrary location. The two communicate through the Internet using thin terminals (PCs with sound and video cards and H323 video-conferencing software), and can share web pages and maps by means of a special White Board. The NESPOLE! system provides for multimodal communication, allowing users to perform gestures on displayed maps, by means of a tablet and a pen.

3.1 Principles and Design

The system uses a client-server architecture to allow a common user, who is initially browsing through the web pages of a service provider on the Internet, to connect seamlessly to a human agent of the service provider who speaks another language, and provides speech-to-speech translation service between the two parties. Standard commercially available PC video-conferencing technology such as Microsoft's NetMeeting is used to connect between the two parties in real-time. The design principles of the NESPOLE! system are described into details in (Lavie et al., 2001).

3.1.1 Scenario

During the project, the NESPOLE! system has been developed in two steps corresponding to two fully functional showcases. After one year and a half, the first showcase (showcase 1) in the tourism domain was completed. For the second showcase the developments were addressed in two directions: enlarging the tourism domain (showcase 2a) and demonstrating system portability to new domains (showcase 2b).

In showcases 1 and 2a, used respectively for the first and the second experiments described here, the scenario is the following: a client user is browsing through the web-pages of APT³ in search of tour-packages in the Trentino region. If more detailed information is desired, the client can click on a dedicated button within the web-page in order to establish a video-conferencing connection to a human agent located at APT. The client is then presented with an interface consisting primarily of a standard video-conferencing application window and a shared whiteboard application. The interface allows the client to carry on a conversation with the agent, where the NESPOLE! server provides two-way speech-to-speech translation between the parties. The agent speaks Italian, while the client can speak English, French or German.

The third showcase (2b) works in the medical domain. It was developed to evaluate the portability of the NESPOLE! STST system to new domains. Within the selected medical domain, the scenario was restricted to a first aid medical assistance service (Mana et al., 2003).

³ the tourism bureau of the province of Trentino in Italy

3.1.2 Interlingua

Translation in NESPOLE! follows an Interlingua approach (see chapter 1). In this chapter we briefly describe the main features of the adopted Interlingua: the Interchange Format (IF). More information on the topic can be found in (Levin et al., 2003, Cattoni et al., 2001). IF is a task-oriented, language independent, meaning representation formalism, aiming at representing the communication intentions of the speaker more than the literal expression of such intentions. An IF representation corresponds roughly to a clause (or fragment of it) called a *Semantic Dialogue Unit* (SDU). The representation consists of four components:

1. the *speaker tag*, where c: indicates the client (in our dialogues the traveler or the patient), and a: the agent (in our dialogues the travel agent, or the doctor);
2. the *speech act*, e.g. thank, give-information;
3. a possibly empty sequence of *concepts*, describing the conceptual focus the utterance, e.g. +hotel, +pain;
4. a possibly empty list of *arguments* as name-value pairs, specifying details of the intended SDU meaning. Arguments are licensed by concepts.

The following are three examples of utterances tagged with their corresponding IF labels:

1. *Thank you very much*
c:thank
2. *And we'll see you on February twelfth*
a:closing (time=(february, md=12))
3. *There is an hotel in the town*
a:give-information+existence+accommodation
(accommodation-spec=hotel, location=town)

The first element is the speaker tag c:, identifying the travel agent. The second component is the give-information speech act, which describe the communication intention of passing some information to the hearer. The speech act is followed by the concepts +existence and +accommodation, which are the two main concepts of the SDU. The combination of a speech act with one or more concepts results in what is

called a *domain action*. In our example the domain action can be paraphrased as “communicating information about the existence of some accommodation”. The domain action licenses a set of *arguments* that are semantically related to the concepts of the domain action. Here the concept +accommodation licenses the accommodation-spec= argument, specifying the type of accommodation the speaker is referring to, whereas +existence licenses the location= argument, specifying that the hotel can be found in the town. Despite being task-oriented, the IF has been conceived with the goal of accommodating as many domains as possible, by clearly distinguishing the IF parts (speech acts, concepts, etc.) that are domain-independent, from those that are domain-specific. This has positively contributed to the portability of STST systems, resulting in the current version of the IF, which covers two very different domains: tourism and medical assistance⁴.

3.1.3 Architecture and HLT Servers

The basic system design is shown in Figure 2.1.

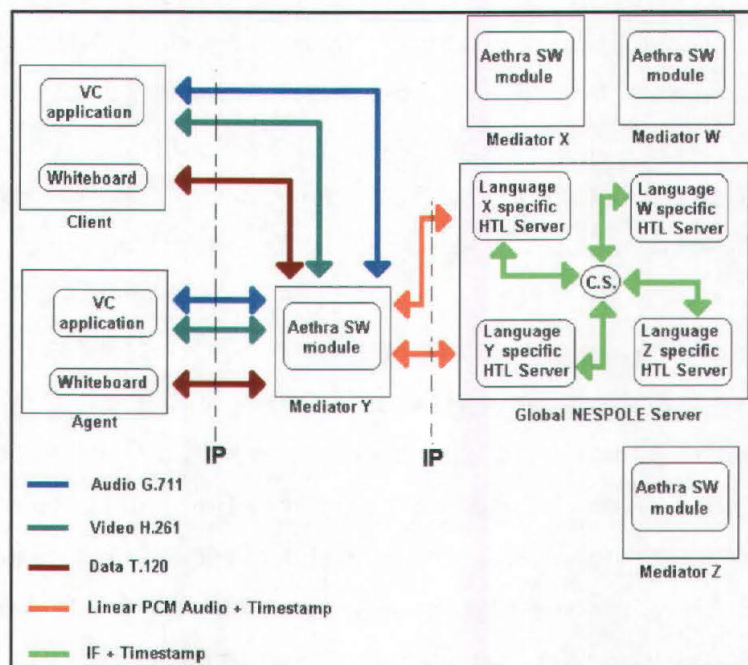


Figure 2.1: The Nespole Architecture.

A key component in the NESPOLE! system is the *Mediator* module, which is responsible for mediating the communication channel between the two parties as well

⁴ For reference: <http://www.is.cs.cmu.edu/nespole/db/>

as interfacing with the appropriate Human Language Technology (HLT) speech-translation servers. The HLT servers provide the actual speech recognition and translation capabilities. This system design allows for a very flexible and distributed architecture: Mediators and HLT-servers can be run in various physical locations, so that the optimal configuration, given the locations of the client and the agent and anticipated network traffic, can be taken into account at any time. A well-defined API allows the HLT servers to communicate with each other and with the Mediator, while the HLT modules within The servers for the different languages are implemented using very different software packages.

For example, let us suppose an English-speaking customer in the US is connecting to an APT agent in Italy. A connection request from the customer's PC (in the US) would be made to the Mediator, which can be physically located anywhere on the net (in practice, located at the agent site in Italy). The Mediator establishes a connection over the internet with both an English HLT server and the Italian HLT server (also physically located anywhere on the internet), before calling the agent in Trento. The Italian HLT server provides Italian speech recognition (*recognizer*), translation from Italian text into our IF (*understanding module*) as well as Italian generation (*natural language generator*) from IF and speech synthesis (*synthesizer*); the English HLT server provides similar functionalities to and from English. The steps between the utterance of a sentence at one site and the reception of the translated sentence from the other site are illustrated in Figure 2.2. When the Italian agent speaks, the Italian recognizer converts the speech signal into text, from which the IF is produced. The IF is sent through the network to the other language HLT servers which produce the output sentence in the target language(s) (English in this case); when the customer answers to the translation of the agent's speech, the same process as already described for the Italian HLT module is activated. The IF of the customer's contribution is sent back to the Italian HLT server where the Italian generation is provided and synthesized. Trace of some of those steps is made available to the users through the feedback window within the user interface (see next paragraphs).

Each user is able to hear both the original audio from the remote user as well as the translation of this audio as provided by the system. The two audio streams are mixed and can overlap. This functionality, provided and managed by Mediator modules, simulates the "simultaneous" translation capabilities that would be provided by a human interpreter. In our case, where network traffic and translation processes

introduce time delays, the ability to hear the original audio provides the users with appropriate feed-back on what is taking place on the other side (the partner is waiting or the partner is speaking). However, in particular situations where there is a need for more control over the transmitted messages, it would be better to disable the original audio. For example, during our experiments (see chapters 3 and 4) we needed to disable the original audio in order to ensure that verbal information was being communicated *only* via the translation (and not via the original language). The interface controlling the Mediator supports the disabling of original audio transmission and controls both original and translated audio volume independently.

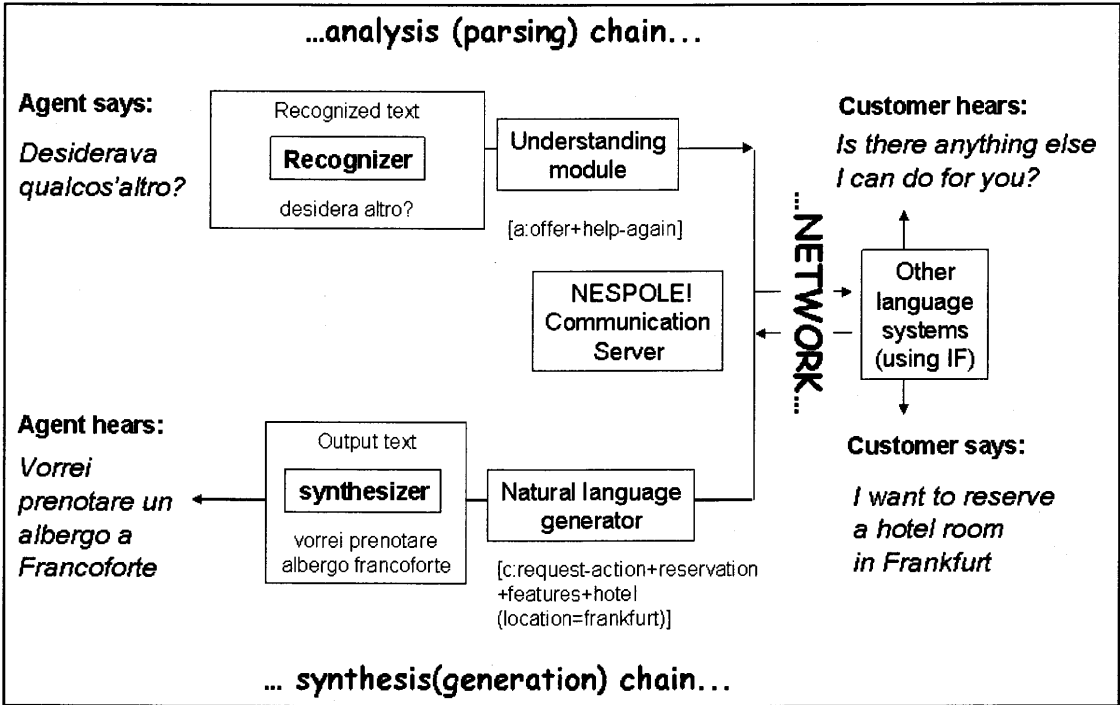


Figure 2.2: The HLT Servers Architecture.

The computationally intensive part of speech recognition and translation is done on dedicated server machines, whose nature and location do not concern to the user. A wide range of client-machines, even portable devices or public information kiosks, are therefore able to run the client software, so that the service can be made available nearly everywhere. The main technical difficulty for VoIP ("Voice over Internet Protocol") applications is coping with adverse internet bandwidth conditions. In order to guarantee real-time communication under insufficient bandwidth conditions, video-conferencing software often drops short segments of speech that were delayed in

transport. This, however, can be very detrimental to the performances of speech recognizers (Metze, 2001). To reduce bandwidth requirements, it is possible to use the NESPOLE! system without video transmission (Taddei et al, 2002).

3.2 The NESPOLE! User Interface: First Version

NESPOLE!'s standard user interface displays three windows, with a fixed size and position on the monitor: 1) The NetMeeting® window, allowing control over the usual features of this application; 2) The Aethra® White Board window, used to display maps and to share pen-based gestures using the White Board drawing functionalities; 3) The Feedback window, displaying feedback to the users concerning the status of the translation process. Each window plays a different role in the interaction between client and agent (see below). In addition, a browser window can be opened when needed. The content and functionalities of the windows has been changed during the project with the aim of providing more usable functions and feedbacks.

In the following paragraphs we describe the interface of showcase 1 and the final version of the interface (showcase 2), together with some of steps of its development.

3.2.1 Activation of the System

The user can start a videoconference with the operator by simply pressing a dedicated hyperlink on the web page. The following sequence of events then take place:

- activation of Microsoft® Netmeeting®;
- establishment of the audio-video-data call to the system of the Tourism Board operator in Trentino;
- transmission of the user's web page address to the agent; this allows the browser on the agent's PC to display the exact same page as the one seen by the user; moreover, if the web page is available in different languages, the agent PC will display the Italian page that corresponds to the French, English or German page of the user;
- activation of AeWhiteboard for graphic information exchange;
- activation of NESPOLE! Monitor to keep track of the translation process provided by the Global NESPOLE! translation server.

These functions have been implemented by using NetMeeting® UI ActiveX Control from Microsoft® and ApplLaunch.ocx developed by Aethra.

3.2.2 Microsoft® NetMeeting®

The NetMeeting (see fig. 2.3, upper right side) window allows control over the usual features of this application. In particular, it establishes the audio-video-data call, it has a button to activate/de-activate the microphone (push-to-talk button) and displays the transmitted video. NetMeeting® delivers additional functions to make the exchange of information and communication easier: audio volume control on the user side and the possibility of muting the local audio. These functions are especially useful in the case of very noisy environments. Moreover, the data channel opened by NetMeeting is in compliance with the T.120 standard, which allows for file transfer and application sharing.

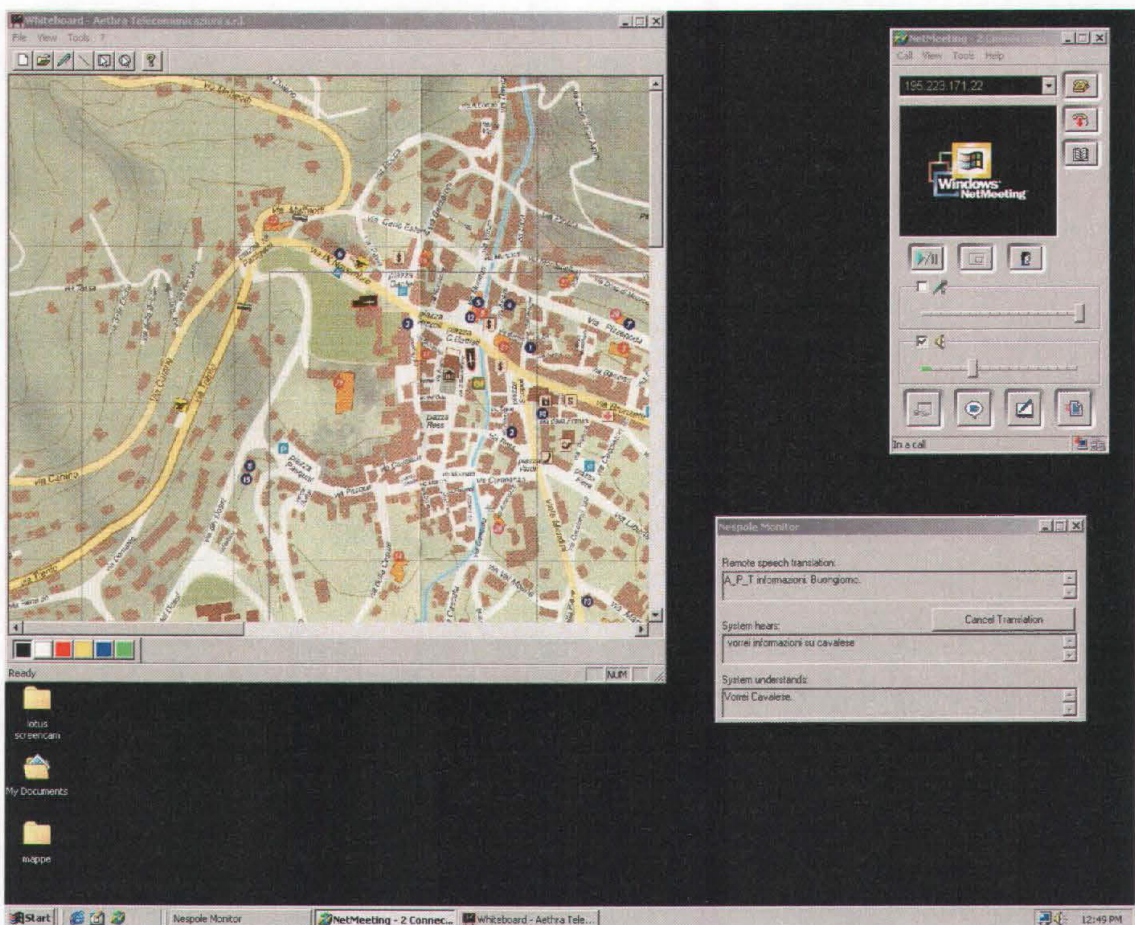


Figure 2.3: The Interface of the NESPOLE! Showcase 1.

3.2.3 The AeWhiteboard

The AeWhiteboard (Figure 2.3, left side) is based on Windows application standards and features menus, a tool bar and a status bar. It allows the user to view bitmaps, such as town maps or maps of tourist areas. The user can also draw gestures on the bitmap to show routes or highlight places, zoom in and out and scroll the bitmap. All the operations performed by one user are shared with the remote user.

The AeWhiteboard drawing functionalities include:

- free hand strokes (only MM condition). By selecting this function, the user can draw arrows, circles and other free hand strokes of her choice on the displayed image, by using a pointing device (mouse, pen+tablet). A palette allows selection among different colors, which helps distinguishing among different gestures performed on the same image.
- lines: the user can select a specific function to draw lines;
- selection of areas on maps (only MM condition). This can be done by enclosing portions of maps in elliptical/rectangular figures drawn with the pointing device. As in the previous case, appropriate colors can be selected among the palette.

The drawings are performed by means of a tablet-pen device. Appropriate colors can be selected among the palette for all types of drawings, to distinguish among different gestures. Finally, the user can save a copy of the map with the drawings performed on it, in order to reload it whenever needed. Another important functionality supported by the AeWhiteboard is the ability to simultaneously display a web page on the browsers of both parties. If the same page is available in multiple languages, the system will display the web page in the appropriate language of each of the two users.

All of the above tools and modalities are available to both parties throughout the communication, interleaved with the ongoing multilingual verbal dialogue that is taking place. The goal is to allow the two users to act and feel as if they were sitting around a table exchanging brochures and illustrative material.

3.2.4 The Monitor Window

We have found it to be extremely important and useful to provide the users with the ability to monitor the recognition, analysis and synthesis implemented by the translation components of the system in order to keep track of the translation process. The NESPOLE! Monitor (see fig. 2.3, lower right side) has been developed to provide this

feedback to the users. If A and B are the two users, then fields and their functions are as follows:

- 'Remote Speech Translation' field (synthesis) field. On each side, this field reports the textual format of the translated message. So, on A's side, it features B's message as translated and spoken by the system.
- 'System Hears' field. This reports the recognized text representation of the last utterance spoken by the local user, as recognized by the speech recognizer within the HLT server for the language of the local speaker. On A' side, it displays the hypothesis string for A's last turn.
- 'System Understands' field. This is about what the system has understood about the hypothesis string. Thus, for an utterance of A, it displays the content of that very utterance, as understood by the system, generated in A's language. The purpose of this field is to provide the user with the ability to identify cases where the translation is likely incorrect, due to incorrect analysis of the spoken input utterance by the translation system.

By monitoring the 'System Hears' field, the user can verify the recognition accuracy of the last spoken utterance. Similarly, by monitoring the 'System Understands' field, the user can verify that the meaning of the utterance was correctly captured by the analyzer within the translation server (by judging whether the paraphrase back into their own language reflects the same meaning as the originally spoken utterance). When a translation failure is detected, the user can click on a 'Cancel Translation' button, which generates a red, flashing message on the monitor of the other party, alerting them to the fact that the incoming translated message should be ignored. The user can then repeat or rephrase the message. This kind of feed-back has been demonstrated to be very helpful, in particular for "expert" users, who are familiar with machine translation technology and have gained some experience with the NESPOLE! system. However, the results of the experiments (Costantini et al, 2002a; Costantini et al., 2002b) however, have demonstrated that even novice users find this type of information very useful and can learn to use the functionalities after a brief training or usage of the system.

3.3 Interface Improvements and Final Version

User studies play a basic role in development of well-designed interfaces capable of providing an effective and pleasant interaction experience for real users. Throughout the first two years of the NESPOLE! project we took advantage of many opportunities to collect data concerning the system (and the interface) usability with actual users - both computer experts and people with little to no computer skills. Most of the interface improvements leading to the final version were based on the comments and suggestions of these users, and on our observation of their behavior (Taddei et al., 2002).

3.3.1 AeWhiteboard

The main improvements in the AeWhiteboard window were about map's saving procedure, since the main problem reported by the agents involved in data collection and multimodal experiments was concerning this functionality (Taddei et al., 2002). The map saving procedure needed in fact too many steps (selecting the saving function, writing the file name, choosing the directory where the map had to be saved, ...) and interrupted the dialogue flow. A new *quick map saving* mechanism was therefore implemented to allow the users to save maps only by clicking a button: by pressing this button, the system saves the map with the gestures performed on it with an appropriate name and a progressive number in a default directory. The old saving mechanism is still available, in case the user needs for some reasons to save the map with a different name or in a different directory. Moreover, the automatic saving mechanism is activated each time the user loads a new bitmap. In fact, when the agent loads a new bitmap, the previous one would be cancelled, unless the agent saves it, and the gestures performed would be lost. The automatic saving procedure reduces the cognitive load of the agent (who does not need anymore to remember to save a map before opening another one), and prevents from the possible loss of information.

One of the most frustrating experiences encountered during the experiment were time delays due to the very long map transfer times (from about half a minute to about two minutes, depending on the bandwidth and on the network conditions). Even previously shared maps were re-transmitted whenever accessed again later in the communication. A new mechanism of map transfer has been implemented: the bitmap file is now actually transferred only if not previously transferred or locally available; otherwise, only the name of the bitmap is transferred and the remote system loads it

locally. In this way the average map transferring time was significantly reduced, and this positively affected the fluency of the conversation.

3.3.2 NetMeeting Window

One of the most frequent requests from the system users was the addition of live video transmission (Taddei et al., 2002): the facial movements and expressions of the person we are speaking with convey relevant information in the natural human-to-human communication, and their availability in a speech-to-speech translation system could increase the quality of communication providing with natural and effective feedback. Since high quality live video (CIF images) requires at least 128kb/s of bandwidth during the connection to have acceptable results, a QCIF video was experimented (QCIF has lower quality than CIF, but requires much less bandwidth). The difference of the quality between CIF and QCIF is relative to the image definition: considering that the window where the live video is available is very small, the QCIF definition is sufficient to have acceptable video images of the remote user, even in case of low bandwidth.

3.3.3 NESPOLE! Monitor and Dialogue History Window

One of the most discussed part of the interface was the NESPOLE! Monitor, the window containing feedback strings from HLT Server. Different system users have different needs concerning this window. On the one hand, people working on the project use it as a debug tool during internal tests and demos, and so they need the more information about translation process it is possible. On the other hand, novice users uninvolved with NESPOLE! use it mainly as a feedback window. They need to monitor quickly the translation process to understand if the partner received their message or not or if they had to repeat a sentence. So they need short and clear messages, and they should be able to understand the messages without having a deep knowledge about what the system is actually doing. The messages needed for the debug reasons were therefore not completely suitable to their needs, because they have not a clear meaning for people who do not know in details how the system works. For this reason we decided to differentiate between an expert user interface and a novice user one: we realized a simplified NESPOLE! Monitor window, and an improved Dialog History Window, which is configurable in two different modalities: "normal mode"

and “expert mode”. The Dialog History Window will be optional and used primarily for system demonstrations.

Another observation was that the “push-to-talk” button positioned within the Netmeeting® window, was perceived to be too small and quite difficult to manage. Thanks to the NetMeeting® SDK we a larger “push-to-talk” button was implemented within the NESPOLE! Monitor window. Some efforts were made in order to avoid the use of the “push to talk” button, since the need to check/uncheck the microphone reduces the naturalness the conversation. A silence detection algorithm was implemented, that recognizes when the user stops speaking even if he does not push the “Audio Disable” button. The algorithm proved to work well in case of quiet environments, but the “Audio Enable/Disable” button is still necessary when the environment is very noisy. That’s why it is still available in the Nespole Monitor Window (figure 2.4).

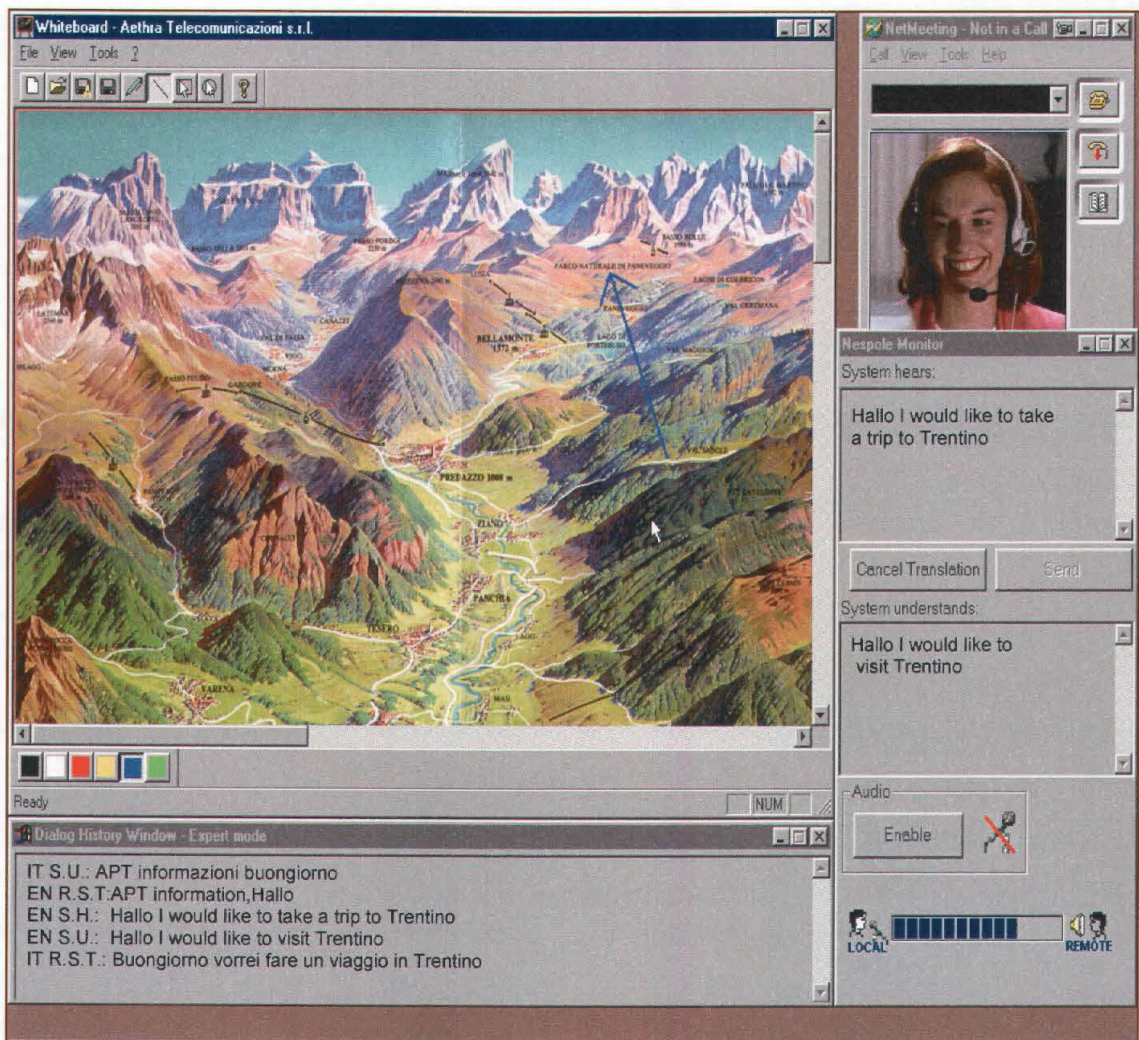


Figure 2.4: Improved interface for showcase 2a

The new NESPOLE! Monitor window contains two fields, instead of the three which were previously available (Figure 1).

1. The 'System Hears' field: it displays the recognized text representation of the last utterance spoken by the local user, as recognized by the speech recognizer within the HLT server for the language of the local speaker.
2. The 'System Understands' field: it displays a textual representation resulting from the translation of the last utterance spoken by the local user back into their own language.

The two displayed strings enable the user to evaluate if the translation process is going on well or not: by monitoring the 'System Hears' field, the user can verify the accuracy of recognition of the last spoken utterance; similarly, by monitoring the 'System Understands' field, the user can verify that the meaning of the utterance was correctly captured by the analyzer within the translation server (by judging whether the paraphrase back into their own language reflects the same meaning as the originally spoken utterance). When the user realizes that the recognizer or the analyzer output is bad, she can click on the "Cancel Translation" button. In this way, a red, flashing message appears on the monitor of the other party, signalling that the incoming translated message should be ignored. After that, the user can repeat or rephrase the message. If multiple recognition attempts of the same sentence fail, the user can manually edit the "System Hears" field, correct the sentence and resend it to the translation server, in order to eliminate the mistake made by the recognizer.

Since recognition and translation processes take some time to produce the speech synthesis output, there could be relevant time delays between a turn utterance and the delivery of its translation to the other party. During that waiting time, a user has no way to know what is happening (e.g. if the remote interlocutor has already received the translated audio) except for the information given by "System hears" and "System Understand" fields. This kind of feedback was not effective in helping to avoid overlapping speech. To give to users some additional and 'easy to read' information about the status of the translation process, we provided them with a visual feedback: a progress bar (see figure 2.4) was added within the NESPOLE! Monitor interface to inform about the sending process status and to signal, with a blinking icon, when translated audio arrives to remote interlocutor. Another progress bar informs about the remote speech processing and therefore about the arriving audio. This visual feedback was of great help to users to avoid overlapping speeches and bad turn taking.

More detailed information about the translation and transferring processes is available within the Dialog History Window, where information is ordered on a temporal base: the user could configure the Dialog History window in two different viewing modality:

- the “Expert mode” modality (see figure 1) displays all text strings produced by the two HLT Servers involved in the communication (the local one and the remote one):
 - IT S.U.: APT informazioni buongiorno
 - EN R.S.T: APT information,Hallo
 - EN S.H.: Hallo I would like to take a trip to Trentino
 - EN S.U.: Hallo I would like to visit Trentino
 - IT R.S.T.: Buongiorno vorrei fare un viaggio in Trentino
- the “Normal mode” modality displays only the text strings produced by the local HLT Server. For example the Dialog History Window of the English user will display:

- EN R.S.T: APT information,Hallo
 - EN S.H.: Hallo I would like to take a trip to Trentino
 - EN S.U.: Hallo I would like to visit Trentino

The Dialog History Window is optional and used primarily for system demonstrations and/or debug reasons. It has been found very useful by novice users during the data collection for the user study [8] and during some demos, in particular in case of low quality of the synthesized audio. In this case, in fact, a speaker could check the text corresponding to the bad quality synthesis in the Dialog History Window instead of asking the other speaker to repeat her turn, hence improving the dialogue effectiveness.

CHAPTER III

FIRST EXPERIMENT

1. Introduction

Previous research using Wizard-of-Oz technique demonstrated that, when interacting on spatial tasks, the performances of users sensibly improve if multimodal input is available, leading to faster task completion, fewer input disfluencies, less complex language and greater satisfaction (Oviatt et al., 1997a). Moreover, it was found that multimodal interaction occurs more frequently in case of spatial location commands (Oviatt et al., 1997b).

These results were obtained in highly controlled experimental conditions, in a monolingual setting. The user interacted with a computer giving command by means of speech, pen-based gestures or combination of the two modalities. A Wizard of Oz (WoZ) technique was used, i.e. in a situation where at least some of the system functionalities were simulated by a human, the wizard, and not performed by the system. In this case, recognition of both speech and gestures was simulated.

It is important to know how robust the mentioned improvements are vis-à-vis disturbing factors such as system's failures, time lag due to network traffic, etc. At the same time, when multilinguality is realized through speech-to-speech translation (STST), it is crucial to ascertain whether the use of pen-based gestures can help to overcome the weaknesses of the underlying Human-Language-Technologies, providing synergies

that the user can exploit to improve the quality and success of the interaction. We designed and executed an experiment, aiming to test:

- whether multimodality increases the probability of successful interaction, even with prototypes of 'real' multilingual systems, when spatial information is the focus of the communicative exchange;
- whether multimodality supports a faster recovery from recognition and translation errors.

The 'real' system we exploited is the first NESPOLE! showcase (see chapter 2). Two kinds of participants were involved: American English and German native speakers played the role of the customer, and native Italian speakers were trained to act as tourist agents.

In this chapter we report about this experiment. We describe methodology (task and instructions; experimental design and setting, group of participants and conventions for recordings, transcriptions and annotations. We then present and discuss the results.

2. METHOD

2.1 Experimental Design

Two experimental conditions were considered:

- a speech-only condition (SO), involving multilingual communication and the possibility for users to share images and maps through a WhiteBoard;
- a multi-modal condition (MM), where users could additionally perform pen-based gestures (pointing, area selection, connection between different areas) on shared maps to convey spatial information.

2.2 Task and Instructions

The scenario of the experiment (Winter accommodation in Val di Fiemme) was modelled after one of the five different NESPOLE! tourism scenarios, enriched with spatial information (Burger et al, 2001).

The scenario features a customer browsing the web pages of the APT, searching for information about winter holidays in Val di Fiemme, Italy. When the customer wants more information, he/she clicks on a special button, which opens a direct connection with a human agent, mediated by a speech-to-speech translation system. The customer's task was to choose an appropriate location and a hotel within constraints specified a priori concerning the relevant geographical area, the available budget, etc. The agent's task was to provide the necessary information.

The experiment involved American English native speakers (located at Carnegie Mellon University, Pittsburgh) and German native speakers (located at University of Karlsruhe), who played the role of the customers. Both interacted with Italian native speakers (located at Irst, Trento), who were trained to act as tourist agents. This resulted in four experimental groups: German customer/SO, German customer/MM, English customer/SO, and English customer/MM.

A research assistant assisted the participants during the experimental session. Customers received written information and instructions about the scenario, the task, system functionalities and interaction modalities (task and instructions are available in Appendixes 1 and 2). Before starting the interaction, we asked clients to write down the information they thought they would need to ask the agents for in order to help clients planning the conversation. In the MM condition, we demonstrated them the whiteboard functionalities, and allowed them few minutes to familiarize with the pen.

Agents were trained by Irst and instructed about how they would better answer (kinds of answers allowed, style, so as to adhere as much as possible to what 'real' agents usually do). Agents training took longer than client training, since they had to cope with be more acquainted with the functionalities of the White Board (in a real setting, one does not expect customers to have previously used the White Board and the pointing devices, whereas this should be part of agents' skills), and be proficient in the task of searching and providing the requested information. Agents were given description cards with information about two resorts in Val di Fiemme, and three hotels for each resort (Appendix 3) The agents received in addition training and instruction (Appendix 4) in proper methods of response (kinds of answers allowed, style, etc.) so as to adhere as much as possible to what 'real' travel agents usually do. For the same reasons, only agents were allowed to send maps and web pages, as it is the tourism operator and not the customer who knows which resources can be helpful at which point, where they can be found, etc.

Subjects wore a head-mounted microphone, using it in a push-to-talk mode: During the MM condition they drew gestures on maps by means of a table-pen device. Each subject could only hear the translated message of the other party (original audio was disabled). The first version of the NESPOLE! user interface was used (see chapter 2). The Aethra® White Board window, set at 600x600 resolution, used to display maps. During the MM condition, the users were allowed to draw gestures on the shared maps using the White Board drawing functionalities, which include free-hand strokes to draw arrows, lines, circles, etc.

Some pre-test dialogues were recorded in order to test task and instructions. The aim of these pre-tests was to make sure that the HLT modules (see chapter 2) were capable of supporting the task and to ascertain whether further data collection for spatial language need to be planned. ITC-irst and CMU collected locally a few monolingual dialogues (Italian-to-Italian at Irst, English-to-English and German-to-German at CMU) using a draft of the experimental task. CMU recorded 13 dialogues, 11 resulting in successful recordings (5 ENG to ENG, 2 FRA to FRA, 4 GER to GER). ITC-Irst collected 10 dialogues (ITA to ITA) with slightly different versions of the main task, in both speech-only and multimodal condition. Each dialogue took on average about 15 minutes. All dialogues were transcribed. The task and the instructions were modified during the pre-test, according to considerations related to users behaviour. Further data collection to train the translation modules to cope with spatial language proved unnecessary. Cross-sites multilingual pre-tests (IRST-CMU and IRST-UKA) were carried out with the aim of testing technical issues (connection, recording tools, etc.). In addition, 29 full dialogues (23 English-Italian, 6 German-Italian) were recorded using the full experimental setting, in order to test the task design and the instructions. 17 English dialogues and 3 German dialogues were also transcribed. The results suggested a number of modifications to the task, and some improvement to HLT modules. The resulting systems and task definition were frozen for use during the experiment.

2.3 Participants

Thirty-nine subjects participated in the experiment: 32 volunteers (16 American English and 16 German native speakers; sex balanced) played the role of the customer, and 7 native Italian speakers were trained to act as tourist agents. The participants who

played the role of the customer were paid (\$ 15 at CMU, DM 10 at UKA); agents were people working at ITC-Irst, uninvolved with NESPOLE!. All candidates were first given an enrolment form and a questionnaire on computer literacy and web expertise (Appendixes 5 and 6). Since the candidates demonstrated approximately the same level of computer literacy, they were subsequently all contacted for scheduling an appointment for the experimental sessions. The average time required for each session, including training, interaction and post-interaction questionnaire was estimated to be one hour.

2.4. Recordings, Transcriptions and Annotations

We recorded 28 successful dialogues: 14 involving an American English customer and 14 involving a German customer; all dialogues involved Italian agents. Each group consisted of 7 SO and 7 MM dialogues.

We captured the audio streams at both sides through Total Recorder¹, so to produce two audio stereo files for each dialogue, containing the voice of the local speaker recorded through the microphone and the translated and synthesized turn of the remote speaker. The audio files were transcribed. Besides orthographic words, transcription files contained annotations for turns, for spontaneous phenomena of speech and for gestures. By aligning and comparing original and translated turns with their replies, we classified all turns into successful, partially successful and non-successful. Turn repetitions (where the speaker repeats her utterance because of errors made by the system) and some other phenomena related to dialogue were counted as well. In Appendix 7 the list of all files produced during recordings, transcriptions and annotations is available.

2.4.1 Transcriptions: Conventions and Tools

Each dialogue was transcribed at both the recording sites. Transcriptions were carried out in accordance to the VERBMOBIL conventions (Burger, 1997; Burger et al., 2001), which offer an established method, a labelling set and the tools necessary for transcription and turn segmentation. The set of transcription conventions is listed in Table 1 (in Table 3.1 the two dots (..) represent any sequence of characters). More

detailed information regarding labelling and clustering of spontaneous phenomena are available in Appendix 8.

;..	Global Comment
..'.	Apostrophe (reduced word)
... (--)	Hyphen (compound word)
<*FOR>..	foreign word, specified if possible
<*ITA>..	Italian word
<*ENG>..	English word
<*GER>..	German word
<*FRA>..	French word
*..	Neologism/Mispronunciation
..%	Unintelligible
..=	Aborted Word Articulation
.._	Interruption of a Word, Left Fragment
..._	Interruption of a Word, Right Fragment
<T >..	Technical Interruption of a Word, Beginning
..< T>	Technical Interruption of a Word, End
<*T>	Technical Interruption within a Turn
<*T>t	Technical Break-off of a Turn
. ? ,	Punctuation; Period, Question Mark, Comma (separated by the rest of the text by a space)
+/..	Beginning of a Repetition/Correction
../+	End of a Repetition/Correction
-/..	Beginning of a False Start
../-	End of a False Start
	Respiration
<uh>	Filled Pause (Hesitation)
<uhm>	Filled Pause (Hesitation)
<hm>	Filled Pause (Hesitation)
<hes>	Filled Pause (Hesitation)
<%>	Unidentifiable Sound Production
<Smack>	Sound: Smacking
<Swallow>	Sound: Swallowing
<Throat>	Sound: Clearing one's throat
<Cough>	Sound: Cough
<Laugh>	Sound: Laughing
<Noise>	Other Sounds
<P>	Pause during Speech
<;..>	Local Comment

Table 3.1: Transcription: the used subset of VERBMOBIL conventions

¹ Total Recorder is a software that records streaming audio and sound card inputs (<http://www.highcriteria.com/products.htm>).

Transcriptions were done by using the TransEdit annotation tool², a Windows-based tool for transcription and segmentation. It features a graphical interface, automatic turn numbering, and format management. It also provides an audio application allowing multiple audio signals to be displayed concurrently, so that transcribers can view both the agent and client audio signals simultaneously. TransEdit also creates additional files for each transcription files with dialogue specifications and time stamps (see Appendix 9).

2.4.2 Annotation of Gestures

We developed an annotation scheme for gestures. In our annotation convention, the term *gesture* has a broad meaning, referring to all Whiteboard (WB) commands concerning shared maps and web pages.³ Thus, the following were annotated as *gestures* (for a description see the “§ 2.2: User Interfaces):

- loading images,
- running a web browser,
- scrolling images,
- zooming of images,
- free-hand strokes,
- selection of areas on the map.

The first four functions are multimedia commands which allow the exchange and exploration of visual information, and are available both in the MM and in the SO condition. Though they are performed through the Whiteboard by means of the pen+tablet device, and involve the manipulation of graphics and images, they are not on a par with *free-hand strokes* and *selection of areas*. The latter, in fact, involve the deictic/referential use of portions of images to indicate relevant locations, connect different places, etc.; hence, they directly contribute to the contents of the interaction. Those *strictu sensu* gestures characterize the MM condition, and consist of:

- free hand strokes: the user can draw arrows, lines, and other free hand strokes of her choice on the displayed image, by using a pointing device (mouse, pen+tablet).

² Burger S., Meier U., “TransEdit. A New Way to Transcribe Speech Data.” Manual by Helman J.

- selection of areas on maps: it can be done by enclosing portions of maps in elliptical/rectangular shapes drawn by means of the pointing device.

For both types of drawings, a palette was made available, yielding a selection of five different colors (black, blue, red, yellow, green).

Gestures were annotated on a copy of the agent-side transcription files. Annotators could resort to videos recorded at the agent side to recover information not explicit in the audio files. For each image used, the files also contained a bitmap including all the drawings the users had performed.

The annotation consists of three line comments placed after the corresponding turn. They include the following information (zooming of images is not included because it was never used):

1st LINE: GESTURE IDENTIFICATION

- progressive number,
- user: *agent* or *customer*, depending on who performed the gesture;
- time: *just before*, *during*, or *just after* the speech turn.

2nd LINE: GESTURE DESCRIPTION

- type: eight possibilities, corresponding to the WB commands, plus *clearing the image* and *closing the web browser*,
- description: shape and color for *free-hand strokes* and *selection*; name of the map/web-page for *loading image*, and *running a browser*, number and type for *scroll*;
- context: name of the map; only for *drawings* and *scroll*.

3rd LINE: GESTURE GOAL (only for drawings, 4 types)

- *selection of an area* — i.e., enclosing portions of maps in a figure through elliptical/rectangular shapes, or free-hand strokes — plus content (items: town name, ski area name, hotel name, bus stop, skating rink or other);
- *pointing at an area*: arrow plus pointed item (see above for items list);

³ Gesture annotation conventions can be found on the project's web site.

- *connection* (line connecting two areas) plus *items* (name of the connected areas - see above for items list);
- *word* plus the word written by the user⁴.

Details concerning annotation conventions for gestures are available in Appendix 10.

2.4.3 Alignment of Transcription Files

As mentioned above, each dialogue resulted into two transcription files: one recorded and transcribed on the customer side (containing the original voice of the customer and the synthesis of the agent-translated message) and the other recorded and transcribed on the agent side (containing the original voice of the agent and the synthesis of the customer-translated message). The two transcriptions of each dialogue were manually aligned and a new transcription file was obtained, in which every genuine turn (for both the agents and the customer) was associated with the synthesized translation, thus giving a sequence like the following (first two turns):

1st turn: - AGENT genuine
 - AGENT synthesis

 2nd turn: - CUSTOMER genuine
 - CUSTOMER synthesis

The resulting file made is possible to compare genuine and translated turns with their replies, and classify genuine turns into successful, partially successful and non-successful.

2.4.4 Annotation for Turn Successfulness

The two halves of each dialogue transcription (containing annotations) were aligned, in order to compare genuine and translated turns with their replies, and classify turns into *successful*, *partially successful* and *non-successful*:

⁴ Sometimes the agents used the free-hand modality to write a word (e.g. "bus stop", or a hotel name) on the map. In this case, the gesture annotation includes the written word like the following: *goal=word:bus stop*.

- **Successful turns** were those which had good translations, from the grammatical, syntactical and semantic point of view.
- **Partially successful turns** had poor or bad translation, 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 managed to preserve (part of) the original message, so that the targeted party could react properly. A typical example is when the translated turn contains less information than the original turn — e.g., it contains the hotel name and the double room price, but the hotel category has been dropped. Another example of a partially translated turn is when many parts of the original utterance have been omitted, but what remains still permits the other party to understand the message. E.g., the original turn states: “you can find a skating rink at Cavalese”, and the translation is “skating Cavalese”.
- A turn was labelled as **non-successful** if the other speaker couldn't understand any component of the original utterance, or else the original utterance produced no translation. The latter cases arose because of system errors: the system often fails to produce a translation and issues a “no-tag” message, or a series of question marks. Another case is that the speaker rejected the hypothesis string (the product of the speech recogniser) by pressing the ‘Cancel Translation’ button (chapter 2).

2.4.5 Other Annotations

Besides the above mentioned speech, gestures and dialogue features, other information was addressed:

- **topics**: number of discussed topics plus, for each topic, its content, the number of turns it took, the **number of returns** to such a topic (see § 3 in this chapter), the number of associated gestures;
- **spatial topics**: the same set of information as before, but limited to spatial topics;
- **ambiguities**: number of times where confusion concerning names (of hotels, towns, ski areas, etc.) developed;
- **illegal questions**: number of questions asked by client which violate the instructions concerning allowed topic.

3. RESULTS

We scheduled 53 appointments. Six volunteers cancelled the appointments. Of the remaining 47 appointments, only 28 resulted in successful dialogues, due to technical problems (system crashes, network failures, etc.) or incomplete recordings (e.g. Total Recorder was not started). In addition 5 German dialogues had to be cancelled because problems with the German HLT modules required some further improvements after they were recorded.

CANCELLED DIALOGUES	
connection problems (connection failed)	4
interrupted (connection or hlt servers crashes)	4
fully recorded and cancelled because the system was changed after recording	5
incomplete recordings	6
TOTAL NUMBER	19
SUCCESSFUL DIALOGUES	
dialogues without technical problems	20
delays due to connection problems (about 20 minutes)	3
interruption and restart during dialogue	3
synthesis crashed about 10 minutes before the end of the dialogue	2
TOTAL NUMBER	28

Table 3.2: Cancelled and successful dialogues

Among the successful dialogues, 8 suffered from technical difficulties. However, these difficulties did not significantly affect the dialogues. In particular, in 2 dialogues the synthesizer crashed about 10 minutes before the end of the dialogue, nevertheless, the users were able to successfully close the dialogue because they could read the translation of the not-synthesized turn in the 'Remote Speech Translation' field, on the Nespole Monitor window.

3.1 Turns, Tokens, Types, Dialogue Length

The total number of spoken turns, word-tokens and word-types (used vocabulary) were counted for each dialogue. A turn is operationally defined as a speaker contribution between a switching-on and a switching-off of the microphone button in the NetMeeting® window of the NESPOLE! monitor. A word-token is an occurrence of a

given word-type — e.g., the sentences “Paul is the brother of John” and “John is the brother of Paul” contains 12 word-tokens and 6 word-types.

The number of turns per dialogue was computed by adding the figures from the customer transcribed speech to those of the agent. It must be noted that the number of synthesized turns is different from the number of spoken (translated) turns, because of:⁵

- turns cancelled and then repeated by the speaker (the ‘Cancel Translation’ option described above);
- turns (e.g. long turns) that were split by the translation modules into multiple turns;
- extra-turns produced by the system in response to noise caught by the microphone, which the users had forgotten to switch off.;
- synthesis messages which were erroneously sent back to the person who produced the original one.

We obtained an average number of 73 turns per dialogue, 37 from agents and 36 from customers (39 for German customers and 33 for English customers), as shown in figure 3.1.

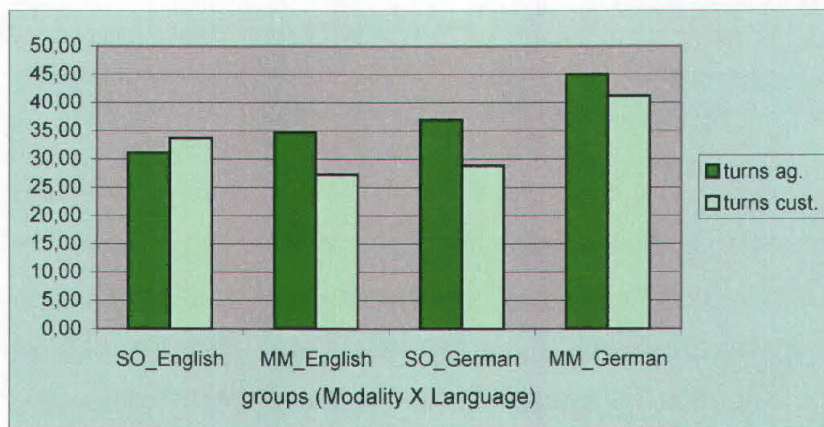


Fig. 3.1 : Average number of turns for all groups, both speakers. “SO/MM_German”: German-speaking customers; “SO/MM_English”: English-speaking customers; the agents were always speaking Italian.

⁵ Similarly, one should notice that the figures we report for word-token and word-types concern the speech actually uttered by the agent and the customer, and might well differ from the corresponding figures computed on translated speech. We do not address this issue here.

The average number of word-tokens uttered by the speakers during each dialogue is 258 for Italian agents (28 dialogues), 254 for German customers (14 dialogues) and 218 for English customers (14 dialogues). The number of word-types is 101 for agents, 103 for German customers and 82 for English customers.

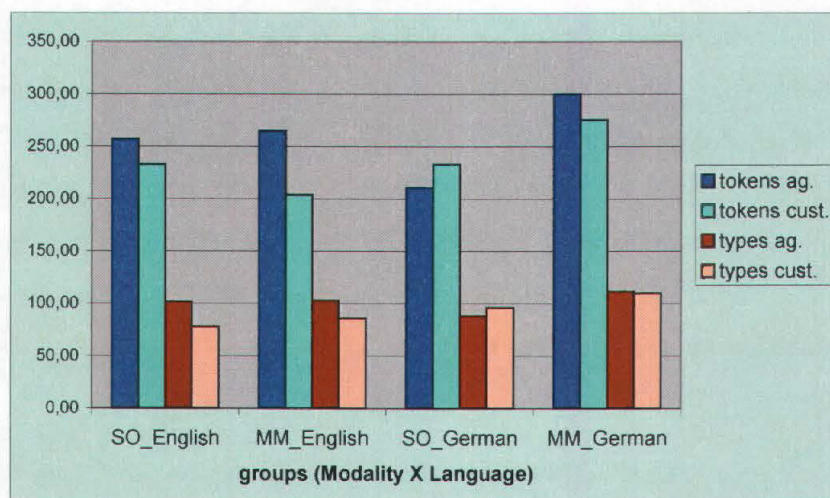


Fig. 3.2: Average number of tokens and types for all groups, both speakers.
“SO/MM_German”: German-speaking customers; **“SO/MM_English”**: English-speaking customers; the agents were always speaking Italian.

By dividing the number of tokens by the number of types, we obtain the average token/type rate, which is 2.56 for agents, 2.47 for German customers and 2.66 for English customers; those values indicate how many words were uttered before a new word was introduced.

	Italian agent	German cust.	English cust.
turns per dialogue	37	39	33
tokens per dialogue	258	254	218
types per dialogue	101	103	82
tokens per turn	6.98	6.50	6.60
token/type ratio	2.56	2.47	2.66

Table 3.3: Average number of turns, tokens, types, plus rates, for each language.

Average values and variance of all measures are similar across agents and customers and across the two conditions (Language and Modality). ANOVA tests ($p=0.05$) ran on the number of turns, agents and customers separately, did not produce significant results. Thence, there is no evidence that modality or language affected the number of words spoken.

Each dialogue lasted 35 minutes on average (36.50 in SO condition and 34.50 in MM condition). Given that the number of turns per dialogue was 73, the time lag between two consecutive turns is 30 seconds (average dialogue length in seconds divided by number of turns). That time span includes: the time during which the first turn is spoken, the translation time (including delays due to the network) and the time during which the translated message is uttered at the other site. Since turns are very brief (6.98 tokens on average for agents and 6,56 for customers) most of the time was 'waiting' time.

3.2 Disfluencies

As mentioned above, some classes of spontaneous phenomena were annotated on transcription files: a-grammatical phrases (repetitions, corrections, false starts), empty pauses, filled pauses, human noises, word interruptions and breaks, incomprehensible utterances, technical interruptions, and turn breaks; (see Appendix 8) for details. For each class of spontaneous phenomena the percentage with respect to the total number of word tokens was calculated. Percentages for the various classes, divided across agents (age) and customers (cust.) and modalities (SO; MM) can be seen in Fig. 3.3 for Italian/German dialogues, and in Fig. 3.4 for Italian/English ones. The average percentages are very low: for seven of the eight classes they are always smaller than 3% (in most of these classes even smaller than 1%). Only the percentage of empty pauses at the customer site is a bit higher, ranging from 6% to 10%.

Spontaneous phenomena were further clustered into two groups: the first includes: empty pauses, filled pauses, human noises, incomprehensible utterances; the second includes: word interruptions/breaks, turn breaks, a-grammatical phrases. This grouping was motivated by the hypothesis that the various disfluencies have different effects on turn fluency. Specifically, pauses are expected to be less disturbing than a-grammatical phrases and turn or word breaks.

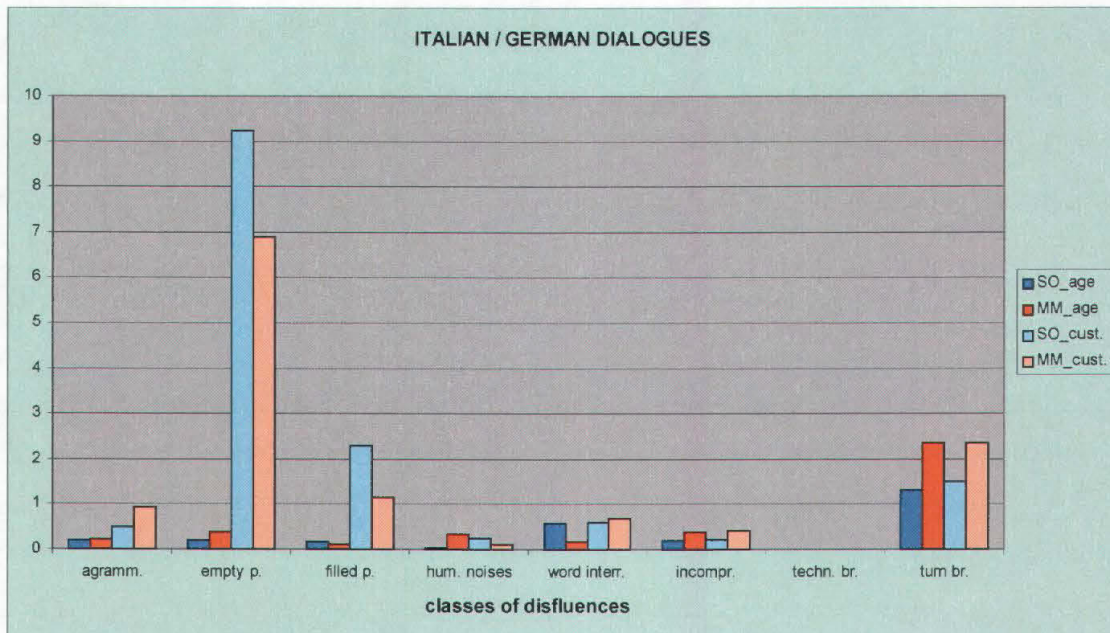


Fig. 3.3: Percentage of all classes of disfluencies for both speakers and interaction modalities, German dialogues

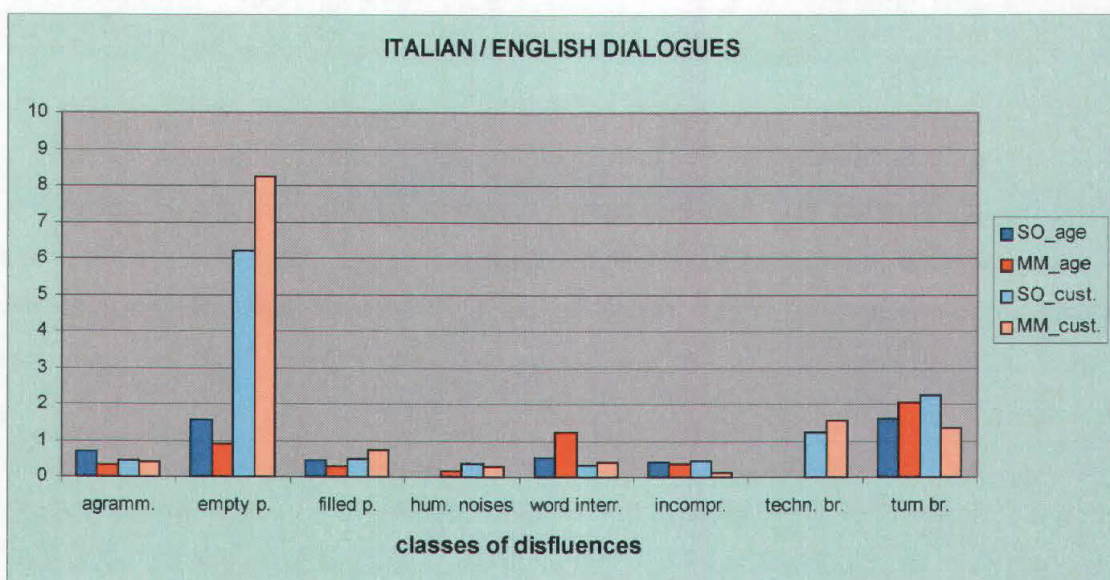


Fig. 3.4: Percentage of all classes of disfluencies for both speakers and interaction modalities, English dialogues

This led to assigning different weights to the two groups: weight 1 to pauses and incomprehensible phrases and weight 2 to the second group. We then computed a turn-fluency score, as the weighted sum of the average frequencies for each class. Notice that the score did not include technical breaks because they are related to

system features and hence do not inform about speech disfluencies. In addition, empty pauses were not included because they were not uniformly annotated across languages. In particular, Italian annotations do not report pauses exceeding a given threshold (600 ms). Hence, the numbers we have reported for agents' pauses are lower than actual figures.

We obtained an average fluency-score of 1.27 for customers (all groups, SD = 1.15) and 1.06 for agents (all groups, SD = 1.48). ANOVA tests ($p=0.05$) run on customers and agents separately didn't detect any effect of modality and/or language on the turn-fluency score. Hence, there is no evidence suggesting that turn fluency is affected by the experimental condition (MM and SO) or by customer's Language (English or German).

3.3 Pen-Based Gestures

We counted the number of *selection*, *pointing* and *connection gestures* for each dialogue, and annotated which of the White Board functionalities (*free-hand*, *line* or *elliptical/rectangular selection*) was used. In addition we counted how often agents used the *free hand* modality to write words on the map, most of these being hotel or town names associated with selection or pointing gestures.

The average number of *drawing gestures* per dialogue (MM condition) was 9. Given that the average number of turns per dialogue is 73, this means that gestures were performed on average every 8 turns. Considering that some gestures were performed together to convey a unique meaning, the number of "meaningful" gestures (sequences) is even lower, e.g. most of the pointing gestures were combined with selection gestures emphasizing the latter, rather than conveying additional information — e.g., an area was first selected and immediately after it was "pointed" at. Counting the number of pointing gestures that are performed in isolation — i.e., not in association with selection gestures — we obtain an average number of performed gestures per dialogue of 6.4. Such low ratios are probably due to the fact that interaction involving spatial information was confined to a few dialogue segments.

Drawings	% on all	Mode	% on class
Selection	61%	free-hand	65%
		elliptical	31%
		rectangular	4%
Pointing	19%	free-hand	100%
Connection	12%	free-hand	47%
		line	53%
Words	8%	free-hand	100%

Table 3.4: Percentage of performed drawing gestures and used White Board functions (MM condition)

Table 2 shows the distribution of each gesture. The figures in the table do not distinguish between the agents' and the clients' contributions, given that the agents performed almost all the drawings (98,1%). A clear preference emerges for area selections among the drawing gestures (61% of the total number of drawings), and for the free-hand mode. The optical pen was used in addition to load maps and web pages, and to scroll or zoom images. Fewer than three maps were loaded on average during each dialogue. Web pages were used rarely (0.4 on average per dialogue): in particular, there are 18 dialogues (9 SO and 9 MM) in which they were not used at all. Probably, this owes to the fact that the two available web pages contained information that was not seen as crucial, the only exception being the description of the town and the phone number of the bus service provider. Scroll was defined as a single scrolling movement or a single scrolling sequence: thus, when users perform a sequence of vertical and horizontal scrolling movements to make specific areas of the map available, we count the whole sequence as one gesture, as with isolated scrolling movements. Zoom was never used.

Three classes of temporal integration patterns between gestures and speech were annotated: immediately before, during or immediately after the corresponding turn. Table 3 reports the relevant figures, for each class of gestures. As can be seen, most of the gestures (79%) followed the speech turn, and none were performed during the turn. The typical sequence occurring when an agent wanted to use drawings (or to load maps or to send web pages), consisted of some kind of verbal anticipation of her intentions — e.g. "I'll show you the ice skating rink on the map" — followed by a

switching off of the microphone, and then by gesture performance and feedback request, for example, “Can you see the skating rink?”. It can be argued that this particular sequence and the absence of gestures during the speech were influenced by the push-to-talk procedure and the time needed to transfer gestures across the Internet. More precisely, the verbal cues were meant to alert the other party that she had to wait for a forthcoming gesture, possibly refraining from speaking in the meanwhile. This procedure allowed agents enough time to perform the gesture and ask for feedback. In addition, both microphone on/off switching and drawing functions were performed by means of the pen device. It can be argued that managing both tasks nearly simultaneously further discouraged the simultaneous execution of speech and gestures.

Drawing gestures	Before	During	After
Selection	19%	0%	81%
Pointing	26%	0%	74%
Connection	20%	0%	80%
Word	33%	0%	67%
Sum drawings	21%	0%	79%

Table 3.5: Percentages of gestures performed before, during and after the speech.

Few or no deictics were used. Sometimes the customer used indicator “here” to inform the agent that the map or the web page was on her screen (“the map is here”). No other relevant uses of deictics could be found. Agents preferred to resort to descriptive phrases that relied on visually available cues — e.g., “the skating rink is at the bottom right of the map”, “I’m selecting it with the red color”.

Those findings, too, seem related to the push-to-talk procedure. As already mentioned, users tend to avoid mixing gestures and speech. Thus, there was always a certain time lag between speech and gestures. Deictics, on the other hand, consist of linguistic markers (almost) concurrent with demonstrations (gestures). In the described situation, they would tend to be infelicitous, and rarely used.

Summing up: few gestures are performed; almost all gestures were performed by agents; gestures always followed the verbal contribution; few or no deictics were used.

3.4 Dialogue Features

In this section we report the results concerning dialogue features, in particular those concerning dialogue fluency, ambiguities, successful turns and turn repetitions.

3.4.1 Dialogue Fluency

During the dialogue the speakers sometimes returned to previously discussed topics. When occurring frequently, those returns complicate the dialogue flow and decrease dialogue fluency. Returns are usually related to difficulties in successfully closing a dialogue segment. For instance, if the customer does not obtain clear answers to her questions, she may abandon the current topic and return to it later on, asking for further clarifications. Our hypothesis that MM positively affects dialogue fluency implies that it could help speakers in successfully close dialogue segments, thus reducing the need to reiterate old topics, and yielding fewer returns. Hence, we expected a lower number of returns in MM than in SO. Moreover, it is also expected that this advantage should be clearer for dialogue segments dealing with spatial information, because MM provides alternative methods of conveying information about cartographic landmarks (e.g. drawings, pointing, etc).

The average number of returns per dialogue is 3.6. We computed two *return rates* by dividing the number of turns by the number of returns: the first over all the turns of a dialogue, and the second limited to the turns conveying spatial information. These rates indicate how many turns were spoken on average from one return to the next, and can be used as an index of dialogue fluency: the greater the index, the better the fluency. Average figures for each combination of language and modality are reported in table 3.6.

MODE	ALL TURNS		SPATIAL TURNS	
	SO	MM	SO	MM
German	21	24	13	11
English	19	31	15	44

Table 3.6: Return rate for all turns, and for turns conveying spatial information

In English dialogues there is a trend for return rates to be higher in MM condition than in SO condition. In all-turns, we have 19 turns spoken on average from one returns to the following in SO, and 31 in MM. In the only-spatial-turn condition, the figures are 15 in SO and 44 in MM. German dialogues have similar return rates in SO and MM conditions, both in the all-turns condition and in the spatial-information-only modality. We can, therefore, conclude that English dialogues show a tendency for MM to be superior to SO in terms of dialogue fluency, specifically when spatial information is conveyed. The German dialogues do not support the conclusion.⁶

3.4.2 Ambiguities

Sometimes during a dialogue agents and customers end up discussing different topics without being aware of that they are not talking about the same thing. The following is a typical example: the topic could be a certain town (Panchià), and the customer asks for information about skating rinks. The agent replies by sending the map of a (different) town (Cavalese) where there is a skating rink, but fails to inform the customer that the rink is not located in Panchià. So the customer does not distinguish between the two towns, and mistakes Cavalese's map for Panchià's. Such a misunderstanding can last for many turns and may not even be clarified by the end of the dialogue.

Direct observations of agent/customer interactions suggested that MM (i.e. gestures on the whiteboard) could aid in the resolution of misunderstandings; to check this we counted the number of dialogues in which topic confusion occurred. The number of dialogues containing ambiguities concerning place names was higher in SO (7 dialogues, 50%) than in MM (3 dialogues, 21%). Thus, multimodality seems to be effective in preventing ambiguities, when compared with speech input alone.

The number of English dialogues containing place name ambiguities is higher in the SO condition than in MM condition: 5 dialogues out of 7 (71%) in the first case, and only 2 in the second case (29%). The fact that fewer ambiguities are found in the MM condition suggests that multimodal input helps to prevent them, when compared with speech input alone.

Qualitative analysis of transcripts sharpens this point: transcripts reveal that some SO dialogues contain more than one ambiguity, which in many cases remained unsolved. In MM, the three dialogues with ambiguities contained only one of each, and

⁶ See § 4 for some hypothesis about why German is different here.

those ambiguities were solved in a couple of turns: as soon as the agent felt that the customer had not properly understood, she availed herself of the MM functionalities to select and show the customer the place she was speaking about. In the same situation, under SO condition, the agent had to resort to language for clarification, this strategy being obviously affected by the limitations of the STST system. In fact, she helped the customer to the effect of: “This town is not Panchià, it is Cavalese”, but attempts to translate this type of utterance usually generated ambiguous messages (e.g. “not Panchià Cavalese”), which were generally unhelpful to the customer. Agents, on the other hand, were usually satisfied by that kind of translation (determinable by checking the ITA generation in the System Understands Window, see chapter 2). This asymmetry usually led the agent to believe that the customer understood well and, based on this assumption, proceeded with the dialogue. As a consequence, the ambiguities remain unsolved.

The frequent failures in this respect seem to show that the paraphrases or whatever used by the speaker to recover from ambiguities were often outside the reach of the STST system. Hence, one of the main hypotheses of our study is further supported: multimodal input can indeed help overcome the limitations of STST systems, when the speech input is not able to convey the needed information. In the case discussed, solving ambiguities in SO would require the system to be capable of supporting complex interaction about the content of the interaction itself. Part of this involves providing appropriate prosodic cues. E.g., it can be argued that the utterance “not Panchià Cavalese” would have been in a better position to help disambiguating if the system were able to put appropriate stress on the word “Cavalese”, explicitly marking it as contrastively stressed. Apparently, the MM condition can circumvent the need for these pragmatic strategies by directly drawing the other party’s attention to the right object.

Inexplicably, these considerations are limited to English dialogues. In the case of the German dialogues, there is no clear indication that multimodality is advantageous over SO. We will put forth some hypotheses to explain these differences in § 4.

3.4.3 Successful Turns and Turn Repetitions

We computed the percentages of successful, partially successful and non-successful turns (see above) both on the total number of turns (“all turns”) and on legal turns only.

Legal turns are defined as turns discussing “legal” matters. A given topic was classified as illegal if it was not among those specified in the written instructions, even if it sounds reasonable within the given domain. For example our written instructions did not provide for questions about whether there is much snow in December, or whether anyone at the hotel speaks German, though these are *reasonable* questions in the tourism domain. Illegal questions were neglected to eliminate factors that could affect dialogue in unpredicted ways. Finally, the same percentages were computed for the turns conveying spatial information (“spatial turns”). The expectation was that possible effects of MM on dialogues could be better demonstrated by focusing on turns containing spatial information.

Figure 3.5 displays average distribution for each class of turns across all turns, legal turns and spatial turns. The percentage of non-successful turns for legal turns is slightly lower than that for all turns, which confirms our hypothesis that illegal topics have a misleading effect. The same values decrease even more clearly when only spatial turns are considered, pointing towards a possible positive effect of MM on turn success. The decrease of unsuccessful turns within spatial segments, in fact, is associated with an increase of partially successful turns, but not of successful turns.

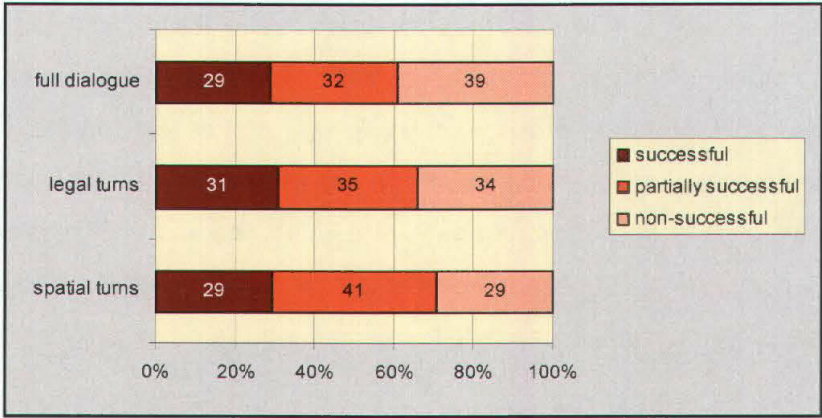


Figure 3.8: Percentages of the three classes of turn success across all turns, legal turns and spatial turns.

This suggests that some factors could improve the communicative effect of otherwise poorly translated spatial turns, enabling the other party to react properly, and permitting to classify the relevant turn as partially successful rather than non-successful. The obvious candidates are gestures in the MM condition. This hypothesis is supported by

table 5, which shows a tendency for MM to reduce the number of non-successful turns with respect to SO. This tendency is more evident in the case of spatial turns.

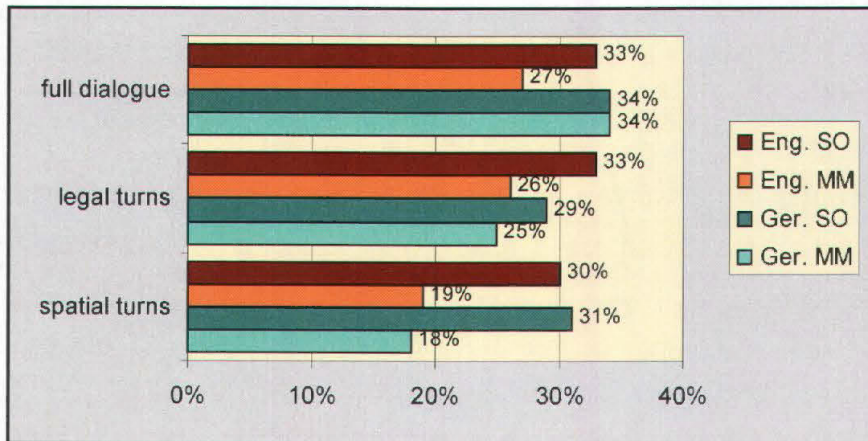


Figure 3.9: Percentages of non-successful turns on all turns, legal turns and spatial turns, split across conditions

Speakers often repeated turns in order to overcome system errors or misunderstandings. In our experiment, each repeated turn was repeated twice, on average. Table 6 reports the distribution of repeated turns. As can be seen, repeated turns tend to diminish in the MM condition (11% vs. 17% for English, and 18% vs. 23% for German), when only spatial segments are considered. This is consistent with the conclusions above: MM increases the number of partially successful turns while decreasing the number of unsuccessful ones.

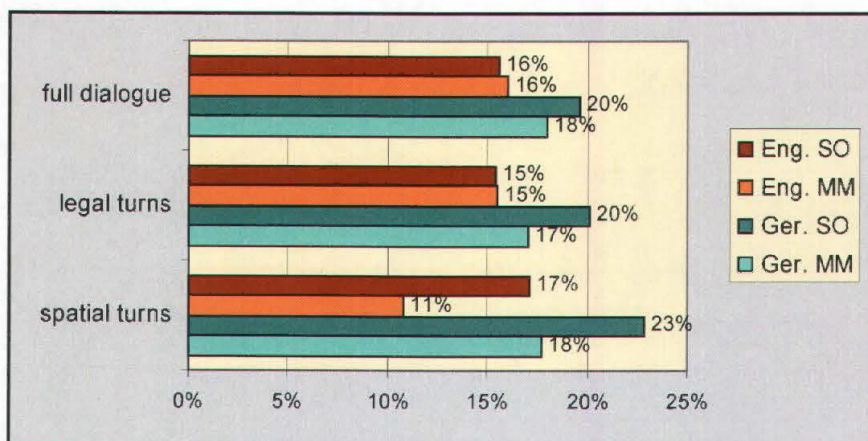


Figure 3.10: Percentage of repeated turns on all turns, legal turns and spatial turns, for all groups

It is clearly possible to conclude that multimodality can increase the probability of successful interaction and support a better recovery from translation errors, as well as reduce the number of turn repetitions.

3.5 Effectiveness

The effectiveness of the two versions of the system was measured by means of a goal attainment index, and a *subjective assessment*. *Effectiveness* is usually defined in terms of the capability of the system to support the user in completing the task, and of the quality of the task's outputs. Our effectiveness index is binary and considers only whether the users reached their goal. As to *subjective assessment*, we used the S.U.S., *System Usability Scale*, a simple ten-items scale developed at Digital Equipment Co. Ltd, Reading, UK. In addition the participants who played the role of the agent, who therefore experienced both the MM and the CO modalities, were asked for their preferences concerning the one over the other.

3.5.1 Goal Attainment

The goal of the customer is to book a hotel, meeting some assigned constraints (three-star hotel with half board accommodation, close to a bus stop or ski-area, no more that 108,5 Euro for a double room per night, etc). All available hotels were three-star hotels, and all prices included half-board accommodation. All possible hotels were out of the budget range, except for the two target hotels. The number of successful dialogues, i.e. dialogues where a target hotel was chosen, was 24 (86%), without relevant differences among modalities.

choice	target		out of budget	
Modality	SO	MM	SO	MM
German	7	6	0	1
English	6	5	1	2
Sum	13	11	1	3

Table 3.7: Reached goal concerning hotel

This demonstrates that the STST system is good enough for novice users to accomplish a task with minimal written instructions, very short initial training on the White Board, and no further assistance during the interaction. At the same time, multimodal communication did not provide any clear advantage on the completion of the particular task we chose.

3.5.2 System Usability Scale

As to *subjective assessment*, we used the S.U.S., *System Usability Scale*, a simple ten-items scale developed at Digital Equipment Co. Ltd, Reading, UK (Jordan et al, 1996). It is a ten-items *Lickert* scale; subjects must express their attitudes toward each statement by using a 5 point scale. The scale was developed starting with 50 items, which were assessed by 20 people using two examples of software systems. The statements were selected on the basis of the obtained answers, taking into account the aim of covering a variety of aspects of system usability, such as the need for support, training, system's complexity, and the need to prevent response biases (the scale is available in Appendix 11). S.U.S. yields a number (ranging from 0 to 100) representing a composite measure of the overall usability of the system under study.

S.U.S was submitted to all customers immediately after each experimental session, while agents filled out the questionnaire at the end of the experimental data collection. Therefore the S.U.S scores for customers reports about single interactions with the system (either SO or MM), while the agents' score informs about multiple interactions, without distinguishing the modality. The average S.U.S. score was 55. We found no difference between the experimental condition among customers: the average S.U.S score and the Standard Deviation for MM dialogues and SO dialogues were the same. We also discerned no difference between the agent and customer groups.

3.5.3 User Preferences

The participants who played the role of the agent, who therefore experienced both the MM and the SO modalities, were asked for their preferences concerning the one over the other. After all dialogues were recorded, we asked them to answer the following question:

If you were asked to take part again in a new experimental dialogue, which condition (interaction modality plus language) would you choose among the following?

1. *Multimodal input, English partner*
2. *Multimodal input, German partner*
3. *Speech-only input, English partner*
4. *Speech-only input, German partner*
5. *I have no preferences*

Table 3.8 reports the results grouping them along the two basic conditions of our study: MM vs. SO, and English vs. German. As to interaction modality, there is a clear preference for MM over SO: five agents preferred MM, while the remaining two were indifferent.

As to language, no preference is expressed for German, but the preference for English is not so strong as that for multimodal interaction. Indeed, only two agents chose the English language. Two additional agents indicated a *weak preference*, answering that “the best condition would be multimodal input with an English partner, but even German could be okay: what is important is that the interaction is multimodal”.

AGENT	Pref. SO	Pref. MM	Pref. Eng.	Pref. Ger.
1		X	x	
2		X	x	
3		X	X	
4		X	X	
5				
6				
7		X		

Table 3.8: Agents preferences concerning condition and language:

X = strong preference; x = weak preference

Summing up, no differences between modalities with regard to reached goals (effectiveness index) and S.U.S. scores have emerged. On the other hand, agents expressed a clear preference for MM over SO.

4 Conclusion

There seem to be few, if any, global differences between the two experimental conditions, MM and SO. Both the linguistic behaviour, as measured by the speech-related variables and effectiveness indexes do not change across conditions.

However those results should not be taken to indicate an absence of differences, nor should we assume that the kind of available multimodal interaction failed to impact at all on STST-mediated interaction. As we have observed, the measures considered so far all address global effects — that is, effects that can be detected at the level of whole dialogues. Conceivably, global effects arise when a suitable number of gestures, and a suitable number of occasions for their use, are facilitated by the task. If this is this case, our findings thus far have simply revealed that the actual number of gestures were too low to significantly affect global variables. The low number of gestures, in turn, is explained by a number of factors, most of them related to the fact that we were using a “real” system prototype, which introduced failures, errors and time delays, and produced a high number of bad turns, resulting in a great variability among dialogues (see chapter 5 for further discussion about this point).

We can conclude then that the absence of global effects of the MM/SO distinction can be traced back to the limited number of occasions in which gestures were needed. The features of the realistic scenario in which subjects operated further limited the impact of multimodality: system errors, time delays, etc.

If this is correct, we have not as yet established a concrete assessment of the impact of multimodality in our speech-to-speech translation setting. The measures and indices we have used are too crude to reveal differences that, if present, only affect subparts of dialogues. Similar considerations can be made for the effectiveness index. It addresses, in fact, only goal achievement on hotel choice; gestures, however, were relevant for spatial information, which is only one factor among many in a realistic hotel selection process. Hence, the absence of differences about effectiveness, as measured in our study, does not independently refute the importance of gesture for conveying spatial information.

When more fine-grained measures are considered, differences do start to emerge. English dialogues showed some tendencies towards better results for multimodal dialogues: shorter dialogues; fewer repeated turns; efficient dialogue

fluency (if returns can be taken as an indicator of dialogue fluency). This is even clearer for dialogue segments dealing with spatial information. In addition, there is a tendency for MM to exhibit less ambiguity, and resorting to MM resources immediately solved sporadic ambiguities.

This gives further support to the idea that gestures can positively affect the interaction and that the lack of differences between MM and SO condition in our dialogues are due to the low number of gestures. In fact, as soon as we carefully analyze small portions of the dialogue, the tendency for gestures to benefit results became increasingly visible.

There is no evidence in German data to favor MM over SO. We believe that this has a spurious effect, due to confounding intervening variables. In particular, we observed that German customers made a greater number of “illegal” questions.⁷ Their average is 3.3 in German, with an average of 28 turns dedicated to illegal question (including repetitions and answers), almost half of the total number of dialogue turns. For English dialogues the average number of illegal questions was much lower (0.4) investing only 3.6 turns. This difference suggests that some element in the German dialogues was different from the English dialogues. Since average values for all measured variables were similar in German and English dialogues, we cannot know the exact meaning of these data. It is nearly indisputable German dialogues customers were not accurate in following instructions, and that many turns were spent on topics that were not related to the task goal.

In the end, the findings so far do not tell in favour of any systematic effect of gestures on quality of interaction. However, for the mentioned reasons, they do not even provide a negative answer. More fine-grained analyses and instruments are needed to provide correct answers.

⁷ A question is illegal if it violates specific written instructions, whether or not it contradicts common sense. For example, asking whether there is much snow in December, or whether anyone at the hotel speaks German, are *reasonable* questions, but are not part of the permissible (*legal*) questions.

CHAPTER IV

SECOND EXPERIMENT

1. Introduction

The first experiment highlighted some advantages of multimodality over the speech-only modality, especially at the level of dialogue fluency. Two main issues concerning the integration of multilingual and multimodal communication were left open by this study (see chapter 3; see also: Costantini et al., 2002a and 2002b).

The first was related to the impact of the specific technique we used. The main open question concerns the extent to which multilingual communication, realized through a speech-to-speech translation (STST) system, differs from 'ordinary' monolingual communication, with respect to its dialogue structure and to the participants' communicative strategies. In particular, it would be interesting to understand 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, and which is the specific impact of the push-to-talk mode (PTT).

The second open issue was concerning methodological aspects of dialogue effectiveness evaluation: analyzing only dialogue length, number of turns, words and disfluencies as well as "classical" measures such as task accomplishment and translation successfulness, proved not to be sufficient enough to show interesting differences between different conditions. Additional efforts had to be spent at the level

of dialogue analysis, and new indexes for dialogue description and evaluation needed to be introduced, as already suggested by previous research (Anderson et al., 1996; Monk et al., 1996).

The second experiment aimed at investigating those issues, by explicitly comparing multilingual dialogues with monolingual dialogues, with and without PTT, and adopting a more structured dialogue analysis.

We start describing methodology, in particular focussing on similarity and differences with the first experiment. The description of the dialogue coding scheme we used is available in this section. We then report and discuss the results of the experiment, and draw some discussion.

2. Method

2.1 Experimental Design

The aim of the second user study was to compare multilingual dialogues with monolingual dialogues, with and without PTT, using a structured dialogue analysis. Hence, the following three experimental conditions were designed:

- 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 the previous experiment did not reveal any important cross-linguistic difference. We expected the multilingual condition to be different from the monolingual conditions with respect to dialogue length, number of words, 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.

2.2 Task and Instructions

The scenario was the same of the first experiment: a customer is browsing the web pages of an Italian tourist board office, searching for information about winter holidays in Val di Fiemme, Trentino, Italy, and opens a direct connection with the travel agency to get more information in order to plan her vacation. The task was slightly different from those used in the first experiment: 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, the planned activities, etc. (what is different here is the choice of an all-inclusive tourist package instead of the choice of a hotel, which was the task of the first experiment). The agent's task was to provide the requested information following the available descriptive cards (Appendix 12). Customers and agents both received written information and instructions about the scenario, the task, system functionalities and interaction modalities, similarly to what is described for the first experiment (instructions and task are available in Appendixes 13, 14 and 15).

For the STST condition 7 English customers located in Pittsburgh interacted with three tourist agents located in Italy through the final version of the NESPOLE! system, resulting in 7 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 both recorded in Italy. This resulted in better network connections and very limited transfer delays (see chapter 2).

The user interface was an improved version of the one used for the first experiment. The main improvements (as described in chapter 2) concerned the windows providing visual and textual feedback about the translation process, some functionalities of the White Board and the availability of a live video of the other party, allowing visual contact. In particular, the Aethra Whiteboard 6.1 was used, with screen resolution set at 1024x768, and Whiteboard size set at 750x600. The Dialogue History

Window was available in the "Normal Mode" version, allowing participants to view a record of their conversation.

Participants drew gestures on maps by means of a mouse, instead of the Whiteboard table-pen device. It was clear, in fact, from the monolingual data collection for the development of the second Showcase that most of the involved subjects preferred to use the mouse, and complained when forced to use the table-pen device (Taddei et al., 2002).

2.3 Participants

Thirteen subjects participated in the experiment: nine American English speaking volunteers played the role of the customer (five female, four male), and four native Italian speakers acted as tourism agents. Customer participants received compensation (\$15 at Carnegie Mellon University); agent participants were real tourist agents working at APT¹.

The customers were located at Carnegie Mellon University (Pittsburgh, PA, USA), while the agents were at Irst (Trento, Italy). Before taking part in the experiment, all candidates were first given an enrollment form and a questionnaire on computer literacy and web expertise (Appendixes 5 and 6). Because the candidates demonstrated approximately the same level of computer literacy, we subsequently invited all of them to schedule an appointment for the recording sessions.

The average time required for each session, including training, interaction and post-interaction questionnaire was estimated to be one hour.

2.4 Recordings, Transcriptions and Annotations

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, as in the first experiment. In the monolingual conditions each file contained both the original speech

¹ APT is the tourist board office of Trentino, a region in northern Italy, and it is a partner of the Nespole! Project.

of both speakers (the clean speech of the local speaker recorded through microphone and the remote speaker's voice transmitted through the network). Transcriptions and annotations followed the same procedure as in the first experiment, resorting to VERBMOBIL conventions for speech and to the NESPOLE! scheme for gestures (see CHAPTER 3).

However, the use of non-PTT modality (free speech) required a different definition of turn. 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. This makes difficult to compare directly measures related to number of turns in the different conditions (with and without PTT), since the figures actually have different meaning. There is a related problem with spontaneous phenomena. Some categories of disfluences may be used by the speakers to signal the end of their contributions (in particular pauses); the same phenomena might be annotated as sponaneous event in a PTT modality but not in a non-PTT modality, where it they are used as cues to decide that the turn should be closed. For those reason we do not report measures concerning number of turns and disfluencies for this experiment.

As to translation successfulness, three bilingual graders² were asked to judge separately each turn according to our scheme (in the first experiment the grading was done by two project people working in conjunction). Like in the previous experiment, turn repetitions were annotated as well.

Alignment of transcriptions was improved, so that all the (textual) information concerning each speech turn and coming from different sources is available in one single file, and easy to filter and extract (see paragraph 2.4.1 in this chapter).

In order to assess the dialogue structure, we resorted to the Dialogue Structure Coding Scheme (DSCS) from the HCRC (Human Communication Research Centre³). Description of the scheme is available in paragraph 2.5 in this chapter).

² The graders were last-year students of the Translation and Interpretation School, University of Trieste.

³ <http://www.hcrc.ed.ac.uk/Site/>

2.4.1 Alignment of Transcription Files

Each recording site had a log file running during each dialogue which reported on system outputs and the instances in time when an output happened. These log files contain textual outputs of ASR units (“system hears”), their representation in Interlingua format (IF) concepts, the generation of new units using the resulting IF concepts (“system understands”) and the translation of these new generated units into the remote language (see chapter 2 for details concerning all those steps). The other party heard the synthesized output of the translation. These synthesized outputs were also recorded and could be heard in same audio signal where the transcription was based on. The transcriber transcribed them as synthesized output using the identifications AGESYN for synthesized agent turns and CLISYN for synthesized client turns.

To get a representation of the way transcribed speaker turns went through the system resulting eventually in the transcription of the synthesized output, we aligned the transcription of both audio files together with filtered information of the log files, using excel sheets. A row of an Excel table, therefore, contained:

- the transcribed turn of speaker’s side transcription of the audio recording,
- log file output of ASR,
- log file output of concept classification in Interlingua Format (IF),
- log file output of generation ,
- log file output of translation of generation into remote language,
- and finally the transcription of the synthesized output of the remote audio recording, together with information on turn duration, token count, channel number and time marks.

Table 4.1 shows those columns of an aligned turn which contained textual information: transcription of the audio recording, system outputs and the resulting transcription of the synthesized output.

1	2	3	4	5	6	7	8
ERN KJA	yes I can see the map	{ yes I can see the map of }	{{ c:affirm} {c:give- information+feasibility+view +information-object (who=i, feasibility=feasible, info- object=(identifiability=yes, map))} }	Yes. I can see the map.	Si. Riesco a vedere la mappa.	si riesco a vedere la mappa	CLISYN

Table 4.1: Example for an aligned customer turn

Legend for Table 4.1:

- 1 = speaker's (customer) shortcut;
- 2 = manual transcription of original audio;
- 3 = speech recognition output ("system hears");
- 4 = IF concepts output;
- 5 = generation feedback output ("system understands");
- 6 = translation output;
- 7 = manual transcription of audio of the output of the synthetic voice;
- 8 = synthetic translation short cut for customer's side (*CLISYN*)

The Excel sheets served as a basis for comparing original turns with their translations. This aimed at classifying turns and annotating dialogue acts and turn topics, and allowed an easy classification of turns into successful, partially successful and non-successful. It also allows for the convenient analysis of system errors, through comparison of original speech with the outputs of all the recognition and translation steps.

2.5 Dialogue Structure Coding Scheme

To assess dialogue effectiveness, we needed to annotate the dialogues at a level different than the Interlingua speech acts (see chapter 2), for two main reasons: we needed ideally one tag per spoken turn (instead of one annotation for each SDU, see chapter 2), and we needed to capture within the same annotations some features concerning dialogue flow and structure, repetitions and reformulations, "unsolved" sequences, situations in which poor or bad translation was at the basis of misunderstandings between the two speakers. For the second experiment described here we resorted to the Dialogue Structure Coding Scheme (DSCS) from the HCRC (Human Communication Research Centre⁴).

The scheme has been developed for use on the Map Task Corpus (Anderson et al. 1991). These dialogue structure distinctions were developed within a larger vertical analysis of dialogue encompassing a range of phenomena beginning with speech characteristics. DSCD differs from previous coding schemes by boasting higher task independence than other contemporary schemes (Carletta et al., 1996; Carletta et al., 1997). In fact, this coding scheme is intended to represent dialogue structure

⁴ <http://www.hcrc.ed.ac.uk/Site/>

generically so that it can be used in conjunction with coding of many other dialogue phenomena. The categories are more independent of the task than the schemes which are devised with particular machine dialogue types in mind, and the coding scheme attempts to classify dialogue structure at higher levels.

Three levels of dialogue structure (similar to the three middle levels in Sinclair & Coulthard, 1975) are distinguished:

1. Dialogues are divided into TRANSACTIONS, which are sub-dialogues that accomplish one major step in the participants' plan for achieving the task. The size and shape of transactions is largely dependent on the task. In the Map Task a typical transaction is a sub-dialogue which gets the route follower to draw one route segment on the map.
2. Transactions are made up of CONVERSATIONAL GAMES, which are often also called dialogue games (Carlson, 1983; Power, 1979), interactions (Houghton, 1986), or exchanges (Sinclair and Coulthard, 1975), and show the same structure as Grosz and Sidner's discourse segments (1986) when applied to task-oriented dialogue. All forms of conversational games embody the observation that questions are followed by answers, statements by acceptance or denial, and so on. Game analysis makes use of this regularity to differentiate between initiations, which set up a discourse expectation about what will follow, and responses, which fulfill those expectations. In addition, games are often differentiated by the kind of discourse purpose which they have – for example, getting or providing information. A conversational game is a set of utterances starting from an initiation and encompassing all utterances up until the purpose of the game has been either fulfilled or abandoned. Games can nest within each other if one game is initiated to serve the larger goal of a game which has already been initiated (e.g. there is need for clarification before answering a question).
3. Games are made up of CONVERSATIONAL MOVES, which are simply different kinds of initiations and responses classified according to their purposes.

The coding schemes for transaction, games and moves are available in Appendix 16

Although devised for the Map Task Corpus, 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 table 4.2 shows the modified schema. A star

“*” marks those moves newly added to the DSCS schema. The proposal, disposition, action and information moves are subclasses of the former information move.

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.

The decision tree for labellers to annotate according to our adapted scheme is in Appendix 17.

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

Table 4.2: Adapted Move Annotation Scheme

3. Results

3.1 Dialogue Length, Turns and Words

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 4.1 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*: numbers, nouns, adjectives, adverbs, verbs), and function words (*func*: particles, determiners, pronouns, conjunctions). 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.

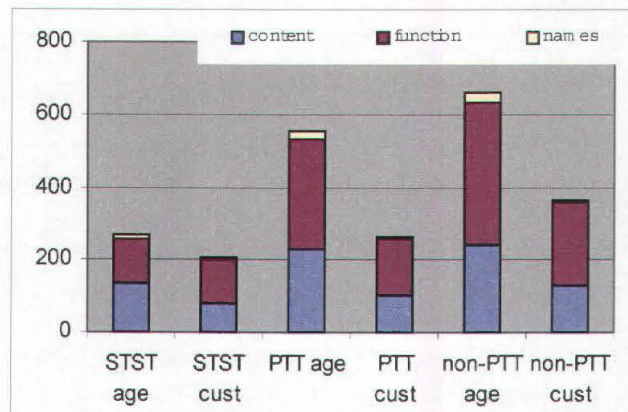


Figure 4.1: Average number of word-tokens for the three conditions, for agent and customer.

3.2 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 (annotation conventions available in Appendix 10)

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 4.3 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 4.3: 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).

3.3 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 4.2).

Figure 4.2 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.

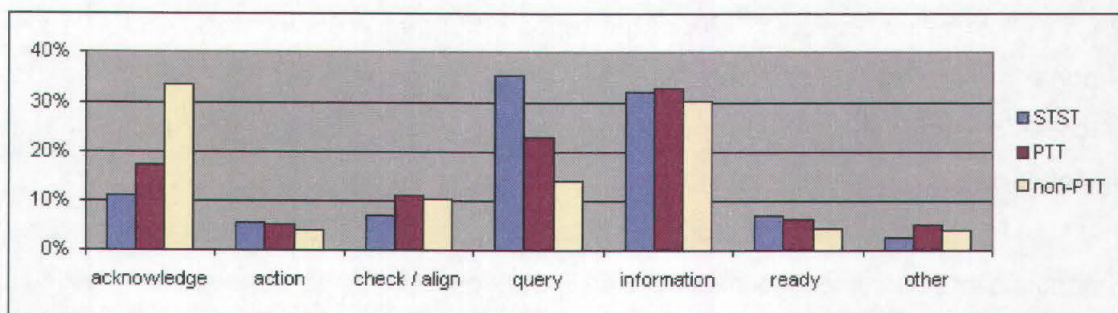


Figure 4.2: Percentages of move categories for the three conditions.

Figure 4.2 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.

3.4 Turn Repetitions

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 in STST condition 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 in the STST condition 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.

3.5 Additional Results for the STST System

3.5.1 Successful, partially successful and non-successful turns

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 (see chapter 2).

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%.

3.5.2 Manual editing of the recognized string

The actual version of the Nespole! interface provided the possibility to correct the output of the speech recognition module by typing in such cases where the recognition was already bad [3]. The user was given access to this output in the Nespole! Monitor window under "system understands". Users were instructed to use it only in cases where the recognition of their turns failed even after several repetitions and because of an obvious non-recognizing of a specific term or expression.

Customers used this option only three times during the entire data collection, each of them in a different dialogue (on a total number of 309 strings sent to the analyzer, 306

were spoken and 3 were edited). As to the Italian agents, on 324 strings 318 came from speech and only 6 from manual editing).

3.6 Questionnaires

A questionnaire was submitted to all customers immediately after each experimental session, while agents filled out the questionnaire at the end of the daily recordings. Therefore customers' answers refer to a single interaction, while the agents' refer to multiple interactions⁵.

The questionnaire consisted of 20 questions investigating system's usability issues and users' interaction experience (agents' version in STST conditions). It was not validated; therefore we cannot produce objective and valid usability scores, but they have an impressionistic value. Four out of 20 items were free-response; the other 16 required users to grade their agreement with given assertions, as follows:

- complete agreement;
- partial agreement;
- partial disagreement;
- complete disagreement.

The version for customers, STST condition, did not include the question about the procedure of map saving maps, since they were not asked to save maps. In addition, the questionnaires for the monolingual versions (both for agents and customers) did not include the item about quality of translation, for obvious reasons.

Some of the items describe the system in a positive way (e.g. "the system is effective") and some others in a negative way (e.g. "the system is difficult to use"). An expressed agreement on the positive items, therefore, translates to a positive score for the system, while an expressed disagreement on the same item means a negative score. The opposite is true for the negative items.

In Appendix 18 the customer's STST version is available. The other are available in the project web site⁶.

⁵ Each of the 4 agents came at the lab in a different afternoon during which they took part to the training session and to 2 recordings (3 for one of the agents).

⁶ <http://nespole.itc.it>

The highest agreement among responders was found on the following items concerning advantages of multimodality:

- I think that the use of maps can improve the communication.
- I think that the possibility to draw on the map can improve the communication.

On both of these items the totality of customers expressed agreement, with most answers being “total agreements”. The same was for agents, except for one who expressed “partial disagreement” with the item concerning drawings. Since both of the items are expressed positively, this means that multimodality in NESPOLE! is perceived as helpful in supporting the communication.

The video and the web pages were rated as useful more frequently in the STST condition, suggesting that they can be of help in conveying information or feedback in particular when the communication is made more difficult by translation errors and time delays.

Almost half of the responders across conditions rated the graphic presentation as not pleasant, even if the function of the elements on the screen is rated as very clear by the majority of responders. This could reflect the fact that in improving the user interface we focussed our efforts on utility and not on aesthetical aspects. Despite of the efforts spent to increase the quality of the user interface in the second year of the project, more than half participants acting as customers judged the system feedback as partially inadequate, suggesting that the interface is a quite delicate aspect of the system, on which developers should put much more attention. However, the agents judged the feedbacks adequate. Since the agents answered the questionnaire only after having participated in more than one dialogue, while the customers answered after a single interaction, the differences here could mean that the feedback provided by the system takes some time to become familiar with.

The overall system is judged by the majority of responders as innovative, effective, helpful and easy to use, without relevant differences across conditions.

4 Conclusion

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

As a general remark, the overall results for the monolingual-PPT condition were usually 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.

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.

CHAPTER V

DISCUSSION AND CONCLUSIONS

1. Summary of Results

In this work we have described two experiments performed through a multimedia multimodal Speech-to-Speech translation (STST) system, called NESPOLE!. The general aim was to investigate how multimodal resources could support effective communication between two humans speaking different languages, when it is mediated by a STST system.

Our first experiment aimed at investigating the added value of multimodality in such a scenario by comparing two versions of the NESPOLE! system: the Multimodal version (MM), where pen-based gestures could be used in addition to speech during a conversation between two speakers, and the Speech-Only version (SO), where people had to rely only on speech.

Previous works in a monolingual scenario demonstrated that, when users interact with a computer on spatial tasks, their performances sensibly improve if combined speech and pen-based inputs are available, leading to faster task completion, fewer input disfluencies, less complex language and greater satisfaction

(Oviatt et al., 1997a). Those results were obtained in a situation where the user interacted with a computer giving command by means of speech, pen-based gestures or combination of the two modalities. A Wizard of Oz (Woz) technique was used, i.e. in a situation where at least some of the system functionalities were simulated by a human, the wizard, and not performed by the system. In this case, recognition of both speech and gestures was simulated. The first experiment aimed at testing: a) whether multimodality increases the probability of successful interaction, even with prototypes of 'real' multilingual systems, when spatial information is the focus of the communicative exchange; and b) whether multimodality supports a faster recovery from recognition and translation errors. The 'real' system we used is the first showcase of NESPOLE!. 14 German-speaking and 14 English-speaking novice users interacted with seven Italian-speaking travel agents in a push-to-talk mode, producing 28 dialogues. We compared two conditions: in the first condition (Multimodal, MM) the users could utilize the multimodal facilities (pen-based drawings); in the second condition (Speech-Only, SO) they had to rely only on speech. The customer's task in the experiment was to choose an appropriate location and a hotel within specified constraints, concerning the relevant geographical area, the available budget, etc. The agent's task was to provide the necessary information.

The results show that multimodal interaction seems not to affect the dialogue length, the number of spoken turns and words, and the number of disfluencies and spontaneous phenomena. On the other hand, it seems quite capable of enhancing dialogue effectiveness. When spatial information is conveyed, multimodal input is better than speech-only in decreasing the number of ambiguities, repetitions and non-successful turns; in addition, it helps in solving misunderstandings and provides for more fluent dialogues. Moreover, when explicitly asked to express a preference between the MM and the SO condition, the users indicated a clear preference for the MM system version.

Two main issues concerning the integration of multilingual and multimodal communication were left open by the first study. The first was related to the impact of technical aspects: how significantly does the specific STST system impact on the way speech and gestures are integrated, given its delays, translation errors and technical problems. What is the impact of the push-to-talk mode (PTT)? The second concerned dialogue structure: the first study only considered dialogue length, number of disfluencies, number of turns, and vocabulary counts, as well as other "classical"

measures such as task accomplishment and translation successfulness. The results showed that these dimensions might not be the right ones to investigate differences at the level of dialogue structure. These considerations motivated a second experiment.

The second study aimed at: a) explicitly comparing multilingual dialogues with monolingual ones, with and without PTT, and b) exploiting a more structured approach to dialogue analysis. It resulted in three experimental conditions: STST condition: multilingual (English/ Italian), using the STST translation in a push to talk mode; PTT condition: monolingual (Italian/Italian) interactions in a push to talk mode; Non-PTT condition: monolingual (Italian /Italian) interactions without push to talk. We expected the multilingual condition to be different from the monolingual ones 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.

For the STST condition seven English customers located at Carnegie Mellon University (Pittsburgh, Pa) interacted with three tourist agents located in Trento (Italy) through the final version of the NESPOLE! system, yielding seven recorded dialogues in a PTT mode. The same three agents acted as agents 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. Task, procedure and annotations were similar to those of the first experiment: we used a slightly modified task, the same recording and transcription modalities, partially different annotations (see chapter 4). In addition, all dialogues were annotated following a dialogue structure annotation schema, which was an adaptation of the Dialogue Structure Coding Scheme (DSCS) of the HCRC (Human Communication Research Centre¹. We annotated dialogue games, which are sets of utterances sharing a common goal, and conversational moves, which are different kinds of initiations and responses classified according to their purposes, e.g. opening, checking, affirmative replies, etc.

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 and reduces the number of words spoken per dialogue, especially for agents. As to dialogue structure, the STST dialogues are characterized by: shorter

¹ <http://www.hcrc.ed.ac.uk/Site/>

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 acknowledgments. 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, and a clear trend for gestures to be more often associated with speech in the monolingual non-PTT condition than in the others. As a general remark, the overall results for the monolingual-PPT condition were usually 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.

2. Discussion

2.1 Added Value of Multimodality

We were able to find evidence in favor of Multimodality over Speech-Only modality. However, most of advantages of multimodal input were weaker than those found in simpler monolingual scenarios (Oviatt et al., 1997a), and far from being supported by statistical evidence. This is not only due to the low number of dialogue per condition (problems in scheduling and managing cross-sites appointments over the network, especially with the US, and the need for accurate transcriptions and annotations prevented us from increasing the number of collected dialogues). The fact that we used a real system prototype instead of, for example, a Wizard-of-Oz is of primary importance to assess the results.

First of all we could not exploit the full power of multimodality because the limitations of the translation modules would have not been supported it. As a consequence gestures could be used only as a support to verbal interaction, and not to convey substantial meaning per se. The task de designed was the best compromise between the system's capabilities at that time, and the need to provide for true pen-based gestures. This might explain the limited number of gestures observed (1 every 8

spoken turns in STST conditions), which, in turn, might account for the fact that multimodality did not strongly affect a number of variables measuring the verbal interaction (average values and frequencies for linguistic phenomena, task accomplishment, etc.).

In addition, the use of a “real” system prevented us from manipulating/controlling all the relevant variables. In particular, it is possible that some system features – e.g., the time required by the translation process and/or to manipulate (transfer, load and save) shared objects, the quality of the translation and of the user interface - affected the interaction more than the targeted variables (e.g. presence versus absence of gestures, in the first experiment).

However, despite our difficulty in finding strong evidences, all the trends were in the direction of an advantage of multimodality. In addition, the observers and participants of the first experiment had the clear impression that the multimodal condition was better (more fluent, effective, less frustrating) than the speech-only one. We had to find different measures to capture what was actually different between the two situations. Methods addressing dialogue structure, such as the Dialogue Structure Coding Scheme adopted for the second experiment, seem to be valid in this sense.

2.2 Speech-Gestures Integration

We used the number of gestures performed before, during and after the corresponding speech turn as an index of the level of integration between speech and drawing gestures in dialogues (see chapter 3). A high frequency of gestures performed during speech was considered as indicating a strict association between speech and gestures; a high number of gestures performed after the speech was interpreted as indicating that the user wanted to inform her partner about the gesture, before its arrival. This, in turn, could turn out to be useful to prevent the partner from speaking while the gesture was underway (chapter 3).

	STST exp1-mm	STST exp2	Monoling. PTT	Monoling. Non-PTT
Before	21%	32%	8%	0%
During	0%	14%	61%	96%
After	79%	53%	31%	4%

Table 5.1. Percentages of turns performed before, during or after the corresponding speech turn.

In table 5.1 we collect data from both the first and the second experiments. The results show that the integration of speech and gestures improves from the first to the second experiment, and within the second experiment progressively from STST to PTT and to NON-PTT condition. In particular, in the first experiment most of the drawings followed the speech turn (79%) and no drawings at all were performed during speech. A small percentage of gestures performed with speech appeared in the STST condition of the second experiment; the percentage increased considerably in the monolingual conditions, reaching 96% in the non-PTT condition. In addition, in the monolingual conditions gestures are rarely anticipated by speech (almost never in non-PTT).

Those data suggest that an appropriate level of integration between speech and gestures can be realized in scenarios of remote computer-mediated communication. However, this seems also to be a quiet delicate feature that can be lost as soon as more tasks have to be handled in parallel, or the overall context of the conversation - e.g., the temporal lag between successive turns – starts differing from the ‘norma’ one.

2.3 Push-to-Talk Procedure

As shown in the previous paragraph, the presence of the Push-to-Talk procedure requires adaptations by the users, resulting in multimodal integration patterns that are different from those observed in free-speech (more “natural”) conversations. It seems possible to generalize this consideration to other dialogue features, e.g., the number of words, number of games and number of moves per game, frequency of single moves per dialogue.

We can reasonably argue that the PTT modality used in the STST conditions played a role in determining the observed results for multilingual dialogues. It would be interesting to investigate free-speech multilingual dialogue to test this hypothesis. An attempt in this direction was made, by using silence-detection techniques; unfortunately, their performances turned out to be insufficient to allow for a true, PTT-free interaction.

2.4 Multilingual versus Monolingual Dialogues

It is an interesting finding that multilingual dialogues can be different from monolingual ones. In our case, the latter featured an higher number of words, in particular function words (determiners, conjunctions, articles, pronouns), indicating a richer and more complex language structure. In addition there are differences even in the kind of sentences (moves) that are used, indicating that the dialogues are more strictly focused on exchange of information (see chapter 6). Those results suggest that the analysis of the communication styles may be of great interest to the STST research community, particularly regarding the choice of training materials. Indeed usually training data for STST systems are collected in monolingual (and/or Wizard-of-Oz) scenarios, while systems are designed to work in a multilingual scenario; if the two kinds of scenario produce dialogues with different structures, this can undermine the system performances. For example, in the scenarios covered by the NESPOLE! system, the worst translated turns were meta-communicative ones, which are highly underrepresented in the monolingual database exploited for NESPOLE! (as well as for other similar projects) and, therefore, left unaddressed by the resulting system.

3. Conclusions

The reported results show advantages of multimodality over speech-only modalities even in the case of speech-to-speech translation technologies and computer-mediated communication. They suggest that in multimodal systems, increased complexity does not always mean greater interaction difficulty, even in case of systems supporting human-human communication. In fact, the addition of input modalities, may actually lead to more efficient and pleasant interaction experiences, when a sufficient integration between modalities is achieved. They show in addition that users may flexibly adapt communication strategies to different contexts and interaction modalities. It is suggested that classical evaluation methods (based on the number of errors made by users, word error rates, task completion time, etc) should be integrated with accurate dialogue analysis to help understand the phenomena, and to reliably evaluate multimodal CMC systems. User studies in the field of STST systems are a quiet recent research topic, which is still at its very first steps. We think we could extend what Oviatt

wrote about speech interfaces (Oviatt, 1996) to the more specific field of speech-to-speech translation technologies:

“To date, the development of spoken language systems primarily has been a technology-driven phenomenon. As speech recognition has improved, progress traditionally has been documented in the reduction of word error rates. However, reporting word error rate fails to express the frustration typically experienced by users who cannot complete a task with current speech technology. Although the successful design of interfaces is essential to supporting usable spoken language systems, research on human-computer spoken interaction currently represents a gap in our scientific knowledge. Moreover, this gap is widely recognized as having generated a bottleneck in our ability to deploy robust speech technology in actual field settings. [...] Many basic issues need to be addressed before technology can leverage fully from the natural advantages of speech—including the speed, ease, spontaneity, and expressive power that people experience when using it during human-human communication. For example, research is needed to evaluate different types of natural spoken dialogue, spontaneous speech characteristics and their management, and dimensions of human-computer interactivity that influence spoken communication. With respect to the latter, research is especially needed on optimal delivery of system confirmation feedback, error patterns and their resolution, flexible regulation of conversational control, and management of users' inflated expectations of the *interactional* coverage of spoken language systems. In addition, the functional role that ultimately is most suitable for speech technology needs to be evaluated further. Finally, assessment is needed of the potential usability advantages of multimodal systems incorporating speech over unimodal speech systems, with respect to breadth of utility, ease of error handling, learnability, flexibility, and overall robustness.”

References

- Ahrenberg, L., Dahlbäck, N. and Jönsson, A. (1995). Coding Schemes for Studies of Natural Language Dialogue, Working Notes from AAAI Spring Symposium, Stanford, 1995.
- Alexandersson, J., Buschenbeck-Wolf, B., Fujinami, T., Maier, E., Reithinger, N., Schmitz, B., Siegel, M. (1997). Dialogue Acts in Verbmobil-2. Technical Report 204, BMBF.
- Allen, J. (1983). Recognising intentions from natural language utterances. In Brady, M. and Berwick R. (eds.) Computational Models of Discourse. MIT Press.
- Allen, J. and Core, M. (1997). Draft of DAMSL: Dialog Act Markup in Several Layers.
- Allen, J and Perrault, C. (1980). Analyzing intention in utterances. *Artificial Intelligence*, 15 (3).
- Allwood, J. (1995): An activity based approach to pragmatics. Technical report (GPTL) 75, Gothenburg Papers in Theoretical Linguistics, University of Göteborg.
- Anderson, A. H., Bader, M, Bard, E.G., Boyle, E., Doherty-Sneddon, G., Garrod, S, Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, H., Thompson, H. and Weinert, R. et al. (1991). The HCRC Map Task Corpus, *Language and Speech* 34 (4), 351-366.
- Anderson, A. , Newlands, A. , Mullin, J. , Fleming, A. , Doherty-Sneddon, G. , Velden, J. Vander (1996). Impact of videomediated communication on simulated service encounters.
- Angelini, A., Cettolo, M., Corazza, A., Falavigna, D., Lazzari, G. (1997). Multilingual Person To Person Communication At Iirst. In Proceedings ICASSP '97.
- Austin, J. L. (1962, 1975). How to Do Things with Words. Second edition, Urmson and Sbisà (eds.). Harvard University Press.
- Bolt. Put-that-there (1980). Voice and gesture at the graphic interface. *Computer Graphics*, 14(3):262—270, August 1980.
- Bub, W.T. and Schwinn, J. (1996). Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. Proceedings of ICSLP (2371—2374).
- Burger, S., Besacier, L., Coletti, P., Metze, F., Morel C. (2001). The NESPOLE! VoIP Dialogue Database, Eurospeech 2001, Aalborg, Denmark, September 2001.
- Burger, S., Costantini E., and Pianesi F. (2003). Communicative Strategies and Patterns of Multimodal Integration in a Speech-to-Speech Translation System. Proceedings of the conference: Machine Translation Summit IX, New Orleans, U.S.A, September 2003.

- Burger, S., Costantini E., Pianesi F., Gerbino W., and Taddei L. (2003). The NESPOLE! Multimodal speech-to-speech translation system: user based system improvements. HAAMAHA 2003 (Human Aspects of Advanced Manufacturing: Agility and Hybrid Automation), Roma, Italy, May 2003.
- Burger, S. (1997). Transliteration spontan-sprachlicher Daten–Lexikon der Transliterationskonventionen – VERBMOBIL II”, Verbmobil TechDoK-56-97, München.
- Carletta J.; Isard, A.; Isard, S.; Kowto, J.; Doherty-Sneddon, G. and Anderson, A. (1996). HCRC Dialogue Coding Manual, HCRC Technical Report, HCRC/TR-82, June 1996.
- Carletta J.; Isard, A.; Isard, S.; Kowto, J.; Doherty-Sneddon, G. and Anderson, A. (1997). The Reliability of a Dialogue Structure Coding Scheme. Computational Linguistics, vol 23, no 1.
- Carlson, L. (1983). Dialogue Games: An Approach to Discourse Analysis. D. Reidel.
- Cattoni, R., Federico, M., and Lavie, A. (2001). Robust Analysis of Spoken Input combining Statistical and Knowledge-based Information Sources. Proceedings of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2001), Madonna di Campiglio (Trento), Italy, December 2001.
- Cassel, J., Sullivan, J., Prevost, S., Churchill, E. (2000). Embodied Conversational Agents. The MIT press.
- Cohen, P. and Levesque, H. (1991). Confirmations and joint action. In Proceedings IJCAI-91.
- Cohen, P. and Perrault, C. (1979). Elements of a plan-based theory of speech acts. Cognitive Science, 3 (3).
- Costantini, E., Pianesi, F., and Burger, S. (2002a). NESPOLE! Multilingual and Multimodal Corpus”, in Proceedings of LREC (Third International Conference on Language Resources and Evaluation) 2002, Las Palmas, Spain, May 2002.
- Costantini, E., Pianesi, F., and Burger, S. (2002b). The Added Value of Multimodality in the NESPOLE! Speech-to-Speech Translation System: An Experimental Study. Proceedings of ICMI 2002 International Conference on Multimodal Interfaces, Pittsburgh, PA, U.S.A., October 2002.
- Costantini, E., Burger, S., and Pianesi, F. (2003). Communication effectiveness in Multimodal and Multilingual Dialogues. Proceedings of DIABRUCK 2003 (7th Workshop on the Semantics and Pragmatics of Dialogue), Wallerfangen, Germany, September 2003.
- Coutaz, J., Nigay, L., and Salber, D. (1993). The MSM framework: A design space for multi-sensori-motor systems. In L. Bass, J. Gornostaev, and C. Under, editors, Lecture Notes in Computer Science, Selected Papers, EWCHI'93, East-West Human Computer Interaction, pages 231--241. Springer-Verlag, Moscow, August 1993.

- Doherty-Sneddon, G. , Anderson, A. , O'Malley, C. , Langton, S. , Garrod, S. , Bruce, V. (1997). Face-to-face and video mediated communication: a comparison of dialogue structure and task performance. *Journal of experimental psychology: Applied*, vol 3(2), pp. 105-125.
- Grosz, B. and Sidner, C. (1986). Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12(3):175-204.
- Hinkelman, E. and Allen, J. (1989). Two constraints on speech act ambiguity. In *Proceedings of ACL*.
- Herring, S. (1996). *Computer-Mediated Communication: linguistic, social and cross-cultural perspectives* (Editor), John Benjamins Publishing co. Amsterdam/Philadelphia.
- Herring, S. (2003). *Computer-Mediated Discourse Analysis: an Approach to Researching Online Behavior*. In Barab, S. A., Kling, R. and Grazy, J.H. (Eds.), *Designing for Virtual Communities in the Service of Learning*. New York: Cambridge University Press.
- Houghton, G. (1986). *The Production of Language in Dialogue: A Computational Model*. PhD thesis, University of Sussex, April 1996.
- Hovy, H., Ide, N., Frederking, R., Mariani, J. and Zampolli, A. (eds), (2001). *Multilingual Information Management: Current Levels and Future Abilities*. *Linguistica Computazionale, Volume XIV-XV*, Publisher: Istituti Editoriali e Poligrafici Internazionali, Pisa, Italy, 2001.
- Jordan P.W., Thomas B., Weerdmeester B.A. and McClelland I.L. (1996). *Usability Evaluation in Industry*. London: Taylor and Francis.
- Kies, J. K. and Williges, R. C. (1997). *Desktop Video Conferencing: a System Approach*. In Helander, M. G., Landauer, T.K., and Prabhu, P. V. (Eds), *Handbook of Human-Computer Interaction*, Elsevier North-Holland, 1997.
- King, M. and Falkedal, K.. Using test suites in evaluation of MT systems (1990). In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 211--216, Pittsburgh, Pennsylvania, 1990. Association for Computational Linguistics.
- J. Kowtko, et al. (1992). *Conversational Games within Dialogue*, Research Paper HCRC/RP-31, Human Research Communication Centre, University of Edinburgh.
- Lazzari G. (2000). *Spoken Translation: Challenges and Opportunities*. In B. Yuan, T. Huang, X. Tang (eds.) *Proceedings of 6th International Conference on Spoken Language Processing [ICSLP 00]*, Beijing, China, vol IV, pp. 430-435, October 16-20, 2000.
- Lazzari, G., Frederking, R., Minker, W. (2001). *Speaker-Language Identification and Speech Translation*, Chapter 7, in Hovy, H., Ide, N., Frederking, R., Mariani, J. and Zampolli, A. (eds), (2001). "Multilingual Information Management: Current Levels and Future Abilities", *Linguistica Computazionale, Volume XIV-XV*, Publisher: Istituti Editoriali e Poligrafici Internazionali, Pisa, Italy, 2001.
- Lavie A., Langley C., Waibel A., Lazzari G. , Pianesi F., Coletti P., Balducci F., Taddei L. (2001). *Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-commerce Applications*. *Proceedings of HLT 2001 Human Language Technology Conference*, San Diego, California, March 18-21, 2001.

- Lavie, A., Metze F., Cattoni, R., Costantini, E., Burger, S., Gates, D., Langley, C., Laskowski, K., Levin, L., Peterson, K., Schultz, T., Waibel, A., Wallace, D., McDonough, J., Soltau, H., Mana, N., Pianesi, F., Pianta, E., Besacier, L., Blanchon, H., Vaufreydaz, D. and Taddei, L. (2002). A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System. Proceedings of ACL 2002 workshop on Speech-to-speech Translation: Algorithms and Systems, Philadelphia, PA, U.S.A., July 2002.
- Lavie, A., Levin, L., Schultz, T., Langley, C., Han, B., Tribble, A., Gates, D., Wallace, D., Peterson, K. (2001). Domain Portability in Speech-to-Speech Translation" [.pdf]. In the Proceedings of HLT 2001 Human Language Technology Conference, San Diego, California, March 18-21, 2001.
- Lea, M. (1992). Contexts of Computer-Mediated Communication. (Editor), Harvester Wheatsheaf, Hertfordshire.
- Levin L., Bartlog B., Font Llitjos A., Gates D., Lavie A., Wallace D., Watanabe T., Woszczyna M. (2000). Lessons Learned from a Task-Based Evaluation of Speech-to-Speech Machine Translation. In the Proceedings of LREC 2000, Athens, Greece, 2000.
- Levin, L., D. Gates, A. Lavie, F. Pianesi, D. Wallace, T. Watanabe, M. Woszczyna (2000). Evaluation of a Practical Interlingua for Task-Oriented Dialogues. In the Proceedings of the AMTA-SIG-IL Workshop On Interlinguas and Interlingual Approaches, Seattle, 2000.
- Levin, L., D. Gates, D. Wallace, Peterson, K., A. Lavie, F. Pianesi, Pianta, E., Cattoni, R., Mana, N. (2002). Balancing Expressiveness and Simplicity in an Interlingua for Task based Dialogue. Proceedings of ACL 2002 workshop on Speech-to-speech Translation: Algorithms and Systems, Philadelphia, PA, 7 July 2002.
- Litman, D. and Allen, J. (1992). Discourse Processing and Commonsense Plans. In Cohen, Morgan and Pollack (eds.) Intentions in Communication. MIT Press.
- Mana, N., Burger, S., Cattoni, R., Besacier, L., MacLaren, V, McDonough, J., Metze, F. (2003). The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domain. In proceedings of Eurospeech 2003, September 1-4, Geneva,
- Mann, W. and Thompson, S. (1988): Rhetorical Structure Theory: Toward a functional theory of text organisation. Text, vol. 8 (3).
- Maybury, M. T. and Wahlster, W. (1998). Readings in Intelligent User Interfaces (Eds). Morgan Kaufmann Press.
- Metze, F., McDonough, J., Soltau, H., Lavie, A., Levin, L., Langley, C., Schultz, T., Waibel, A., Cattoni, R., Lazzari, G., Mana, N., Pianesi, F., Pianta, E. (2002). Enhancing the Usability and Performance of NESPOLE!: a Real-World Speech-to-Speech Translation System, HLT 2002, San Diego, California U.S., March 2002.
- Metze, F., McDonough, J., Soltau, H., Langley, C., Lavie, A., Levin, L., Langley, C., Schultz, T., Waibel, A., Cattoni, R., Lazzari, G., Mana, N., Pianesi, F., Pianta, E. and Costantini, E. (2002). The NESPOLE! Speech-to-Speech Translation System, HLT 2002, San Diego, California U.S, March 2002.
- Metze F, McDonough J, Soltau H (2001) Speech Recognition over NetMeeting Connections", University of Karlsruhe, Germany, 2001.

- Monk, A. F., McCarthy, J. C., Watts, L. A., & Daly-Jones, O. (1996). Measures of process. In P. Thomas (Eds.), *CSCW Requirements and Evaluation* (pp. 125-139). Berlin: Springer Verlag.
- Nickerson, R. S. and Landauer, T.K. (1997). *Human-Computer Interaction: Background and Issues*. In Helander, M. G., Landauer, T.K., and Prabhu, P. V. (Eds), *Handbook of Human-Computer Interaction*, Elsevier North-Holland, 1997.
- Olson, G.M., Olson, J. S., (2003). *Human-Computer Interaction: Psychological Aspects of the Human Use of Computing Annual Review of Psychology*, February 2003, Vol. 54: 491-516
- Oviatt, S. (1996). *Usability and Interface Design*. In Mariani, J, Uszkoreit, H., Zaenen, A. and Zue, V.: *Survey of the State of the Art in Human Language Technology*. Report for NSF and UE.
- Oviatt, S.L. (1997a). *Multimodal Interactive Maps: Designing for Human Performance*, *Human-Computer Interaction*, 1997, pp. 93-129 (special issue on "Multimodal interfaces").
- Oviatt, S. L., De Angeli, A. and Kuhn, K. (1997b). *Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction*, in *Proceedings of CHI '97*. ACM Press, New York, 1997, pp. 415-422.
- Oviatt, S. L. (1999a). *Mutual Disambiguation of Recognition Errors in a Multimodal Architecture*, in *Proceedings of CHI '99*, ACM Press, New York, 1999, pp. 576-583.
- Oviatt, S. L. (1999b). *Ten myths of multimodal interaction*, *Communications of the ACM*, Vol. 42, No. 11, November, 1999, pp. 74-81.
- Oviatt, S. L., Bernard, J. and Levow, G. (1999). *Linguistic adaptation during error resolution with spoken and multimodal systems*, *Language and Speech*, 1999, Vol. 41, nos. 3-4, 415-438 (special issue on "Prosody and Conversation").
- Oviatt, S.L., Cohen, P.R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J. & Ferro, D. (2000) *Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions*, *Human Computer Interaction*, 2000, vol. 15, no. 4, 263-322 [Reprinted in *Human-Computer Interaction in the New Millennium* (ed. J. Carroll), Addison-Wesley Press, Reading, MA, 2001; chapter 19, pp. 421-456].
- Oviatt, S.L., (2000) *Taming Speech Recognition Errors Within a Multimodal Interface*, in *Communications of the ACM*, Sept. 2000, 43 (9), 45-51 (special issue on "Conversational Interfaces").
- Oviatt, S.L. and Cohen, P.R. *Multimodal Interfaces That Process What Comes Naturally*. *Communications of the ACM*, Vol. 43, No. 3, March, 2000, pp. 45-53.
- Power, R. J. D. (1979). *The Organization of Purposeful Dialogues*. *Linguistics*, 17:107-152.
- Raisamo, R., (1999). *Multimodal Human-Computer Interaction: a constructive and empirical study*, Ph.D. dissertation, University of Tampere, Tampere(1999).
- Ruhleder, K., and Brigitte J. (2001). *Co-Constructing Non-Mutual Realities: Delay-Generated Trouble in Distributed Interaction*. *International Journal of Computer Supported Cooperative Work*, 10(1): 113-138

- Salisbury, M. W. (1990). Talk and draw: Bundling speech and graphics. *IEEE Computer*, pages 59--65, August 1990.
- Sacks, H., Schegloff, E. and Jefferson, G. (1974). A simplest systematic for the organisation of turn-taking for conversation. *Language*, 50.
- Saeed, J. (1997). *Semantics*. Blackwell Publishers.
- Samuel, K., Carberry, S. and Vijay-Shanker, K. (1998). Dialogue Act Tagging with Transformation-Based Learning. *Proceedings of CoLing 98*.
- Searle, J (1969). *Speech Acts*. Cambridge University Press.
- Searle, J. (1979): *Expression and meaning*. Cambridge University Press.
- Searle, J. and Vanderveken, D. (1985). *Foundations of Illocutionary Logic*. Cambridge University Press.
- Sinclair, J. and Coulthardt, R. (1975). *Towards an Analysis of Discourse: The English used by teachers and pupils*. Oxford University Press.
- Traum, D. (1999). *Speech Acts for Dialogue Agents*. In Wooldride, M. and Rao, A. (Eds.) *Foundations of Rational Agency*. Kluwer.
- Taddei, L., Costantini, E. and Lavie, A. (2002). The NESPOLE! Multimodal Interface for Cross-lingual Communication - Experience and Lessons Learned. *Proceedings of ICMI International Conference on Multimodal Interfaces*, Pittsburgh, PA, U.S.A., October 2002.
- Vanderveken, D. (1992): *On the Unification of Speech Act Theory and Formal Semantics*. In Cohen, Morgan and Pollack (eds.) *Intentions in Communication*. MIT Press.
- Waibel, A. Interactive translation of conversational speech. *Computer, IEEE Comput. Soc.* 29 (1996) H. 7 S. 41-48. 1996.

APPENDIXES

APPENDIX 1: Customer's Instructions (Experiment 1)

(customer, multi-modal condition)

1. We want to evaluate the system, not the test person.

The aim of the experiment is to evaluate differences in the performance between two versions of the system prototype. You are confronted with one of the two versions. We are not interested in differences among users. We are not evaluating your knowledge or capabilities. Please remember that the system is not perfect, and that if something is going wrong it is not your fault!

Please try to speak naturally and to concentrate on the task.

2. Read the Task.

- You will be pre-conditioned for the task on an extra sheet.
- Basically, you would like to get information about:
 - an appropriate town or village to stay
 - an appropriate hotel
- Please prepare a couple of questions you would like to ask.

3. How the system works

- You ask a question.
- The automatic speech recognition transforms your question into text.
- The text is processed into a paraphrase to prepare it for translation. (This is not a word-to-word translation. Remember, there are differences in languages such as sentence structure, different expressions and speaking styles!)
- The automatic translation generates an Italian output of your question.
- The Italian agent hears your question in Italian produced by a synthetic voice.

On the Italian side:

- The Italian agent answers your question in Italian.
- The answer is again transformed into text.
- The text is prepared for translation.
- You hear a synthetic voice with the English translation of the agent's answer.

4. Using the system

There will be three user windows on the screen:

a, Netmeeting window:

- **click the microphone button** when you want to talk.
- **un-click the microphone button** when you are finished.

Please, make sure that you first click the microphone box, when it is really clicked you start to speak, when you are done, stop talking and than un-click the box. (Otherwise your speech gets cut.)

b, Monitor window:

- You can read the translation of the agent's contribution (also produced by a synthetic voice)
- You can read what the "**system hears**" when you speak (speech to text recognition)
- You can read how the "**system understands**" your utterance (a paraphrase of what you said for translation).

As soon as you can see what the system understood, decide if it represents the main issues of what you have said.

Please remember the following issues:

The agent is supposed to be cooperative and tries to understand even if it is not an exact representation of what you said.

We would rather like to keep the dialogue going by using clarification questions, asking for repetition or telling the other side that you didn't understand.

Only, if you think it is a **completely wrong and irritating paraphrase** and there is **no change to understand anything of your utterance**, click the cancel button and try again.

(When the Italian side is not content with their contribution and, therefore, clicked their cancel button, the text "**ignore**" will blink in red letters.

You can click the "okay" button to confirm that you read it, or leave it, and it will disappear on its own in a moment.)

c, White board:

When the Italian side sends you some visible information, it will appear at the White-board. The agent is able to mark on the sent images using a colored pen. You will see these drawings.

Please, use the drawing functions, too, and refer to details on the images: this may help the agent to understand.

Do NOT scroll, zoom or open URLs at the White-board. The agent will do that for you.

5. Problems

- Every utterance is going through different steps. Every step can produce errors. Therefore, it may happen that the translation is wrong and you would have to repeat your question. You may also have to ask the agent to repeat an utterance.
- **Shorter sentences** are less likely to have errors.
- You should only ask for
 - hotel information,
 - location of towns, ski-areas, ice-skating possibilities and hotels,
 - transportation between hotels and ski-areas or ice-skating facilities.

Questions and comments concerning other topics, as well as street names, are not supported by the system.

- Make use of sentences such as:
Please repeat that.
I didn't understand.
- The system sends sound recordings to Italy and back. Sometimes it takes about 1 minute to send or receive the information.
Please, just be patient and wait.

If you already waited longer than a minute just try again.

APPENDIX 2: Customer's Task (Experiment 1)

Try to imagine being in the following situation.

It is the end of November. You are going to spend a holiday in Val di Fiemme with a friend. Val di Fiemme is a region in northern Italy where you can find several ski-areas, towns and villages.

You are planning to go during the second week of December.

You wish to go alpine-skiing (down-hill skiing) and ice-skating.

You would like to sleep in a three-star-hotel for 7 nights.

You want to have half board accommodation (bed-and-breakfast and dinner)

You are planning to go during the second week of December.

Your available budget is at about 200 000 Italian Lire per night for the hotel room (*this is about 90 US dollar*).

You want a double room.

You will reach Val di Fiemme by airplane and bus. You already know about flight connections and bus transfer to Val di Fiemme.

In Val di Fiemme, you plan to use public transportation.

Your task is to ask the APT agent for more information.

You have to choose

- **a town** where you want to stay - close to a ski-area and with an ice-skating facility.
- **a hotel** close to a bus stop or a ski-area. It should meet your budget and your demands.

*Please remember, you are calling an information center – **you can't make reservations there.***

Please write below a couple of questions you would like to ask the APT agent in order to complete your task.

APPENDIX 3: Descripton Cards for Agents (exp. 1)

localita'	area sciistica raggiungibile con bus	collegamenti con area sciistiche	pattinaggio	collegamenti con localita' con pattinaggio
CAVALESE	Alpe Cermis	20 km 35 min. bus un bus ogni ora	SI 15min x centro 5min x bus stop	/
PANCHIA'	Pampeago	15 km 25 min. bus un bus ogni ora	NO	un bus ogni ora per Cavalese 15 min.
località	alberghi	posiz. su mappa	navetta privata	da fermata bus
	Hotel Bellavista	n 14 centro	si	800 mt -15 min
CAVALESE	Hotel Astoria	n 2 centro-ds	no	500 mt - 10 min
	Hotel Lagorai	n 22 alto-sn	no	50 mt - 1 min
	Hotel Belvedere	n. 5 margine ds	si	800 mt -15 min
PANCHIA'	Hotel Cimon	n 3 centro ds	no	500 mt - 10 min
	Hotel Lucia	n 4 centro sn	no	50 mt - 1 min
località	alberghi	costo singola	costo doppia	numero telefono
	Hotel Bellavista	250.000	350.000	0462-832507
CAVALESE	Hotel Astoria	140.000	240.000	0462-838102
	Hotel Lagorai	110.000	210.000	0462-830125
	Hotel Belvedere	150.000	250.000	0462-798341
PANCHIA'	Hotel Cimon	120.000	220.000	0462-795697
	Hotel Lucia	100.000	200.000	0462-797729

APPENDIX 4: Agent's Instructions (experiment 1)

1. VALUTAZIONE DEL SISTEMA

Lo scopo dell'esperimento e' quello di valutare differenze di performance tra due versioni del prototipo del sistema: quella senza multimodalita' (chiamata speech -only) che permette lo scambio di mappe e pagine web, oltre all'input vocale, e quella con la multimodalita' (intesa come possibilita' di tracciare segni, linee, selezioni sulle mappe). Non valutiamo differenze di performance tra utenti, ne' le loro capacita' o conoscenze. Ricorda che il sistema non e' perfetto e che se qualcosa sembra non funzionare (o oggettivamente non funziona) non e' colpa tua.

Ti chiediamo di contribuire al buon funzionamento del sistema (e dell'esperimento) tramite l'osservazione di una serie limitata di "regole di comportamento".

2. USO DEL SISTEMA

Finestra di Netmeeting:

Ricordati di **accendere/spengere il microfono** sempre prima di parlare/appena terminato di parlare usando il mouse o la penna elettronica; compi questa operazione **con calma**, aspettando qualche attimo prima di parlare dopo averlo acceso e prima di spegnerlo dopo aver parlato (altrimenti la parte iniziale o finale della tua frase potrebbe venir tagliata);

se ti capita di accendere il microfono per errore, non spegnerlo subito: cerca di dire comunque qualcosa prima di spegnerlo. Qualsiasi cosa succeda in sala o sul monitor (rumori, messaggi, ...) **cerca di concludere sempre la tua frase.**

1 Finestra "Monitor"

La finestra "Monitor" in alto e' divisa in tre sezioni.

- puoi leggere la traduzione del contributo del cliente nella prima sezione in alto;
- puoi leggere cosa il sistema "ascolta" quando tu parli (riconoscimento vocale) nella sezione centrale;
- puoi leggere cosa il sistema "comprende" della tua frase nella sezione piu' in basso (una parafrasi per la traduzione di quello che hai detto).

Utilizza la stringa della parafrasi del tuo messaggio (system understands) per capire se il sistema ha "compreso" il nucleo del tuo messaggio. Se pensi che la parafrasi sia completamente sbagliata e/o che non ci sia la possibilita' di comprendere qualcosa della tua frase originale, puoi interrompere il processo di traduzione cliccando sull'apposito bottone (cancel translatio) e ripetere la frase. Se ripetendo con parole simili la parafrasi continua a non corrispondere in alcun modo al tuo messaggio, prova a cambiare l'espressione che hai usato con una dal significato simile.

Quando il cliente non e' soddisfatto della propria parafrasi e clicca "cancel translation", sulla tua finestra Monitor compare il messaggio rossa lampeggiante IGNORE; puoi eliminare la scritta cliccando su OK, o lasciare che scompaia da sola. La comparsa del messaggio indica che il cliente sta provando a riformulare la frase; in questo modo tu sai che qualche altro messaggio sta arrivando (la sintesi vocale del messaggio del cliente

puo' arrivare anche se egli l'ha "cancellato": potresti percio' ricevere una sequenza di messaggi "senza senso" prima che arrivi un messaggio "comprensibile"

2 Whiteboard

Attraverso la whiteboard tu puoi inviare mappe e pagine web. Il risultato di ogni tua operazione viene trasmesso al cliente.

Nella **condizione speech-only** le operazioni permesse sono il caricamento dei mappe e pagine web, la pulizia della whiteboard e lo scroll (per visualizzare parti della mappa non visibili).

Nella **condizione con multimodalita'** puoi inoltre "scrivere" sulla mappa utilizzando una delle funzioni grafiche disponibili (penna, linea, selezione ellittica, selezione rettangolare) e uno dei colori disponibili. Il default e' nessuna funzione grafica selezionata e colore nero.

Le **mappe** si caricano cliccando sull'apposita icona o selezionando la voce dal **menu "file"** della whiteboard; esse si trovano nella **cartella Mappe, sul disco C** del computer; quando carichi una mappa, essa viene inviata anche al cliente, che la puo' ricevere dopo un minuto circa.

Prima di caricare una mappa, ricordati di "**pulire la whiteboard**" selezionando la funzione "new" dal menu "file" della whiteboard o cliccando sull'apposita icona.

Per i dialoghi con multimodalita' e' necessario **salvare il contenuto della whiteboard prima di pulirla**. La voce "salva come" si trova nel menu "file"

Le **pagine web** si inviano selezionando la funzione "**open URL**" dal menu "**tools**" della whiteboard e vengono visualizzate su una finestra del browser, che si apre automaticamente.

Ricordati di **chiedere sempre al cliente se ha ricevuto** la mappa o la pagine web inviata e aspetta (con pazienza) la sua risposta.

3 Gestione dei turni

Tu puoi ascoltare solo la voce sintetica che "legge" la traduzione dell'intervento del cliente, non la voce originale del cliente. Il tempo che intercorre tra quando il cliente parla e quando tu ascolti la traduzione puo' essere piuttosto lungo. Dopo che hai parlato se c'e' un lungo silenzio non hai modo di sapere se il cliente sta parlando o sta aspettando qualcosa da parte tua, a meno che non egli non abbia "cancellato" il proprio intervento per ripetere la frase (comparsa messaggio IGNORE).

In caso di ritardo nella risposta da parte del cliente il suggerimento e' di aspettare circa un minuto, e poi riprovare a parlare. Lo stesso consiglio viene dato al cliente. Se hai l'impressione che ci sia una sovrapposizione di interventi fra te e il cliente, prova ad aspettare un po' di piu'.

Normalmente comunque una situazione di sovrapposizione di parli si risolve nell'arco di qualche turno.

3. ALCUNE ISTRUZIONI AGGIUNTIVE DA SEGUIRE SCRUPOLOSAMENTE:

Inizio del dialogo: inizi tu il dialogo, una volta che la connessione e' pronta e tu ti senti pronto/a, con una **frase di apertura** di presentsaion (APT del Trentino, buongiorno; Trentino Informazioni, buongiorno).

Nomi di alberghi: far sempre precedere ai nomi degli alberghi la parola hotel o albergo

Prezzi: far preceder o seguire sempre all'indicaizione di un prezzo la parola **lire**

Non fare riiferimenti a nomi di vie e piazze: non fanno parte del repertorio del sistema

Cerca di **evitare domande contenenti la parola dove:** produrranno certamente degli errori

Il cliente non ha chiaro cos'e' la val di Fiemme, come e' fatta, quali localita' ci sono. Una delle prime cose che devi fare (appena chiede qualcosa sulla val di Fiemme) e' di **inviare subito la mappa delle ski area.**

Come rispondere a:

- richieste generiche (o poco chiare) su una localita': inviare pagina web della localita'
- richieste generiche su impianti sciistici: inviare mappa Ski-area
- richieste generiche su alberghi: chiedere informazioni piu' precise sull'albergo (es: categoria desiderata)
- richieste specifiche di alberghi (se incluse nelle informazioni a tua disposizione): suggerire il primo albergo della lista (se non si e' ancora scelta la localita', suggerire la localita' seguendo l'ordine indicato, a partire dalla prima)
- richieste specifiche su alberghi (se non incluse nelle informazioni a tua disposizione): offrire il numero di telefono dell'albergo. In particolare questo vale per richieste su disponibilita' di stanze e per prenotazioni (tu non sai ma soprattutto non puoi prenotare!)

IMPORTANTE!

A. Non dare informazioni non richieste esplicitamente:

esempi: se ti chiede il tipo di servizio di un albergo non devi aggiungere il prezzo o il numero di telefono; se la situazione e' di stallo, non aggiungere informazioni per sbloccarla (se possibile): cerca piuttosto di stimolare altre domande con frasi del tipo: ha bisogno di altro? Di che informazioni in particolare ha bisogno?

B. Non dare informazioni che non siano contenute nella scheda delle localita' e degli alberghi: in caso ci siano domande su apetti riguardo ai quali non avete informazioni, rispondete che non potete rispondere.

4. PROBLEMI

Ogni frase passa attraverso diverse fasi di elaborazione; ogni fase potrebbe produrre errori, pertanto potrebbe succedere che **la traduzione e' sbagliata** e che tu debba **riperere la frase**. Potresti anche aver bisogno di chiedere al cliente di ripertere una sua frase.

Fai uso di frasi come:
puo' ripetere per favore?
Non ho capito
...

E' meno probabile che frasi corte producano errori rispetto a frasi lunghe.

Cerca di essere il piu' collaborativo/a possibile. Il cliente vede il sistema per la prima volta e si aspetta che tu prenda in mano la situazione nel caso di una situazione di stallo. Cerca di intuire quello che ha detto il cliente nel caso in cui il messaggio arrivato non sia chiaro (raramente e' chiarissimo); ovviamente se non ti sembra di capire nulla, non tirare a indovinare e chiedi di ripetere.

APPENDIX 5: Enrollment Form

The experiment in which we ask you to participate is part of the Nespole! project. The Nespole! partners are research institutes and technology providers from Europe and USA. They are: ITC-irst (Italy), UJF (France), UKA (Germany), CMU (USA), APT (Italy), Aethra S.r.l. (Italy). The project manager is Gianni Lazzari from ITC-irst. The project is aimed at building and evaluating three prototypes of a multimedial, multimodal and multilingual video-call-centre (web site: <http://nespole.itc.it>).

The first prototype works in a tourist scenario. Through this system a potential client can visit the web site of APT- Trentino (APT-Trentino is an Italian tourist board office, located in a particular region called Trentino) and, clicking on a button, he/she can open a videoconferencing session with a tourist agent to ask for more information.

The main goal of the experiment is to compare the efficiency and usability of two versions of the system. In both versions the participant acting as a client and the agent can see each other and share web pages and maps; in one version they could in addition draw free strokes on maps loaded on a particular tablet. Each participant will use only one of the two versions of the system during the experimental session.

Taking part in the experiment will require you to:

- fill in a questionnaire on your computer literacy and web expertise (we need our participants group to be homogenous in terms of computer literacy and web expertise);
- take part in one experimental session. We will make you an appointment with an APT agent. You will be connected to each other through Nespole! system; each of you will speak in his/her mother language: a machine interpreter will translate the speech into the other language. You will ask the agent for information you need to take a certain decision regarding your holidays. You will receive detailed information on system functionalities and on your task. You will get in touch with the system through a short training task before the main task. We will not measure your skills or knowledge, but the capability of the system to support the communication in different situations: the situations can differ among them in terms of implied languages and interaction modalities.
- to fill in a questionnaire about your impressions concerning the system and your experience with the system, once the interaction will be ended.

We inform you that we guarantee your anonymity. Your personal data will be treated with the maximum reserve from the researchers involved in the experiment and will not be disseminated without your written authorization in any case. From now on you will be identified through a code. We inform you that we will record video and audio stream of your experimental session; the records will be disseminated only among Nespole! partners and exclusively for experimental needs (in particular for data analysis). You may authorize us to use records regarding you in communications to scientific community (e.g. conferences).

Now choose and write here a code of 4 letters. You will have to write the same code on the questionnaire on computer literacy and web expertise.

— — — —

Full name: _____

E-mail address (*): _____

Telephone number (*): _____

Do you authorize Nespole! consortium to use records regarding you in communications to scientific community?

(thick the correct answer)

YES NO

Signature (in case of authorization given) _____

(*) We will use your e-mail address and phone number exclusively to contact you for the experimental session

APPENDIX 6: Questionnaire on computer literacy and web expertise

Please, fill in the following blanks with your personal data (only if your mother language is American English).

Age: _____

Sex: _____

Mother language: _____

Educational qualifications: _____

Job: _____

Field of study (only for students): _____

Identification code: _____

The following questionnaire is aimed at measuring computer literacy and web expertise in the group of participants. Please find below a list of questions regarding your personal experience with computers and with the web. Think about the activities you have been carrying out with the computer from the beginning of this year until now, and answer to the questions giving an average value.

We are interested in your answers because we have to make sure that the group of participants is homogeneous in terms of computer literacy and web expertise. Your contribution is greatly appreciated.

For each question enter a number corresponding to the correct answer using the following evaluation scale:

1 = never used

2 = less than 3 hours a month

3 = from 3 to 10 hours a month

4 = from 11 to 30 hours a month

5 = more than 30 hours a month

1. How many hours a month have you used a computer to play videogames? ____

2. How many hours a month have you used a computer to write and edit text? ____

3. How many hours a month have you used a computer to program? ____

4. How many hours a month have you used a computer to carry out data analysis? ____
5. How many hours a month have you used a tablet and a electronic stylus as device (instead of the mouse)? ____
6. How many hours a month have you used a computer for your e-mail? ____
7. How many hours a month have you used a computer to connect to a chat-line? ____
8. How many hours a month have you used Microsoft Netmeeting (or other videoconference applications)? ____
9. How many hours a month have you used a computer to carry out searches on the Internet? ____
10. How many hours a month have you used the Internet to search travel information (train/air schedules and links)? ____
11. How many hours a month have you used the Internet to book train/air tickets? ____
12. How many hours a month have you used the Internet to book hotel rooms? ____

Answer to the last question using the following evaluation scale:

1 = less than 1 hour a day

2 = from 1 to 2 hours a day

3 = from 3 to 5 hours a day

4 = from 6 to 8 hours a day

5 = more than 8 hours a day

13. How many hours a day have you spent (on average) on the computer? ____

APPENDIX 7: List of Recorded Files

mod	type	EG		ITA				Extension
rec	audio	cus-EG	a.syn-EG	age-ita	(*)			.wav
rec	commswitch	EG		ITA				_hlt.txt
rec	video	cus		age				(not digital.)
rec	gesture_data	---		age				_age/_clidata
rec	maps	---		age				.gst
prot	speaker_data	cus		---				.spr
prot	recording_prot	local		---				.rpr
an	transcription	cus	a.syn	age	c.syn			.trl
an	time_stamps	cus	a.syn	age	c.syn			.mar
an	gest_annot	cli		age				.glb
an	transcr. alignm.	---		age	c.syn	cli	a.syn	.mix
an	coding tables	---		age	c.syn	cli	a.syn	_sco.xls

EG = English or German;

age = agent, cus = customer;

a./c.syn = synthetic translated output of agent or customer.

Recorded files (rec):

- 1 audio stereo file, 22 (16) kHz, 16 bit, microphone recording of customer plus synthesized translation of agent's turns, recorded at the customers side (CMU, UKA), audio wav file;
- 1 audio mono file, 22 (16) kHz, 16 bit, microphone recording of agent, recorded at lrst, audio wav file;
- 2 CommSwitch files from each location;
- 2 video recordings (customer and agent side); the videos have not been digitalized;
- 2 files containing the white board gestures of the multi-modal task, (agedata and clidata), text file;
- Several bitmaps of maps with drawn gestures

Protocol files (prot):

- 1 speaker files containing information about the customers: identification short-cut, sex, age, comments, information from enrollment form and computer literacy questionnaire, text file format

- **1** recording protocol file client's side, containing information about the recording environment and experiment setting (recording date, sampling frequency, microphone type, experimental condition), text file format

Transcription and annotation files (an):

- **1** transcription file of the customer side's recording, containing customer's and synthetic agents dialogue contributions (turns), text file format;
- **1** transcription file of the agent side's recording containing agent's and synthetic customer's dialogue contributions (turns), text file format; (*) the synthetic customers' contributions are taken from the hlt.txt files.
- **2** marker files: containing the time stamps according to the transcribed turns, text file format;
- **1** text files: manual annotation of drawing gestures, text file format;
- **1** alignment file: containing all original and synthesized turns from agents and customers;
- **1** coding file: containing coding for turns and discussed topics, excel file.

APPENDIX 8: Labelling of Spontaneous Phenomena

A. LABELLED SPONTANEOUS EVENTS: CLUSTERS

The following spontaneous events were labelled in NESPOLE! experiment data transcriptions:

CLUSTERS	SPONTANEOUS EVENTS	LABELS
WORD INTERRUPTIONS	Aborted_Articulation	wor=
	Articulatory_Interruption	in_<P>_terruption
INCOMPREHENSIBLE UTTERANCES	Hardly_Identifiable	word%
	Unidentifiable	<%>
EMPTY PAUSES	Empty_Pause	<P>
	Breathing	
FILLED PAUSES	Filled_Pauses	<hes> <uh> <uhm> <hm>
HUMAN NOISES	Human_Noise	<Noise> <Laugh> <Cough> <Throat> <Smack> <Swallow>
A-GRAMMATICAL PHRASES	False_Start	-/I want/- can you please ...
	Repetition_or_Correction	+/the green/+ the green apple, +/the green/+ the red apple
TECHNICAL BREAKS	Technical_Interruption	<T_>ord <*T> wo<_T>
TURN BREAKS	Turn_Break	<*T>t

(please look at http://www.is.cs.cmu.edu/trl_conventions/ for detailed descriptions).

B. LABELLED SPONTANEOUS EVENTS: DESCRIPTION

Word Interruptions

- **Aborted_Articulation:**

= word break:

Broken words have a “=” at the position of the break. e.g.

good mor= Mond= , Donners= perfor=

- **Articulatory_Interruption**

_ word interruption:

Interrupted words (by any other audible event such as e.g. a filled pause)

Pre_ <uh> dazzo, buon_ _giorno, Val-di_ +/Fiu=/+ _-Fiemme

Incomprehensible Utterances

- **Hardly_Identifiable**

word%

Often words can hardly be identified because of audio quality or sloppy speech. As long as it is clear what the word should have been, a “%” is added at the end of this word (it indicates that even the human transcriber had problems, therefore, speech recognition might also not doing well)

I would like to say% something% .

- **Unidentifiable**

<%>

Word or sound which is completely non-identifiable

I would like to <%> something .

Empty pauses

- **Empty_Pause**

<P> silence:

Speaker stops speaking for a while. Nothing can be heard.

let's say <P> Tuesday .

- **Breathing**

a pause filled with a breathing sound (either exhalation or inhalation)

well , I was

Filled pauses

- **Filled_Pauses**

<hes> hesitation or filled pause:

Since there is so much variety at the occurrence of filled pauses (even differences in languages), it is easier for transcription to use categories:

<uh> just a vowel (e.g. /ee/ /ou/ /aeh/ etc)

<uhm> a vowel/nasal combination (e.g. aehm, aahm, ohm, annn)

<hm> just a nasal (but without meaning. Not to get confused with agreeing mhm or negating m'm). e.g. mmmm, hmmm, nnnn

<hes> anything else e.g. pffff, sssss, schhhh

well , <uhm> what did you say ?

<hm> I don't know .

Human noises

▪ Human_Noise

<Noise> articulatory noise:

Some human noises which might be kept within the transcript

<Smack> lip smack) (Maybe too detailed)

<Swallow> swallowing) (Maybe too detailed)

<Throat> clear one's throat

<Cough> coughing

<Laugh> laughing

<Noise> any human noise, not identifiable

<Laugh> that is so funny . <Throat>

A-grammatical phrases

▪ False_Start

-/../- Interruption and start of a new thought:

The interrupted part is between -/../- brackets.

-/can you please/- what did you say ?

-/I would li=/- oh , something happened .

▪ Repetition_or_Correction

+/../+ Repeating or correcting a word or more:

The repeated or corrected part stands between +/../+ brackets

I will have +/the green/+ <uh> the green apple

I will have +/the green/+ <uh> the red apple

Technical breaks

▪ Technical_Interruption

<T_>ord

wor<_T>

<*T> technical word break:

Technical problems can cut words in the audio recording.

Since this is not an articulatory event produced by the speaker, it should be distinguished by using a different sign for technical word break. The audio signal shows only a straight line.

Initial silence:

<T_>rning <T_>iorno

Final silence:

morn<_T> buongi<_T>

longer silence:

<*T> technical silence period in case of longer periods of technical disturbances

Turn breaks

- **Turn_break**

<*T>t

A turn break occurs when a speaker contribution just stops in the middle of a sentence or obviously unfinished. This happens often when another speaker interrupts, a technical problem occurred. In these situations neither period nor question mark can end the turn.

e001_1_0004_ITL_00: could I plea= <*T>t

e001_1_0005_ABC_00: I have to interrupt you .

APPENDIX 9: Output of the TransEdit Annotation Tool

The files are explained below:

- **file.trl**: the actual transcription file in ascii text
- **file.trl.set**: this small file contains the header data for the transcript file
- **file.mar**: this is the 'marker' file which contains the time marks for the segmentation of the audio

Turns are identified in the transcription (.trl) and marker (.mar) files in a consistent way. To build the turn identifiers the following information are used: file name, channel number, turn number, speaker identifier and time stamp. Speakers are identified with a six-letter speaker ID which is automatically generated, as in the following examples.

A. Turn identifier as it would appear in the *transcription file*:

i114j_1_0001_RIFSTP_00: A P T del Trentino buongiorno ?

i114j: file name
1: channel number
0001: turn number
RIFSTP: speaker ID
00: constant number generated by TransEdit tool

B. Turn identifier as it would appear in the *marker file*:

208880 252894 RIFSTP_0001_1
208880: begin time stamp
252894: end time stamp
RIFSTP: speaker ID
0001: turn number
1: channel number

This example shows the use of the speaker ID to associate a spoken turn in the transcription file (**A**) with the correct time stamps in the marker file (**B**). The transcription's turn identifier provides the file name, channel number, turn number and speaker ID. The turn identifier of the marker file provides the beginning and end time stamps, speaker ID, turn number and channel number.

Here is a short example of transcription of two turns:

e109_1_0018_ALDYNT_00: yeah , +/it is/+ <uh> it is okay . <uh> the accommodation% is% <uhm> camping , but <uh> it is excursion . for example , there are <uh> in the price <uh> <%> <*T> the package <uh> it is included <uh> two admission% to the swimming pool in Predazzo . +/the/+ the camping site is in Predazzo . <%> Valle-di-Fiemme and then admic_T> <*T> <%> <uh> at the center at Cavalese% , a mountain bike excursion and visit <uh> at the local musc_T> <*T> excursion in the mountc_T> an Alpine guide , and one discount card% at the campsite shops .

e109_2_0019_OEIZJ_00: okay , I'll be coming by camper . <uhm> how should I reach the place +/by/+ by camper . is that possible ?

APPENDIX 10: Conventions for Annotation of Gestures

(Erica Costantini, September 2001)

Gestures have to be annotated on a copy of the transcription files. The files containing annotation of gestures will have the extension “.glb” (gesture labels).

The transcribers at Irst have to annotate all gestures performed by the agents using the video recordings and (only for drawings performed during the MM condition dialogues) the saved .gst files (“photos” of maps plus the performed drawings).

The annotation of the clients gestures is done “from the agent point of view”: the transcriber could insert information on received gestures watching at the videos recorded in Irst. If she needs, she could have a look to the clidata.txt files (which contain information concerning clients gestures).

After each turn the transcribers take note of the gestures composed before/during the turn following the special rule for global comments in the .trl files.

Global comments follow a turn. After the turn a new line starts with a semicolon and a blank. Each new line needed for the comment is starting with semicolon and a blank. Between the last line of the comment and the next turn there is an empty line. Using this format would allow us to filter the comments if they are not needed.

Example:

```
e726_2_0001_FRANK_00: good day .  
; gesture=G1_A_before  
; type=loading: skiarea  
  
e726_2_0002_FRANK_00: I would ...
```

Gestures annotations include the following information:

- first line: gesture identification
- second line: gesture description
- third line: gesture goal (only for free-hand strokes, pointing and writing words)

4 FIRST LINE: GESTURE IDENTIFICATION

1. PROGRESSIVE NUMBER:

gesture=G(n)

2. USER:

agent > A
client > C

3. TIME

The gesture is performed JUST BEFORE or DURING or JUST AFTER the speech turn:

just before > before
during > during
just after > after)

4. EXAMPLE:

gesture=G1_A_before

5 SECOND LINE: GESTURE DESCRIPTION

1. TYPE:

free-hand strokes > freehand;
elliptical/rectangular selection > selection;
loading a map > loading
running a browser > running
scrolling > scroll
zooming > zoom (never used).

I would like to add:

clearing the WB > clear
closing the web page > close

2. DESCRIPTION

Description features are written between brackets.

FREEHAND;

Type:

circling an area > circle
line > line
arrow > arrow

writing letters/words > letters

other > other

Color:

black > black

red > red

blue > blue

green > green

yellow > yellow

SELECTION

Shape:

elliptical > el

rectangular > re

Color:

black > black

red > red

blue > blue

green > green

yellow > yellow

LOADING:

Name of the map. Three maps have been used: cavalese, panchia, skiarea.

RUNNING

Name of the web page. Two web pages have been used: APT cavalese, APT panchia

SCROLL:

Number of scrolls:

single scroll > single

one sequence of scrolls > sequence

more than one scroll not in a single sequence > multiple

Type of scroll:

vertical > vertical

orizental > orizental

both vertical and orizental (in case of sequence or multiple) > mix

ZOOM (NEVER USED).

3. CONTEXT

Name of the map. Three maps have been used:

Map of Cavalese > cavalese;

Map of Panchià > panchia;

Map of Val di Fiemme with skiareas > skiarea.

4. EXAMPLES

a. type=freehand (circle_red); context: cavalese

b. type=scroll (sequence_mix); context: skiarea

c. type=selection (el_blue); context: panchia

6 THIRD LINE: GESTURE GOAL (only for free-hand strokes and pointing)

1. SELECTION OF AN AREA (circle, square):

goal=selection: class (item)

class1 = town: cavalese, panchia, other

class2 = ski areas: pampeago, cermis, alpe lusia, latemar, belvedere

class3 = hotels: bellavista, astoria, lagorai, cimon, lucia, belvedere

class4 = bus stop

class5 = skating rink

class6 = other

e.g. goal=selection: town (cavalese)

2. POINTING OF AN AREA (arrow):

>goal=pointing: class (item)

See above for the list of classes and items.

e.g. goal=pointing: hotel (bellavista)

3. CONNECTION BETWEEN TWO AREAS:

> goal=connect class I (item I) to class II (item II)

See above for the list of classes and items.

e.g. goal=connect: town (cavalese) to skiarea (pampeago)

4. WRITING NAMES:

> goal=word: (word)

class1 = town name: cavalese, panchia, other

class2 = ski area name: pampeago, cermis, lusia, latemar, belvedere

class3 = hotel name: bellavista, astoria, lagorai, cimon, lucia, belvedere

class4 = other

e.g. goal=word: hotel lucia

7 COMPLETE EXAMPLE

The case is the following:
the agent loads the ski area map;

the agent uses free-hand strokes;
the agent first circles the town Cavalese and then circles the skiarea Alpe Cermis, and
then draws a line to connect the two;
all gestures are performed before he starts speaking.

e009yi_1_0008_mari_00: If you are in Cavalese, you can easily reach the ski area Alpe Cermis. I show you them on the map of Val di Fiemme?

```
; gesture=G1_A_before
; type=loading: skiarea
;
; gesture=G2_A_before
; type=freehand (circle_red); context=skiarea
; goal=selection: town (cavalese)
;
; gesture=G3_A_before
; type=freehand (circle_blue); context=skiarea
; goal=selection: skiarea (cermis)
;
; gesture=G4_A_before
; type=freehand (line_green); context=skiarea
; goal=connect: town (cavalese) to skiarea (cermis)
```

e009yi_1_0008_mari_00: can you see them?

APPENDIX 11: S.U.S. (System usability Scale)

We would like to know what you think about the Nespole! Project system and your experience with it.

Please let us know your opinion by indicating the level of your agreement/disagreement with each of the following statements. You can answer by choosing a number from 0 to 4, with 0 meaning totally disagreement and 4 meaning totally agreement and marking the correspondent box. If you feel that you cannot respond to a particular answer, you should mark the central point of the scale.

1. I think that I would like to use this system frequently.

0	1	2	3	4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

2. I found the system unnecessarily complex.

0	1	2	3	4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

3. I thought the system was easy to use.

0	1	2	3	4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

4. I think that I would need the support of a technical person to be able to use this system.

0	1	2	3	4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

5. I found the various functions in this system were well integrated.

0	1	2	3	4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

6. I thought there was too much inconsistency in the system.

0	1	2	3	4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

7. I would imagine that most people would learn to use this system very quickly.

0	1	2	3	4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

8. I found the system very cumbersome to use.

0	1	2	3	4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

9. I felt very confident using the system.

0	1	2	3	4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

10. I needed to learn a lot of things before I could get going with this system.

0	1	2	3	4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

APPENDIX 12: Descripton Cards for Agents (exp. 2)

DESCRIPTION OF TOWNS

localita'	area sciistica di riferimento	collegamenti con area sciistica	pattinaggio	collegamenti con pattinaggio
CAVALESE	Alpe Cermis	10 km (20 min. bus) uno skibus ogni ora	SI (a piedi 20 min dal centro, 5 da bus stop)	/
PANCHIA'	Latemar	15 km (25 min. bus) un skibus ogni ora	NO	1 bus ogni ora per Cavalese (15min)
PREDAZZO	Alpe Lusia	5 km (10 min. bus) un skibus ogni ora	SI (a piedi 10 min dal centro, 5 da bus stop)	/

DESCRIPTION OF THE HOTELS

(all of them are three-stars hotels)

località	alberghi	posiz. su mappa	da fermata bus	numero telefono
	Hotel Bellavista	n 30 centro	800 mt -15 min	0039-0462-832507
CAVALESE	Hotel Belvedere	n 21 centro-sn	500 mt - 10 min	0039-0462-838102
	Hotel Lagorai	n 6 alto-sn	50 mt - 1 min	0039-0462-830125
	Hotel Belvedere	n. 5 margine ds	800 mt -15 min	0039-0462-798341
PANCHIA'	Hotel Cimon	n 3 centro ds	500 mt - 10 min	0039-0462-795697
	Hotel Lucia	n 4 centro sn	50 mt - 1 min	0039-0462-797729
	Hotel Astoria	n. 2 centro	400 mt - 5 min	0039-0462-502531
PREDAZZO	Hotel Montanara	n. 3 centro	100 mt - 1 min	0039-0462-504522
	Hotel Excelsior	n. 8 centro	10 mt - 1 min	0039-0462-503123

DESCRIPTION OF PACKAGES

	sistemazione	sport	altro	costo x persona, x 1 settim. in doppia
Pacchetto 1	albergo (pensione completa)	skipass; 4 lezioni sci discesa; 1 ingresso pattinaggio Cavalese o Predazzo; ingresso libero piscina Cavalese o Predazzo. 1 ingr. centro benessere Cavalese	cena tipica trentina; visita caseificio con degustazione formaggi; Ingr. museo Cavalese.	€ 550
Pacchetto 2	albergo mezza pensione	skipass; 1 ingresso pattinaggio Cavalese o Predazzo; ingresso libero piscina Cavalese o Predazzo.	2 cene ristoranti locali; visita caseificio con degustazione formaggi; attivit� per bambini.	€ 350
Pacchetto 3	albergo mezza pensione	skipass; 4 lezioni sci discesa; 1 ingresso pattinaggio Cavalese o Predazzo; ingresso libero piscina Cavalese o Predazzo.	2 cene ristoranti locali; visita caseificio con degustazione formaggi; Ingr. museo Cavalese.	€ 400

APPENDIX 13: Customer's Instructions (Experiment 2)

"We want to evaluate the system, not the test person."

The aim of the experiment is to evaluate:

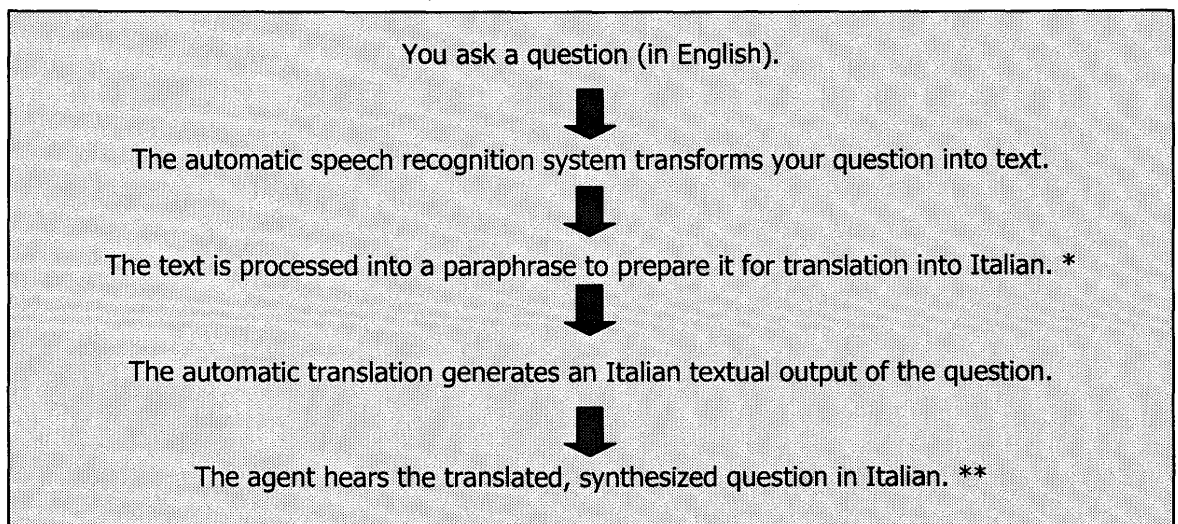
- The role of multimodality in the communication of the NESPOLE! System
- The effectiveness of the system
- The usability of the system

We are not interested in differences among users. We are not evaluating your knowledge or capabilities. Please remember that the system is not perfect and that if something is going wrong **it is not your fault!**

The Task

- You will be pre-conditioned for the task on another sheet.
 - Basically, you would like to get information about:
 - an all included tourist package for a winter vacation in Italy;
 - an appropriate town or village to stay in;
 - an appropriate hotel.
-

HOW THE SYSTEM WORKS



* This translation, based on the Interchange Format approach, is not a word-to-word translation.

** The same steps repeat vice-versa when the agent speaks (in Italian).

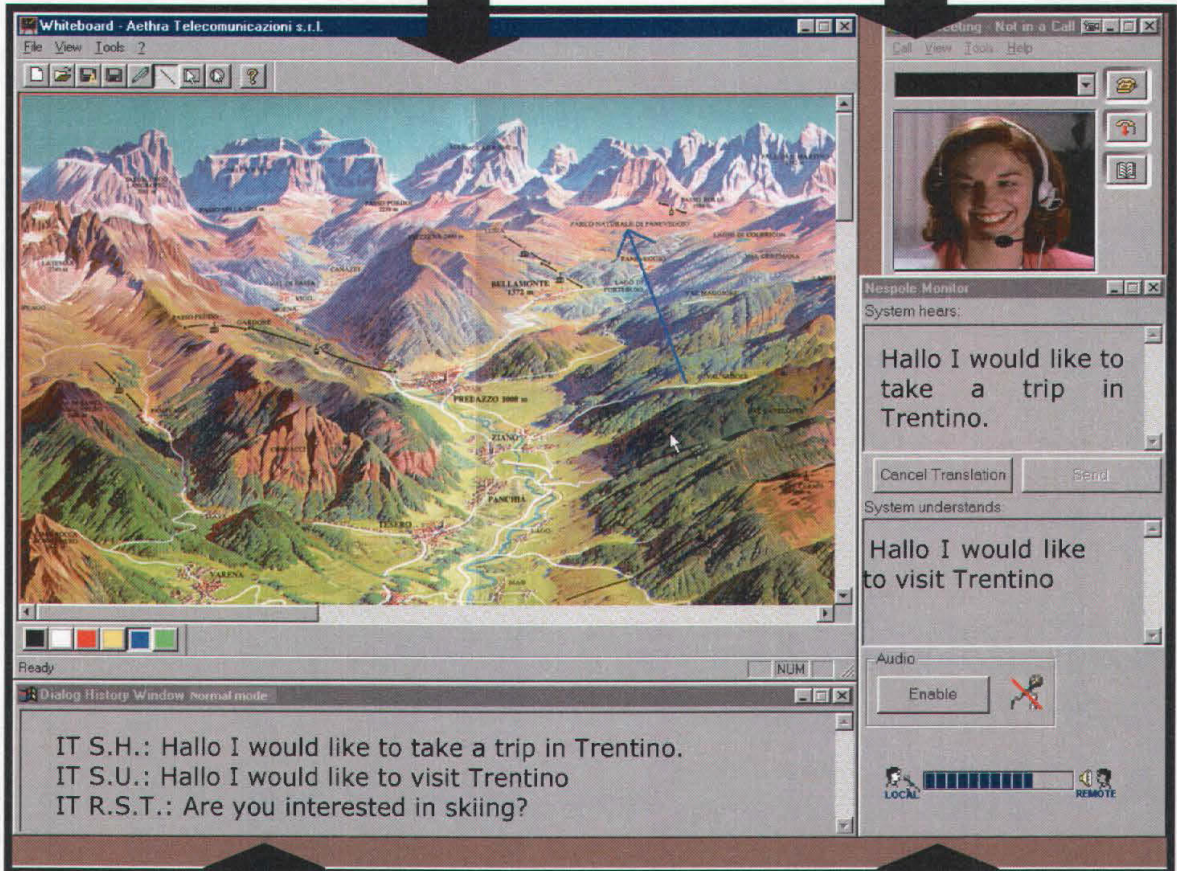
THE NESPOLE! INTERFACE

WHITEBOARD WINDOW

In this window you can see the maps loaded by the Italian agent. You (as well as the agent) can use some drawing functions to mark different colors and shapes on the images.

NETMEETING WINDOW

Here you can see the video of the agent on the other side.



DIALOGUE HISTORY WINDOW

This window visualizes in a progressive way the following phases of the translation process:

- The text of your recognised speech field (also available in the "System hears" field);
- The paraphrase of the recognised speech field, also available in the "System understands" field;
- The text of the remote speech translation (what you hear from the synthetic voice).

MONITOR WINDOW

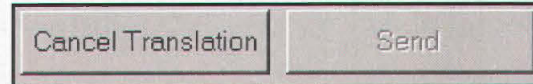
In this window you can see:

- The text of your recognised speech field ("System hears"), which is editable;
- The cancel translation button;
- The paraphrase of the recognised speech field ("System understands");
- The microphone button;
- A progress bar indicating the status of the speech transferring process (one bar for your speech and one for the agent speech).

USING THE SYSTEM

NESPOLE MONITOR WINDOW

"Cancel Translation" button



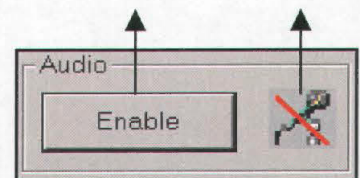
- Immediately after having spoken, look at the "system hears" and "system understands" field fields and read the text. If the meaning of the text is very different from that of your speech, you can **press the "Cancel Translation" button** and repeat your sentence (try using different words).
- Only if there is no way for you to have the system recognize some of your words, you can **edit** what you said **manually** in the "system hears" field rather than repeating it (after having clicked the cancel button), and click the "send" button.
- When you press the cancel button, the text **IGNORE** will blink in red letters on the agent screen to inform her that you are repeating (or editing) the sentence and that therefore a next synthesized speech is arriving (it will appear on your screen when the agent is not content with her contribution and tries again).

Please, click the cancel button and try again only if you think it is a completely wrong and irritating paraphrase and there is no chance to understand anything of what you said, and use the manual editing of the recognized text as rarely as possible.

Microphone Button

microphone button status icon

Please, make sure that you first click the microphone box. When it is really clicked you start to speak. When you are done, stop talking and then un-click the box. (Otherwise your speech gets cut.) The status says you if the microphone is switched on or off.



WHITE BOARD WINDOW

When the Italian agent sends you some visible information, it will appear on the White-Board. The agent is able to mark on the sent images using a colored pen. You will see these drawings.

Please, use the drawing functions, too, and refer to details on the images: this may help the agent to understand.

Do NOT scroll, zoom or open URLs on the White-Board. The agent will do that for you.

DIALOGUE HISTORY WINDOW

The main objective of this window is to help you in remembering what happened. In particular, use this window to check for the content of the synthesized audio in case

the audio quality is poor and you did not understand it well, or in case you forgot what the agent said.

BEFORE THE DATA COLLECTION SESSION ...

- Some **drawing functions** are available on the White-Board:
 - you can freehand draw by selecting the pen icon
 - you can draw circles or rectangles by selecting their icons
 - you can choose a color (red, blue, yellow, back, white and green)**Try to use those drawing functions** until you feel familiar with them.
- **Read the task description and the script very carefully**
- **Please prepare a couple of questions to ask**
- As soon as you are ready **call the APT Operator** by clicking on the specific button in the NetMeeting window (the button with the yellow phone icon).
- When the APT operator answers your call, **start the conversation** according to the task (please try to **speak naturally** and to **concentrate on the task**).

ERRORS

Every utterance is going through different steps. Every step can produce errors. Therefore, it may happen that the translation is wrong and you will have to repeat your question. You may also have to ask the agent to repeat an utterance (make use of sentences such as: please repeat that, I didn't understand)

- **Shorter sentences** are less likely to have errors.
- You should only ask for
 - An all inclusive package (price, ski lessons, ...)
 - Hotel information,
 - Location of towns, ski-areas, ice-skating facilities and hotels,
 - Transportation between hotels and ski-areas or ice-skating facilities.

Questions and comments concerning other topics, as well as street names, are not supported by the system.

TIME DELAYS

The system sends sound recordings to Italy and back. Sometimes it takes some seconds to send or receive the information. **Please, just be patient and wait.** If you wait longer than a minute, please try again.

In particular, it takes a couple of seconds for the progressive bar and the ignore message to appear on your screen when the agent is speaking (or cancelling).

Please wait a couple of seconds before speaking when you receive the synthesized speech to be sure that the agent has actually finished.

APPENDIX 14: customers' task (experiment 2)

Try to imagine being in the following situation.

This winter you are going to spend a holiday in **Val di Fiemme** with a friend.

Val di Fiemme is a region in northern Italy where you can find several ski-areas, towns and villages.

- You are interested in the following activities:
 - **ski lessons (particularly down-hill skiing)**
 - **skating**
- You prefer to stay in a **three stars hotel**.
- You want a **double room**.
- You are looking for an **"all included" package** for **one week**, including at least:
 - **Half board accommodation** (bed-and-breakfast and dinner);
 - **Ski lessons and ski-lift;**
- Your available **budget** is at about **€ 400** (per person) for a one-week "all included" package.
- You will reach Val di Fiemme by airplane and bus. You already know about flight connections and bus transfer to Val di Fiemme. In Val di Fiemme, you plan to use **public transportation**.

Your task is to ask the APT agent for more information.

You have to choose:

- **an "all included" package** meeting your demands
- **a town** where you want to stay - close to a ski-area and an ice-skating facility.
- **a hotel** close to a bus stop or a ski-area.

*Please remember, you are calling an information center – **you can't make reservations** there.*

Please write below a couple of questions you would like to ask the APT agent in order to complete your task.

APPENDIX 15: Agent's Instructions (Experiment 2)

FINESTRA DELLA WHITE BOARD

Userai questa finestra per inviare pagine web al cliente e condividere mappe.

Pagine web

Per **aprire una pagina web** usa la funzione **"Open URL"** dal menù a tendina **"Tools"** della whiteboard.

Selezionando **"Open URL"** vi comparirà l'elenco delle pagine web a disposizione. Dopo la selezione della pagina web che interessa, si apre automaticamente una finestra di Browsing su cui la pagina web viene visualizzata.

Per aprire una **nuova pagina Web** chiudi prima la finestra di browsing aperta e ripeti la stessa procedura.

Eventuali **operazioni di scrolling** su una pagina web **non sono condivise con il client** (la pagina web del cliente rimane ferma).

Mappe

Prima di caricare una mappa...

... assicurati che la **whiteboard** sia **pulita**.

Se c'è già una mappa, **pulisci la Whiteboard** cliccando sull'apposita icona oppure selezionando dal menù **"File"** la funzione **"New"** della finestra di Whiteboard.

Non c'è bisogno di salvare le mappe prima di pulire la Whiteboard (vengono salvate automaticamente).

Per caricare una mappa...

... seleziona dal menù **"File"** la funzione **"Open"** oppure utilizza la corrispondente icona.

Compare una finestra: cercherai la mappa che ti serve nella **cartella Mappe** (attenzione, non in "mappe salvate), **sul disco C** del tuo computer.

Per utilizzare le funzioni grafiche...

- **seleziona una funzione grafica** (penna, linea, selezione ellittica, selezione rettangolare)
- **scegli uno dei colori** a disposizione (nero, bianco, rosso, giallo, blu, verde – di default c'è il nero)
- **traccia segni** sulla whiteboard. Puoi:
 - **indicare un oggetto** sulla mappa (es. hotel, parco, pattinaggio), disegnando una freccia oppure cerchiando l'oggetto in questione.

- **selezionare un'area** (es. un'area sciistica)
- **connettere 2 punti** distanti sulla mappa (es. 2 località)

Puoi inoltre:

- fare un **zoom** su un'area particolare
- usare le barre di **scroll (verticale e orizzontale)** per visualizzare parti diverse della mappa

Ricorda...

... che se apri una mappa o una pagina web, anche il cliente la potrà vedere, ma potrebbe riceverla con un certo ritardo o anche non riceverla in caso di problemi tecnici). Sarebbe quindi opportuno da parte tua:

- **preannunciare** il caricamento di una mappa o una pagina web per avvisare il Client (es. "Le mostro la mappa di...", "Le invio la pagina web di...", ...)
- **chiedere conferma** al cliente se vede la mappa e/o la pagina web (es. "riesce a vedere la mappa?")

IL TUO COMPITO

- **Rispondere, con frasi chiare e semplici**, fornendo solo le informazioni espressamente richieste dal cliente.
- Presentare i pacchetti (e le informazioni in genere) **seguendo le schede** che ti saranno messe a disposizione su carta (e non sulla base della tua conoscenza delle località, dei pacchetti, etc.).
- **Indicare sulle mappe** a disposizione **la dislocazione** di paesi, impianti sciistici, hotel, fermate dell'autobus, etc.

Ricorda che...

...non puoi fare prenotazioni (puoi solo dare informazioni sulla base del materiale che hai a disposizione; nel caso di richieste di prenotazioni o disponibilità di stanze, lasciate il numero di telefono dell'albergo)

... in caso di richiesta di **informazioni che vanno al di là del materiale a disposizione o che non sei autorizzata a fornire** (es. la disponibilità di camere in albergo), puoi:

- **inviare pagine web** in cui il cliente può trovare informazioni utili;
- **fornire un numero di telefono** (es. dell'hotel o dell'Ente Turistico Locale) che il cliente può chiamare direttamente.

MATERIALE A DISPOSIZIONE

In formato elettronico

- **Mappe**
 - Cavalese (**cavalese.bmp**)
 - Panchià (**panchia.bmp**)
 - Predazzo (**predazzo.bmp**)
 - Val di Fiemme invernale (ski-area) (**skiarea_fiemme.bmp**)
- **Pagine Web** (a fianco delle località sono riportati in grassetto i nomi dei file)
 - Cavalese (**Cavalese.htm**)
 - L'indice dei castelli (**castles_index.htm**)
 - Panchià (**Panchia.htm**)
 - Predazzo (**predazzo.htm**)

In formato cartaceo

- **La stampa delle mappe elettroniche**
- **Le schede contenenti informazioni sui pacchetti e sugli alberghi.**

APPENDIX 16: HCRC CODING SCHEME

The Transaction Coding Scheme

Transactions coding gives the sub-dialogue structure of complete task-oriented dialogues, with each transaction being built up of several dialogue games and corresponding to one step of the task. In most map-task dialogues, the participants break the route into manageable segments and deal with them one by one. Four transaction types were identified: normal, review, overview and irrelevant. Other types of sub-dialogue are possible, but were not included in the coding scheme because of their rarity.

NORMAL	A task related segment which opens and closes (? - My definition)
REVIEW	participants return to parts of the route which were previously discussed
OVERVIEW	participants overview an upcoming segment in order to provide a basic context for their partners
IRRELEVANT	segments which have nothing to do with the task (e.g. about the experimental setting)

The Game Coding Scheme

Although some natural dialogues are *well ordered* (once a game is opened the participants work on it without opening new games, and the intention of starting new games is explicitly shared), participants are free to initiate new games at any time (even while the partner is speaking), and these new games can introduce new purposes. In addition, natural dialogue participants often fail to make clear to their partners what their goals are. This makes it very difficult to develop a reliable coding scheme for complete game structure.

The game coding scheme simply records those aspects of embedded structure which are of the most interest. First the beginning of a new game is coded, naming the game's purpose according to the game's initiating move, not all initiating moves begin games, Second, the place where games end or are abandoned is marked. Finally, games are marked as either occurring at top level or being embedded (at some unspecified depth) in the game structure.

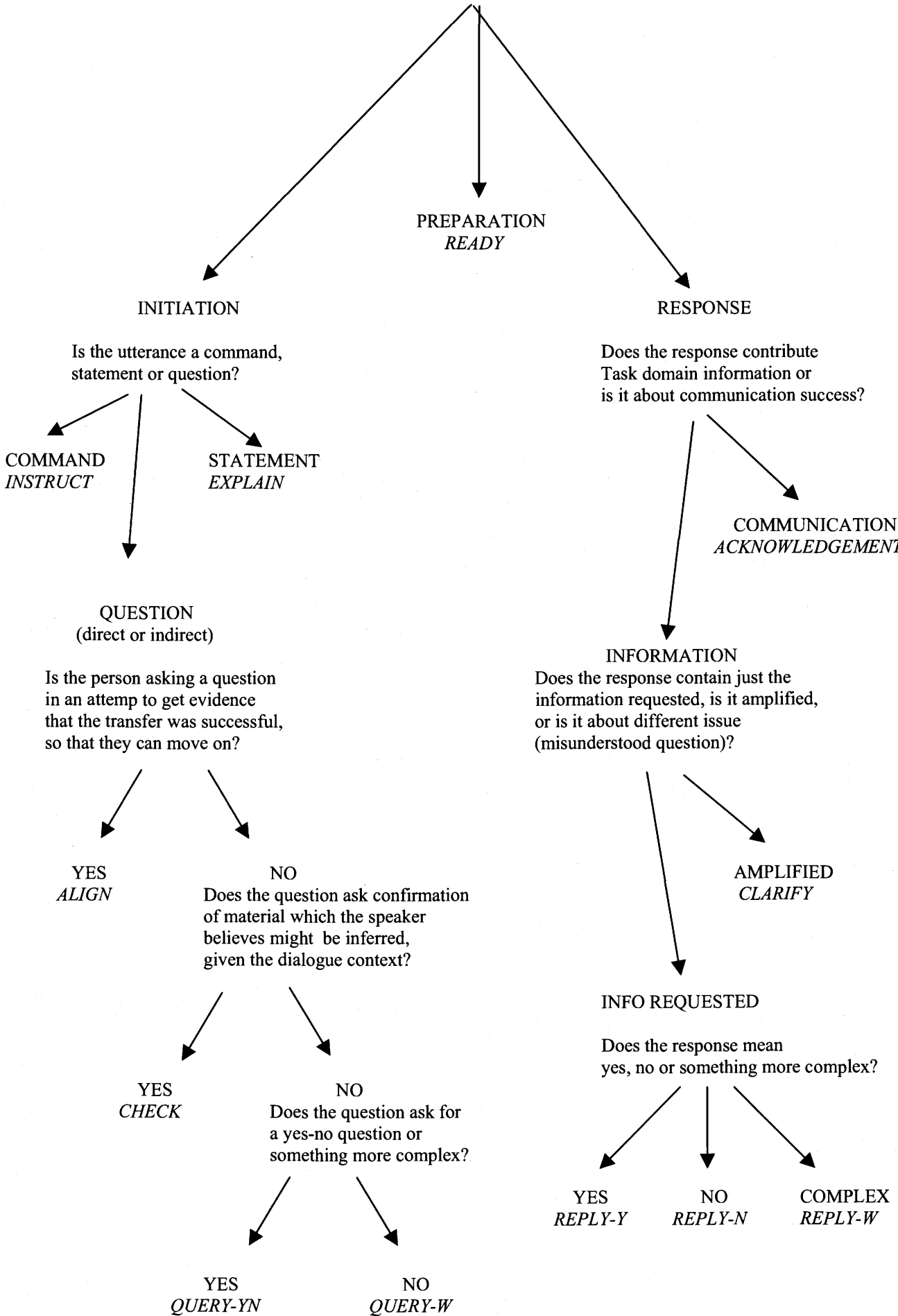
The Move Coding Scheme

The list of moves is available in the following table.

The HCRC decision tree for labelers is available in the following page, while the adapted version used in NESPOLE! is available in the following page.

INITIATING MOVES:	HCRC definition	EXAMPLES
INSTRUCT	direct or indirect request or instruction , to be done immediately or shortly	1. Go right round, ehm, until you get to just above them. 2. I fyou come in a wee bit so that you're about an inch away from both edges. 3. Say it... start again.
EXPLAIN	describes status quo or position in task with respect to the goal, freely offered , not elicited; provides new information	I'm in between the remote village and the pyramid. 2. I have to jump a stream. 3. I've got a great viewpoint away up in the top left-hand corner.
CHECK	checks self-understanding of a previous message or instruction by requesting confirmation directly or indirectly	1. to my right? 2. ok, up to the top of the stile?
QUERY-YN	yes-no question	1. Do you have a stone circle at the bottom? 2. Is it written underneath the tree?
QUERY-W	open-answer Wh-question . As the previous one, asks for a new or unknown detail; does not request clarification about instruction	1. towards what? 2. left of the bottom or left of the top of the chestnut tree?
ALIGN (+META?)	checks the other participant's understanding or accomplishment of a goal; elicits a positive response which closes a larger game; checks attention, agreement, or readiness	1. ok? 2. This is the left-hand edge of the page, right?
OTHER MOVES:		
REPLY-Y; REPLY-N	affirmative or negative - reply , elicited response to QUERY-YN, CHECK, or ALIGN; also indicates agreement, disagreement or denial	1. Yeah. 2. I do. 3. No, no at the moment.
REPLY-W	elicited response to QUERY-W or CHECK; can be a response to QUERY-YN that is not easily categorizable as positive or negative	1. it is the red one.
ACKNOWLEDGE	Verbal response/ vocal acknowledgement of having heard and understood ; not specifically elicited but often expected before the other speaker will continue; announce readiness to hear next move (in essence, a request of "please continue"); may close a game.	1. Mmhmm.
CLARIFY	clarifies or rephrases what has previously been said; usually repeats given or known information.	
READY	indicates intention to begin a new game and focuses attention on oneself, in preparation for the new move; an acknowledgement that the previous game has just been completed, ore leaving the previous level or game; consist of a cue-word (e.g. now, right)	1. Okay. Now go straight down.
UNCODABLE	It is not possible to categorize it, since it is impossible to understand.	

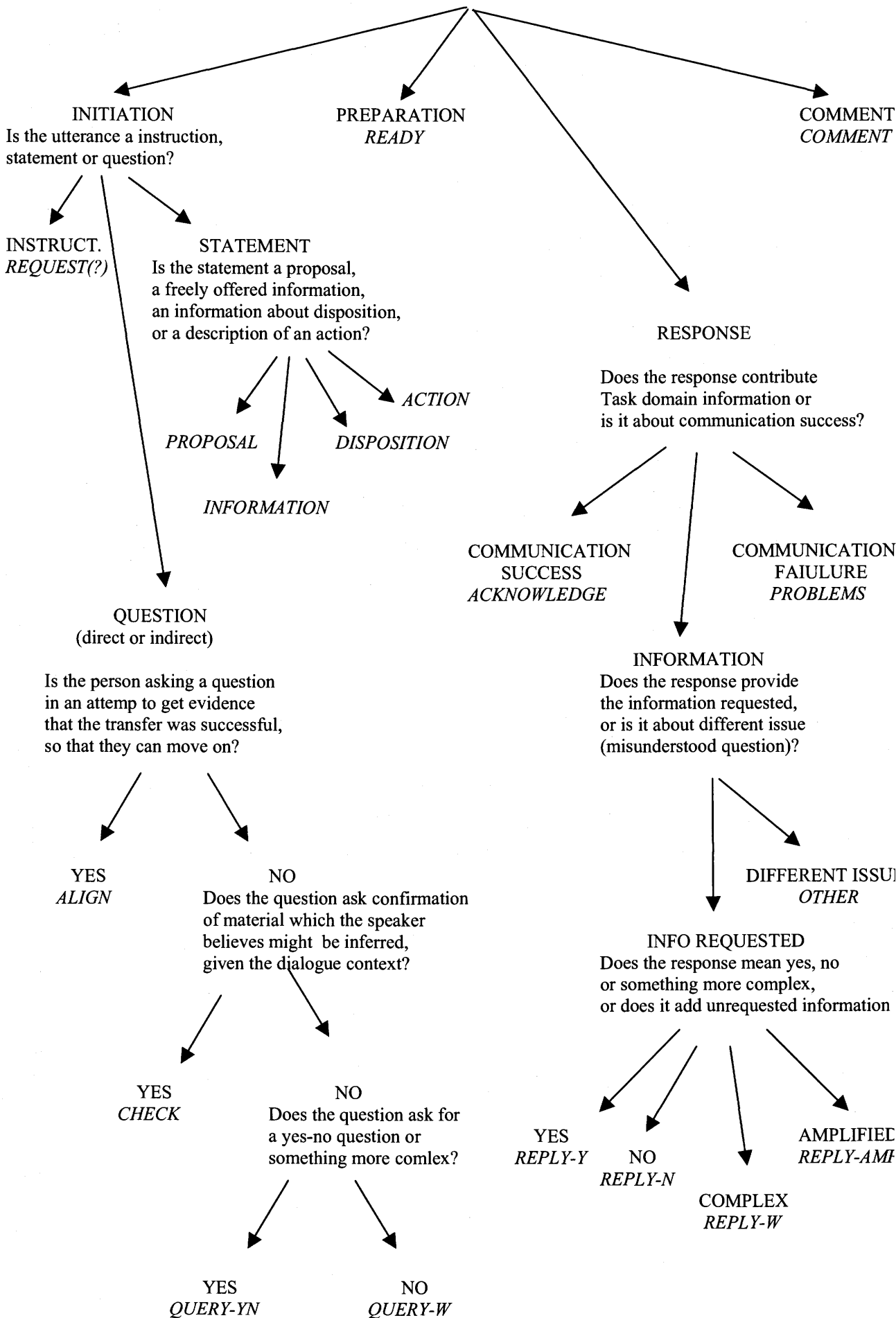
INITIATION, RESPONSE OR PREPARATION?



APPENDIX 17: Decision Tree for Dialogue Annotation

(NESPOLE! Version)

INITIATION, RESPONSE, PREPARATION, OR COMMENT?



Usability Questionnaire on the NESPOLE! System

Experiment 2, STST condition, customer

This questionnaire aims at collecting useful information to evaluate the usability of the NESPOLE! system.

The first 16 answers of the questionnaire should be given referring to a graduate scale that measures your personal agreement or disagreement with reference to the assertions reported below

- complete agreement*
- partial agreement*
- partial disagreement*
- complete disagreement*

After that, you will find 4 open questions.

We inform you that we guarantee your anonymity. The questionnaire will be identified though the file name of the dialogue you took part. Your personal

Your code: _____

1. The function of each element present on the screen is easily understood.

complete agreement	partial agreement	partial disagreement	complete disagreement
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

2. The graphic presentation is pleasant.

complete agreement	partial agreement	partial disagreement	complete disagreement
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

3. The feedback given by the system is inadequate.

complete agreement	partial agreement	partial disagreement	complete disagreement
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

4. I think that the use of maps can improve the communication.

complete agreement	partial agreement	partial disagreement	complete disagreement
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

5. I think that the use of web pages is useless.

complete agreement	partial agreement	partial disagreement	complete disagreement
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

6. I think that the possibility to draw on the map can improve the communication.

complete agreement	partial agreement	partial disagreement	complete disagreement
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

7. I think that the video is useless.

complete agreement	partial agreement	partial disagreement	complete disagreement
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

8. The translation is accurate.

complete agreement	partial agreement	partial disagreement	complete disagreement
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

9. The system is difficult to use.

complete agreement	partial agreement	partial disagreement	complete disagreement
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

10. The system is effective.

complete	partial	partial	complete
----------	---------	---------	----------

agreement agreement disagreement disagreement

11. The system is not helpful.

complete partial partial complete
agreement agreement disagreement disagreement

12. The system is innovative.

complete partial partial complete
agreement agreement disagreement disagreement

13. I received all the information I needed.

complete partial partial complete
agreement agreement disagreement disagreement

14. Which are the main difficulties you encountered during the dialogue?

15. Which elements/functionalities would you add to improve the communication?

16. Which elements/functionalities would you remove to improve the communication?

17. Would you use this system as substitute for others (phone, e-mail) to ask for tourist information? Why?

Thanks for your contribution!